TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Использование линейной регрессии для оценки риска заемщика при потребительском кредитовании

УДК 004.65:004.451:519.233:336.77:330.567.22

Студент

Группа	ФИО	Подпись	Дата
8ПМ0И	Юй пайшэн		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОСГН ШБИП	Меньшикова Е. В.	к.ф.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ООД ШБИП	Антоневич О. А.	к.б.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Савельев А.О.	к.т.н.		

Томск – 2022

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП

по направлению 09.04.04 «Программная инженерия»

Код компетенции	Наименование компетенции
Универсальные компетенции	
УК(У)-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке (-ах), для академического и профессионального взаимодействия
УК(У)-5	Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
Общепрофессиональные компетенции	
ОПК(У)-1	Способен самостоятельно приобретать, развивать и применять математические, естественно-научные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ОПК(У)-2	Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач
ОПК(У)-3	Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ОПК(У)-4	Способен применять на практике новые научные принципы и методы

	исследований
ОПК(У)-5	Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем
ОПК(У)-6	Способен самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности
ОПК(У)-7	Способен применять при решении профессиональных задач методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях
ОПК(У)-8	Способен осуществлять эффективное управление разработкой программных средств и проектов
Профессиональные компетенции	
ПК(У)-1	Способен к созданию вариантов архитектуры программного средства
ПК(У)-2	Способен разрабатывать и администрировать системы управления базам данных
ПК(У)-3	Способен управлять процессами и проектами по созданию (модификации) информационных ресурсов
ПК(У)-4	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий
ПК(У)-5	Способен осуществлять руководство разработкой комплексных проектов на всех стадиях и этапах выполнения работ

TOMSK POLYTECHNIC UNIVERSITY
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
__________ _________ Савельев А.О.
(подпись) (дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	ФИО
8ПМ0И	Юй пайшэн

Тема работы:

Использование линейной регрессии для оценки риска заемщика при потребительском кредитовании	
Утверждена приказом директора (дата, номер)	№ 45-47/с от 14.02.2022

Срок сдачи студентом выполненной работы:	15.06.2022

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	Объектом исследования является алгоритм обработки данных.

Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).*	1.Обзор методов анализа данных. 2. Исследование предметной области, выбор метода анализа данных. 3. Проектирование ПО (веб-сайта, мобильного приложения). 4. Тестирование ПО. 5. Работа над разделом по финансовому менеджменту, ресурсоэффективности и ресурсосбережения. 6. Работа над разделом по социальной ответственности. 7. Работа над разделом на английском языке.
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. UML-диаграммы. 2. Скриншоты веб-форм. 3. Карта веб-платформы. 4. Диаграмма Исикавы. 5. Диаграмма Ганта.

Консультанты по разделам выпускной квалификационной работы

(с указанием разделов)

Раздел	Консультант
Основная часть	Доцент ОИТ ИШИТР, к.ф.-м.н., доцент Губин Е. И.
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Доцент ОСГН ШБИП, к.ф.н., доцент Меньшикова Е. В.
Социальная ответственность	Доцент ООД ШБИП, к.б.н., доцент Антоневич О. А.
Английский язык	Доцент ОИЯ, к.ф.н., доцент Айкина Т. Ю.

Названия разделов, которые должны быть написаны на русском и иностранном языках:

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	1.03.2022

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н., доцент		1.03.2022

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ0И	Юй пайшэн		1.03.2022

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Уровень образования магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения весенний семестр 2021 /2022 учебного года

Форма представления работы:

Магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
10.06.2022	Основная часть	70
10.06.2022	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
10.06.2022	Социальная ответственность	10
10.06.2022	Английский язык	10

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

СОГЛАСОВАНО:
Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Савельев А. О.	к.т.н.		

Tomsk – 2022

TASK FOR SECTION
«FINANCIAL MANAGEMENT, RESOURCE EFFICIENCY AND RESOURCE SAVING»

To the student:

Group	Full name
8PM0И	Юй Пайшэн

School	ИШИТР	**Division**	ОИТ
Degree	Master	**Educational Program**	09.04.04Software Engineering

Input data to the section «Financial management, resource efficiency and resource saving»:	
1. *Resource cost of scientific and technical research (STR): material and technical, energetic, financial and human*	– Salary costs – 81766 – STR budget – 140954
2. *Expenditure rates and expenditure standards for resources*	– Electricity costs – 5,8 rub per 1 kW
3. *Current tax system, tax rates, charges rates, discounting rates and interest rates*	– Labor tax – 27,1 %; – Overhead costs – 30%;
The list of subjects to study, design and develop:	
1. *Assessment of commercial and innovative potential of STR*	– comparative analysis with other researches in this field;
2. *Development of charter for scientific-research project*	– SWOT-analysis;
3. *Scheduling of STR management process: structure and timeline, budget, risk management*	– calculation of working hours for project; – creation of the time schedule of the project; – calculation of scientific and technical research budget;
4. *Resource efficiency*	– integral indicator of resource efficiency for the developed project.

A list of graphic material *(with list of mandatory blueprints):*

1. *Competitiveness analysis*
2. *SWOT- analysis*
3. *Gantt chart and budget of scientific research*
4. *Assessment of resource, financial and economic efficiency of STR*
5. *Potential risks*

Date of issue of the task for the section according to the schedule	

Task issued by adviser:

Position	Full name	Scientific degree, rank	Signature	Date
Associate professor	E.V. Menshikova	PhD		

The task was accepted by the student:

Group	Full name	Signature	Date
8PM0I	Yu Paisheng		

Tomsk – 2022

«SOCIAL RESPONSIBILITY»

Student:

Group	Name
8РМ0И	Yu Paisheng

School	ИШИТР	Division	ОИТ
Educational level	Magistracy	Course/Specialty	09.04.04 Software Engineering

Topic of FQW:

Use linear regression to assess the borrower's risk Consumer loan	
Initial data for the chapter «social responsibility»:	
1. Characteristics of the researched object (substance, material, device, algorithm, technique, working area)	Scope of application: software engineering Work area: computer classroom Research Algorithms: Machine Learning Work area equipment number and name: computer
List of questions to be researched, designed and developed:	
1. Legal and organizational issues of occupational safety consider special (specific to the projected work area) law norms of labor legislation. indicate the features of the labor legislation in relation to the specific conditions of the project.	list normative documents 1.GOST 12.1.003-2014 SSBT. Noise. General safety requirements, 2.GOST 30691 (ISO 4871:1996) Machine noise. 3.artificial lighting. Updated edition of SNiP 23-05-95*; 4.SP 60.13330.2020 Heating, ventilation and air conditioning SNiP 41-01-2003 5.GOST 9241-4-2009 and GOST 9241-4-2007
2. Occupational safety: 2.1. Analysis of the identified harmful and dangerous factors: the sourse of factor, the impact on human ' s body 2.2 Suggest measures to reduce the impact of identified harmful and dangerous factors	Possible harmful and dangerous factors: 1.Exceeding the noise level 2.Insufficient illumination of the working area 3.Temperature in the work area 4.Physical overload
3. Environmental Safety: Influence on the atmosphere, hydrosphere, lithosphere	Analysis of Electronic wastes Analysis of Hazardous Pollutants Analysis of Waste Electronic Components Recycling
4. Emergency Safety: describe the most likely emergency situation	СанПиН 1.2.3685-21 и СП 12.13130.2009 ГОСТ Р 12.3.047-98 ССБТ «Пожарная безопасность технологических процессов. Анализ пожарной безопасности
Date issue of the task for the chapter	

Consultant:

Position	Full name	Scientific degree, rank	Signature	Date
Associate professor	Antonevich O. A	PhD		

Student:

Group	Name	Date	Signature
8РМ0И	Yu Paisheng		

Tomsk – 2022

Abstract

Final qualified work 78 pages, 13 figures, 20 tables.Keywords: data analysis, data preparation, data cleaning, linear regression, random forest, neural network, credit scoring, credit rating, scorecard, credit risk model.

The object of the study is the credit data of borrowers.

This topic studies the credit risk model and makes a flask framework for prediction.

The goal of this project is to use python programming to compare several models, and then get the optimal solution linear regression model to help banks make decisions, control risks, select more good borrowers, and remove bad borrowers from banks.

In research projects, judge models and datasets through analysis and saliency testing, and use Python to create programs that calculate credit risk scores.

The results show that the use of linear regression can improve the accuracy of credit evaluation. According to the predicted repayment month, to judge the repayment situation, the conclusion is that a program calculator is made.

Basic Design, Technology and Technical and Operational Characteristics: The developed methodology allows assessing the creditworthiness of potential borrowers.

Scope: The technology developed can be used to improve the accuracy of bank credit scoring, and the calculator interface makes credit risk models easier to use.

Contents

INTRODUCTION

The development of consumer finance can effectively promote people's consumption levels and promote economic circulation, but the credit system is imperfect, so there have been long-term consumer defaults.Credit scoring models have become more common as banks realize the importance of credit risk management. How to choose a model to evaluate also affects the bank's judgment of customers. How to assess the credit risk of debtors has become an important factor in the development of consumer finance.Among them, models such as linear regression and random forest are widely used to predict the repayment situation.

First through a preliminary analysis of the dataset,and the compare with other models, and finally the linear regression model is selected as the prediction model. Then use the linear regression model and the flask framework to realize the prediction of the repayment month.Through the predicted repayment month, the relevant repayment risk can be judged according to the specific business content.

1.Organize data

Because the financial industry needs to process huge amounts of data every minute, machine learning has proven beneficial in improving services and operations in the financial sector. It's hard to imagine the world of finance without machine learning. Many financial institutions are leveraging artificial intelligence technologies, including machine learning, to improve efficiency in their day-to-day operations.Excellent customer service is one of the key indicators of a financial institution's success.

When a bank's customer service is terrible, customers tend to abandon them in search of better service providers. Machine learning also has specific applications in the field of customer service. Ordinary chatbots have limited information and cannot effectively solve customer problems. However, machine learning enables chatbots to learn and solve customer queries based on customer behavior.

First we need to download a dataset about credit cards.In this dataset, each entry represents a person who takes a credit by a bank. Each person is classified as good or bad credit risks according to the set of attributes. Below are the values in the dataset.

which includes:

1. Age
2. Sex (1 - male,2 - female)
3. Job (0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
4. Housing (1 -own,2 - rent, 3 - free)
5. Saving accounts (1- little, 2 -moderate,3 - rich)
6. Checking account (1- little, 2 -middle,3 - high)
7. Credit amount
8. Duration (month)

1.1Data preprocess method

Due to the fact the actual collected machine learning data sets will inevitably have data shortages, imbalanced data sets, and various types of data in the datasets are not of the same magnitude, etc., the missing data is completed and the abnormal data is cleaned and balanced. Preventing class imbalance and data normalization is critical for machine learning models.

1.1.1Data Completion method

Effectively recovering missing data is an important preparatory work for machine learning modeling. On the one hand, it can make the data more complete, which is convenient for further analysis and research; on the other hand, data completion itself is a way of mining information.

1.1.2 Data balance method

Due to practical situations, such as in the credit card fraud detection dataset, most types of credit card transactions are not fraudulent, and only a few types are fraudulent transactions, so the ratio between non-fraudulent transactions and fraudulent transactions reaches 50:1, resulting in the dataset is unbalanced, it is necessary to balance the data to balance the various types, so as to avoid the negative effects caused by the unbalanced types. Common equalization methods include "undersampling" and "oversampling". Randomly discarding some data in the class to ensure class balance, while "oversampling" is to increase the sampling frequency of a smaller number of samples or interpolate the training set data to ensure class balance. "Undersampling" will reduce the size of the training data and may lose data, while "oversampling" will lead to severe overfitting if the initial data is directly sampled multiple times.

1.1.3 Data normalization method

Data normalization is a basic work of mining data in machine learning. The conversion becomes a relative value with a relative relationship, and the amount of data is reduced to a specific range.

1.2Data process

Before data cleaning, the distribution of the data should be analyzed.By observing the boxplots in the figure below,shows the variables Age, Sex, Job, Housing, Saving accounts, Checking account, Credit amount, Duration.

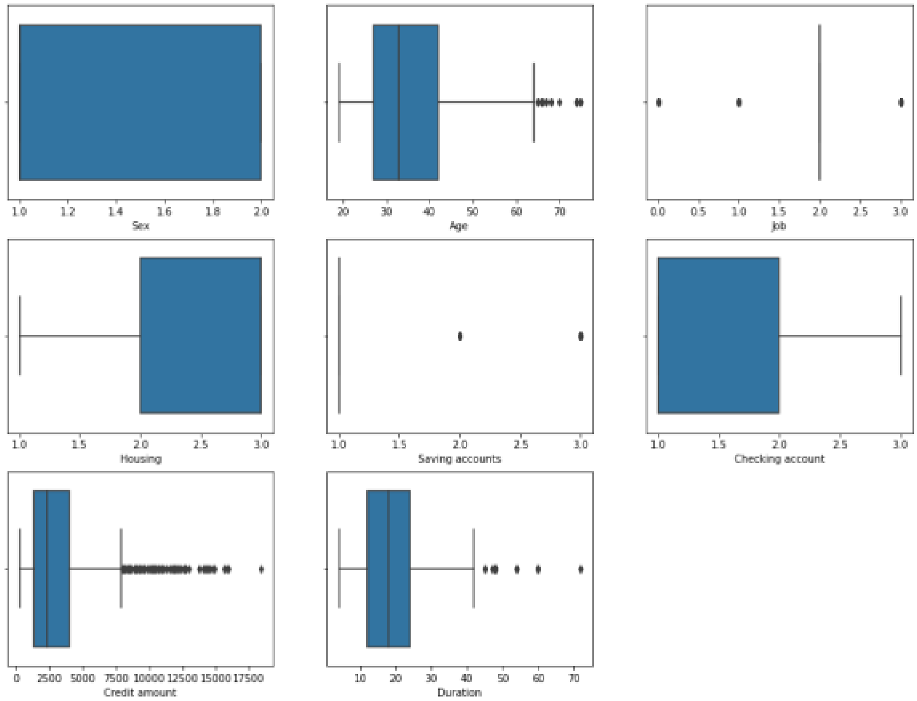


Figure 1 –Box plot

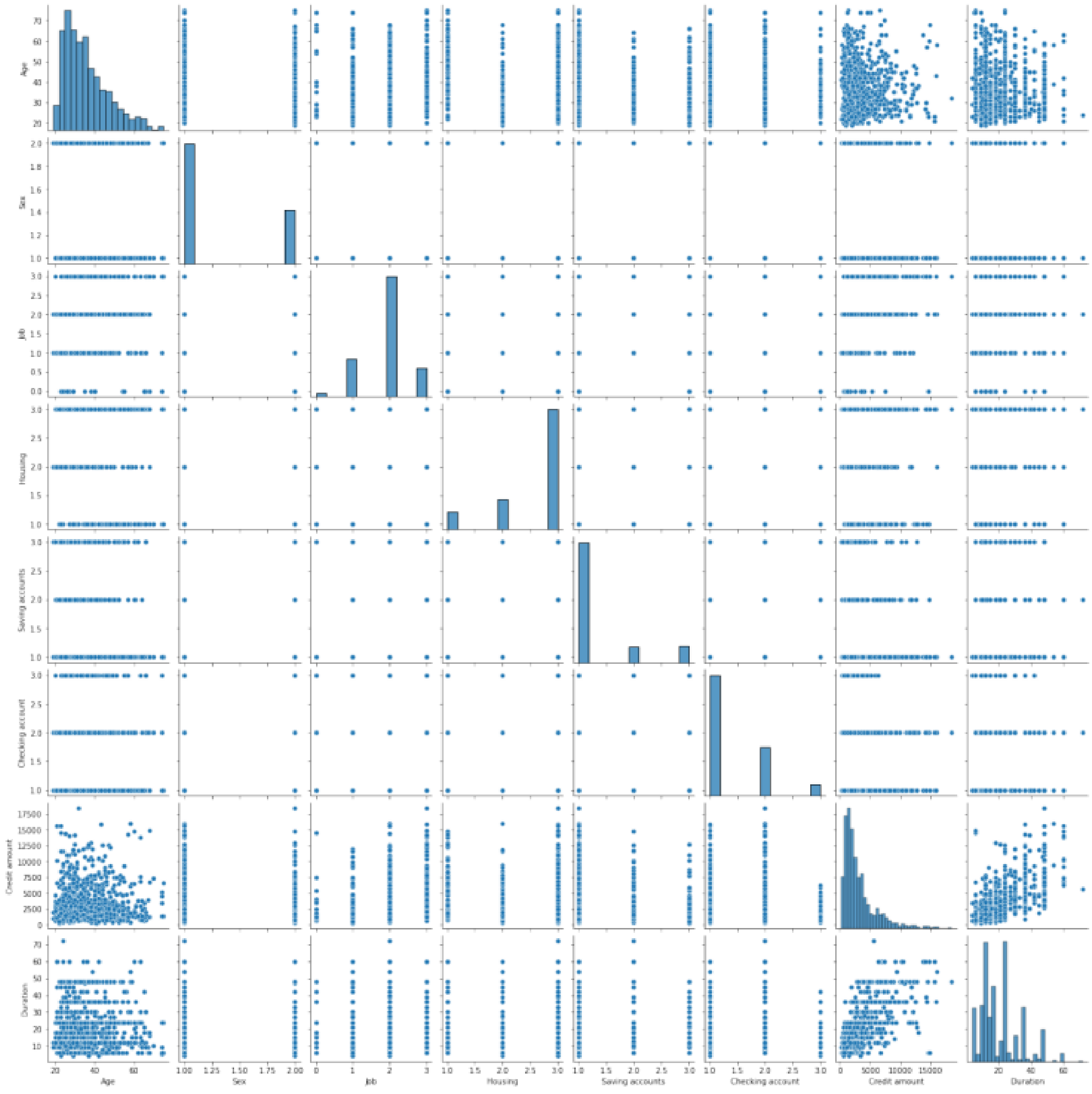


Figure 2 Scatter diagrams of raw data

We process gender, housing, Saving accounts, Checking account separately and turn them into numbers. And delete the feature of Purpose.For this i edited a py file to process the data and store it in csv format(clean.py).

After cleaning the data, we use describe function to know the mean,standard deviation of each data.

Table 1 -mean and standard deviation of each value

	Mean	Standard deviation
Age	35.546	11.375
Sex	1.310	0.463
Job	1.904	0.654
Housing	2.605	0.675
Saving accounts	1.325	0.665
Checking account	1.395	0.604
Credit amount	3271.258	2822.737
Duration	20.903	12.059

Through the method of sns.displot, make a histogram about Duration to observe the normal distribution of Duration.From the histogram, it can be seen that Duration basically conforms to the normal distribution, and the value of Duration is mainly concentrated in 20 to 30 months.

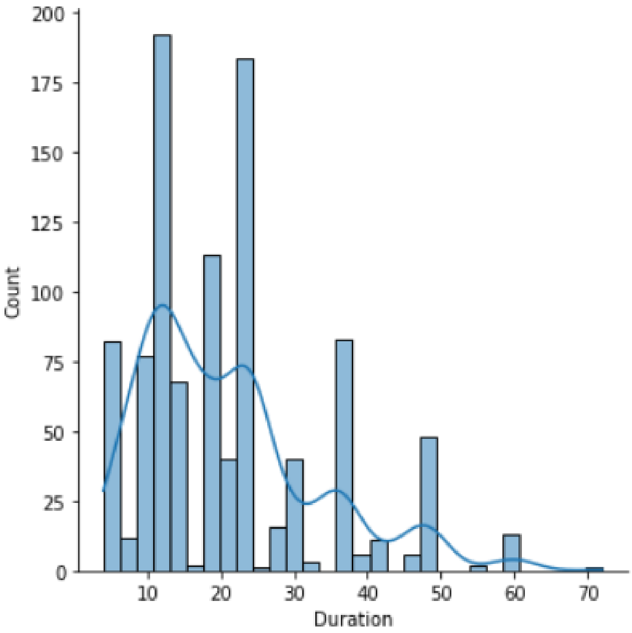


Figure 3 –Histogram of Duration

Finally,use the pearson correlation coefficient is to determine whether each feature is closely related. Through the following picture, the correlation of each feature value can be observed.

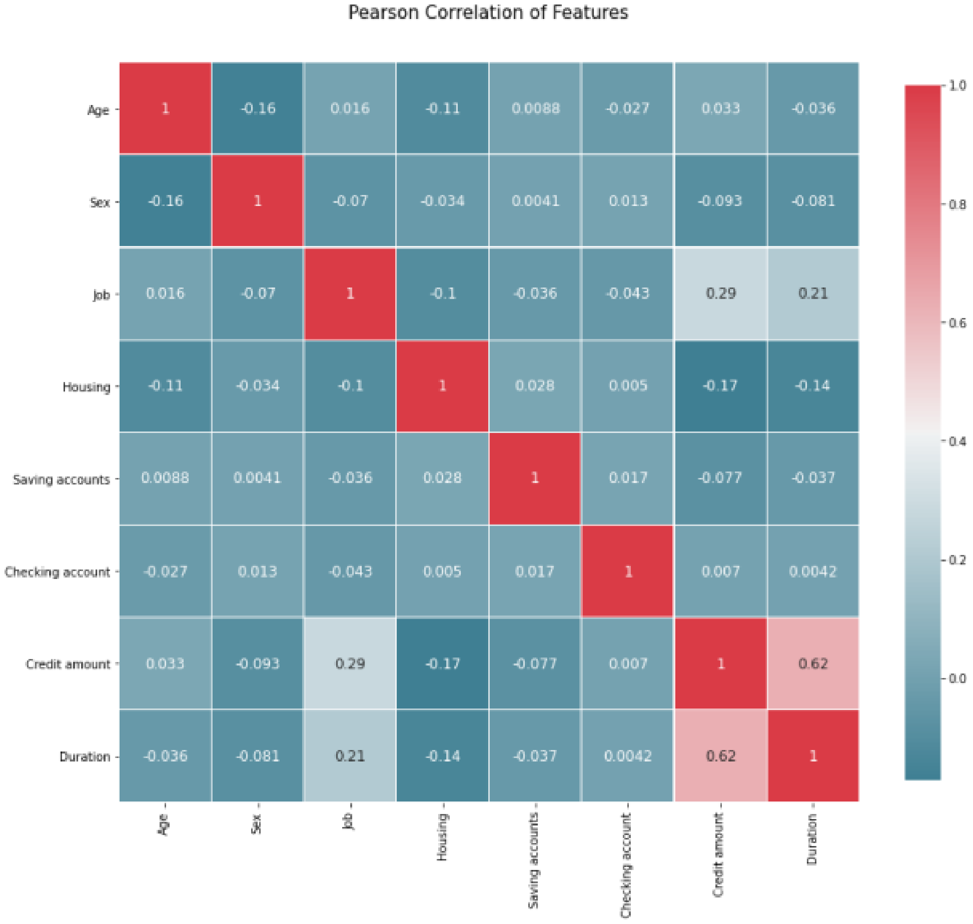


Figure 4 –correlation heatmap

2.Research method

The most common method for assessing a borrower's creditworthiness in banks is credit scoring. Credit scoring is an automated system based on a predictive mathematical model that uses a bank's credit history to predict the likelihood that a potential borrower will repay the loan on time . The forecast is based on information about the credit history, social and demographic parameters, data on the requested loan. Because we are analyzing the repayment month, we choose the Regressions in sklearn.

Regression analysis is a predictive modeling technique that studies the relationship between a dependent variable (target) and an independent variable (predictor). This technique is often used for predictive analysis, time series modeling, and finding causal relationships between variables.

2.1Compare models

Machine learning is the extraction of knowledge from data and is the main method to achieve artificial intelligence. Machine learning is a field of study at the intersection of statistics, artificial intelligence and computer science, also known as predictive analytics or statistical learning. As a core component of artificial intelligence, machine learning is generally defined as the behavior of a computer to judge and automatically improve its own performance through its own experience. It is not difficult to understand that this is actually letting the machine simulate human behavior and extract improvements in the existing operation of the computer system. Machine learning represented by deep learning is the current intelligent learning method and cognitive process closest to the human brain. Layer analysis and processing mechanism, self-adaptive, self-learning powerful parallel information processing ability has achieved great commercial success in many application fields.

In this article, we use regression models ,because the month of repayment is studied. The goal of regression is to predict a continuous value. A simple way to differentiate between classification and regression is to ask the question: Does the output have some kind of continuity. If there is continuity between the possible outcomes, then it is a regression problem. On the contrary, it is a classification problem. This paper mainly discusses the regression problem.

When solving machine learning problems, data scientists often tend to pay attention to model performance metrics such as accuracy, precision, and recall. However, metrics can only tell a fraction of what a model is predicting decisions about. Over time , performance may change due to model conceptual drift caused by various factors in the environment. Therefore, it is extremely important to understand what drives the model to make certain decisions.Understanding the reasons behind predictions is important in assessing trust, and trust in models is critical if planning to act on predictions, or choosing whether to deploy a new model.

In this part, we train linear regression model, random forest model and neural network model separately. Compare several models to screen out the best model as the model for predicting Duration.

By consulting the official documentation of sklearn, we get mean_squared_error, mean_absolute_error, r2_score as a reference to compare several regressor models.

mean_squared_error

mean_squared_error calculate the mean squared error regression loss.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

mean_absolute_error

mean_absolute_error is the average value of the absolute value of the error can accurately reflect the size of the actual prediction error.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \bar{y}_i|$$

r2_score

r2_score function is computes the coefficient of determination, usually denoted as R^2.The larger the value of R^2, the better the fit; the smaller the value of R^2, the worse the fit.R^2 measures how well the model predicts unseen samples by goodness of fit, and thus the proportion of variance explained.

2.1.1 Random Forest

Random forest is an ensemble algorithm, Because the algorithm has high accuracy and generalization performance, it is widely used in the field of machine learning. Where "random" prevents data overfitting, and "forest" makes the results more accurate.

Random forest is a more advanced algorithm based on decision trees. Random forests can be used for both regression and classification. As can be seen from the name, random forest is a forest constructed in a random way, and this forest is composed of many unrelated decision trees. Real-time random forests are essentially a very important branch of machine learning called ensemble learning. Ensemble learning solves a single prediction problem by building a combination of several models. It works by generating multiple classifier models, each of which learns and makes predictions independently. These predictions are finally combined into a single prediction, thus outperforming any single-class prediction. Therefore, in theory, Because the results of random forests are voted by multiple decision trees, random forests generally outperform decision trees in terms of performance.Simply put, each decision tree in the random forest has its own result, and the random forest selects the result with the most votes as its final result by counting the results of each decision tree.

Random forest has many advantages. For example, due to the use of ensemble algorithms, its accuracy is better than most single algorithms; in industry, due to the introduction of two randomness, random forest has a certain anti-noise ability, compared with other algorithms. Advantages; default values can be processed without additional processing; since each tree can be generated independently and at the same time, it is easy to make a parallel method.

Random forests also have some disadvantages. Overfitting on some noisy classification or regression problems may affect the accuracy of the results. In problems with multiple categorical variables, random forests may not improve the accuracy of the base learner.

In order to get a more accurate model, the parameter adjustment of the random forest regressor also needs to be carried out according to the relevant smoothness.

Table 2 - Parameters of RandomForestRegressor

Parameters	meaning
max_depth	maximum depth of the trees
min_samples_leaf	The minimum number of samples required to be at a leaf node
max_features	The number of features to consider when looking for the best split

By adjusting n_estimators,max_depth,min_samples_split and max_features.Finally, the r2_score, mean_squared_error, mean_absolute_error of rf are 0.448, 87.65, 7.071.

Table 3 - Comparison before and after adjusting parameters

	r2_score	MAE	MSE
before adjustment	0.316	7.922	104.8
After adjustment	0.448	7.071	87.65

Another advantage of the random forest algorithm is that the relative importance of each feature to prediction can be easily measured. Sklearn provides a great tool for this by looking at using this feature.The feature reduces the degree of impurity of all trees in the forest to measure the importance of the feature. It automatically computes a score for each feature after training and normalizes the results so that the sum of the importance of all features equals 1.

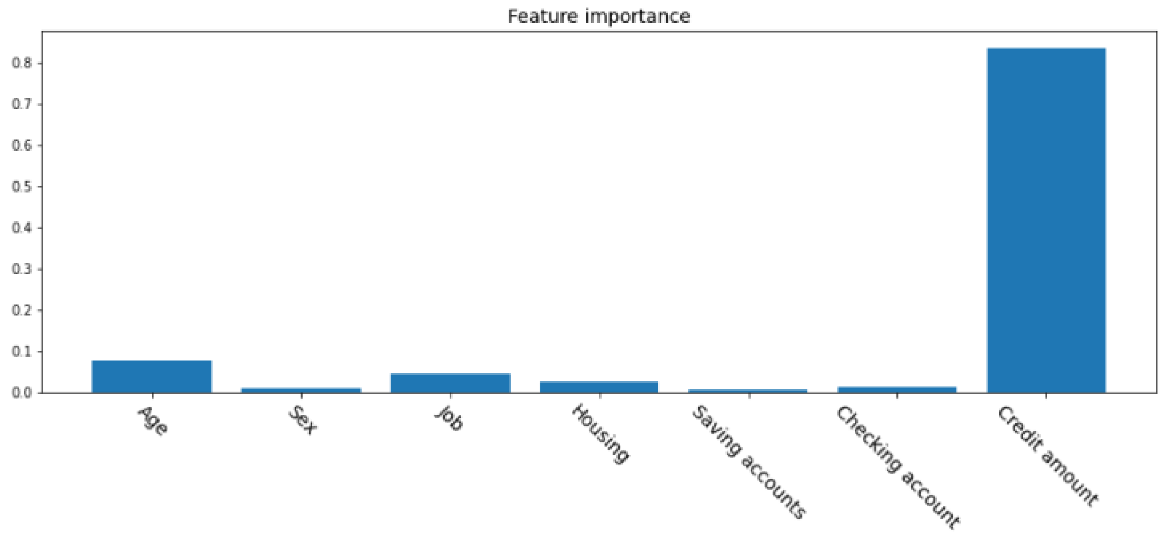


Figure 5 – Feature importance of Random Forest

By observing the proportion of different eigenvalues, it can be seen that the predict amount has a great influence on the change of Duration. Therefore, when there are few features available, we can only make relevant models for a few features such as credit amount to predict Duration.

2.2.2.Neural Networks

Neural networks have been proven to have good predictive effects in many fields. Especially in the three fields of modeling and prediction, signal processing, and expert systems, neural networks have played their advantages. The predictive ability of the neural network is related to its automatic association memory ability. Once the network is trained, it can input new data for prediction and output the prediction result.

The overall purpose of neural network analysis is to map them to output variables by capturing the differences between input variables. The data enters the network through the input node, then transmits the data through the hidden layer node, and finally passes to the output node.

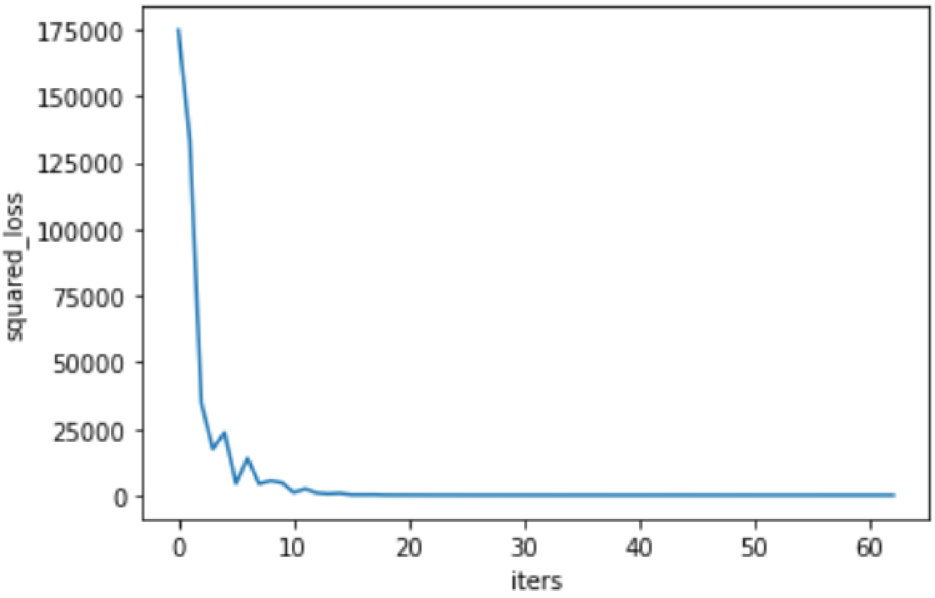


Figure 6 –Visualize the loss function

The advantage of neural network is that it can use mathematical formulas to express the complex relationship between input nodes and output nodes, that is, the relationship between independent variables and dependent variables. In order to build a neural network model, the network must be trained when the output nodes are known. The training process is to input a part of the known data into the network to "teach" the network to recognize the type of data.

While neural networks have many advantages, neural networks also have some disadvantages. Probably the most well-known drawback of neural networks in the first place is their "black box" nature, which means don't know how and why a neural network produces a certain output. In comparison, algorithms like decision trees are very easy to understand. This is important because in some domains, interpretability is very important. That's why a lot of banks don't use neural networks to predict whether a person is creditworthy, because they need to explain to customers why they didn't get a loan. Otherwise, the person might feel wrongly threatened by the bank because he doesn't understand why he didn't get the loan, which could lead him to change his mind about the bank. If they decide to delete a user account because of a machine learning algorithm, they need to explain to the user why they have done it.

At the same time neural networks are more computationally expensive than traditional algorithms. State-of-the-art deep learning algorithms, enabling the successful training of truly deep neural networks, can take weeks to train entirely from scratch. Most traditional machine learning algorithms take less than minutes to hours or days. The computational power required for a neural network depends largely on the size of the data, but also on the depth and complexity of the network. For example, a neural network with one layer and 50 neurons will be much faster than a random forest with 1000 trees. In contrast, a neural network with 50 layers will be much slower than a random forest with only 10 trees.

Regarding the parameters of MLPRegressor, hidden_layer_sizes is The ith element represents the number of neurons in the ith hidden layer.activation is activation function for the hidden layer.In the neural network, the job of the hidden layer is to convert the input into something that the output layer can use. The output layer converts the hidden layer activations to whatever scale you want the output to be at.Through the following table, can see the changes before and after adjustment.

Finally, the hidden_layer_sizes of the model are adjusted to (1000,500).And the r2_score, mean_squared_error, mean_absolute_error of MLPRegressor are 0.4033, 94.75, 7.42.

Table 4 - Comparison before and after adjusting parameters

	r2_score	MAE	MSE
before adjustment	0.1605	8.197	128.7
After adjustment	0.4033	7.42	94.75

2.2.3 Linear Regression

As a classic algorithm, linear regression is widely used in many theories. Linear regression is based on the least squares method. By studying the relationship between independent variables and dependent variables, an equation is obtained to establish a model.

In linear regression, the data is modeled using a linear prediction function, and unknown model parameters are also estimated from the data. Linear regression is a linear method to simulate the relationship between the dependent variable and one or more independent variables; for the model, the independent variable is the input value, and the dependent variable is the output value of the model based on the independent variable, suitable for x and y Application scenarios of data types that satisfy linear relationships.

There are two main scenarios in which linear regression is applied to data analysis:

1.Driving force analysis: a dependent variable indicator is affected by multiple factors, and analyzes the strength of the driving force of different factors on the dependent variable (driving force refers to correlation, not causality);

2.A prediction in which the independent variable has a linear relationship with the dependent variable;

Finally, after analysis r2_score, mean_absolute_error and mean_squared_error of the three models in the table 4, it can be seenLinearRegression is the best model for the dataset.This is also the reason why I choose linear regression to analyze.

Table 5 - Comparison of three machine learning models

	r2_score	MAE	MSE
Linear Regression	0.4472	7.209	87.78
RandomForest Regressor	0.448	7.071	87.65
MLP Regressor	0.4033	7.42	94.75

2.2Analysis of linear regression model

Through the comparison of the above machine learning models and the comparison of different eigenvalues, it can be concluded that there is a great correlation between the Credit amount and the Duration. In order to study the correlation between the two, we use python and r respectively to analyze.

For this reason, we calculate the correlation between Credit amount and the Duration, and the result is 0.63. It can be seen that the values of the two columns are moderately correlated and positively correlated.

The formula for calculating the correlation coefficient is

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 * \sum (y - \bar{y})^2}}$$

The significance test is to test the hypothesis we make about the population, and its principle is the "principle of practical impossibility of small probability events". We don't know the population, then, make a hypothesis about the population, then use the sample to judge the hypothesis, and see the difference between the hypothesis and the sample, this process is the significance test.

The hypothesis to be tested is recorded as H0, called the null hypothesis, and the hypothesis opposed to H0 is recorded as H1, called the alternative hypothesis.

1.When the null hypothesis is true, it is decided to abandon the null hypothesis, which is called the first type of error, and the probability of its occurrence is usually recorded as α;

2.When the null hypothesis is not true, it is decided not to abandon the null hypothesis, which is called Type II error, and the probability of its occurrence is usually recorded as β.

We use the value of the significance test to judge whether it conforms to the H0 test(The value of P-value is greater than 0.05). So as to understand the correlation of columns.

Before detection, we can observe the distribution of Credit amount and the Duration. So as to preliminarily judge the relationship between Credit amount and the Duration.

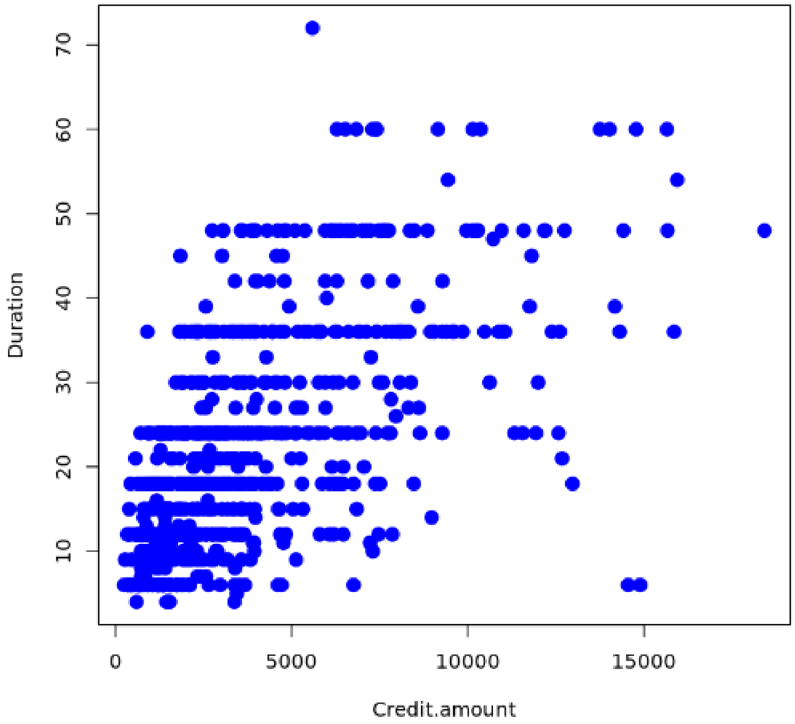


Figure 7 –Credit amount and the Duration

2.2.1 T-test

The significance test is to make an assumption about the parameters of the population (random variable) or the overall distribution form in advance, and then use the sample information to judge whether the hypothesis (alternative hypothesis) is reasonable, that is, to judge the true situation of the population and the original. Hypothetically whether there is a significant difference.The significance test is to determine whether the difference between the sample and the assumptions we make about the population is purely chance variation, or is caused by the inconsistency between the assumptions we make and the truth of the population. The significance test is to test the hypothesis we make about the population.

The T-test (or Student's test) solves the problem of proving the presence of differences in the average values of a numeric variable in the case when there are only two compared groups

We use the t.test function in the r language to judge Credit.amount and Duration. The p-value is less than 2.2e-16. alternative hypothesis: true difference in means is not equal to 0.And 95% confidence intervalis 3075.189 to 3425.521.

The formula for T test is

$$t = \frac{\overline{X_A}-\overline{X_B}}{SE(\overline{X_A}-\overline{X_B})}$$

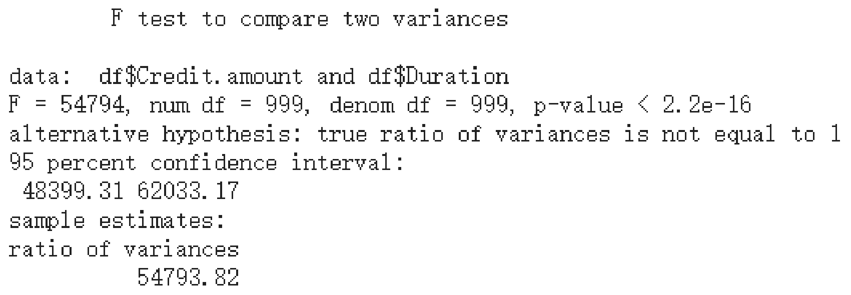
$\overline{X_A} - \overline{X_B}$ is the difference between the two sample means

$SE(\overline{X_A} - \overline{X_B})$ is the standard deviation of the difference between the two sample means.

Beacause the p-value of the test is less than the significance level alpha = 0.05. We can then reject the null hypothesis, Credit amount and the Duration is not significantly different.

2.2.2 F-test

F test is the most commonly used alias is called joint hypothesis test (, also known as variance ratio test, variance homogeneity test. It is a test under the null hypothesis that the statistical value obeys the F-distribution. It is usually used to analyze A statistical model with more than one parameter is used to determine whether all or some of the parameters in the model are suitable for estimating the population.



```
        F test to compare two variances

data:  df$Credit.amount and df$Duration
F = 54794, num df = 999, denom df = 999, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 48399.31 62033.17
sample estimates:
ratio of variances 
          54793.82 
```

Figure 8 – result of F test

Finally, p-value < 2.2e-16 is obtained, the null hypothesis is not accepted, and the two variances are considered to be different.

2.2.3 Correlation analysis

The correlation coefficient is a statistical indicator first designed by statistician Carl Pearson. It is a measure of the degree of linear correlation between variables. It is generally represented by the letter r. Due to different research objects, there are many ways to define the correlation coefficient, and the most commonly used is the Pearson correlation coefficient.

Correlation is the tendency of one investment target to influence the trend of another investment target. Simply put, it is whether the two investment products rise and fall in synchronization. If they rise and fall at the same time, they have a high correlation, and repeated allocations cannot diversify risks.

Correlation is divided into three situations: one is positive correlation, if the change of one investment product is in step with the change of another product, then the two are positively correlated, and the correlation coefficient is 1. The second is negative correlation. If two investment products move in opposite directions and move in different directions, then the two are negatively correlated. The third is irrelevance. If the changes of the two investment products are independent and unrelated to each other, it is said that the two are not related.

After verifying that there is a moderate correlation between Credit amount and the Duration and there is no significant difference, I constructed a simple linear regression analysis on the two variables, and calculated the regression equation of the two variables.

As a method widely used in the field of machine learning. We can observe the correlation between Credit amount and the Duration through the heatmap.

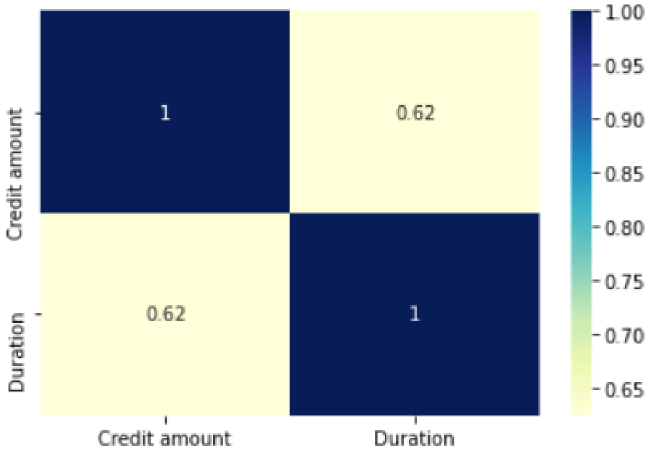


Figure 9 –Heatmap of Credict amount and Duration

It can be seen from the picture and the previous conclusion, that the two variables are moderately correlated, and the correlation is about 0.62. As daily data, it is very suitable for linear regression models to predict.

After calculating the correlation, and after observing the situation of the picture. It can be concluded that a linear regression model can be used, and for the data set, it is also very suitable for us to analyze.

To ensure better interactivity with subsequent files, I used jupyternook to build a simple model to analyze the model equations. Compared with the r files compiled by rstudio, notebooks can be read, written and modified on the pycharm compiler better, which is convenient for subsequent maintenance and updates.

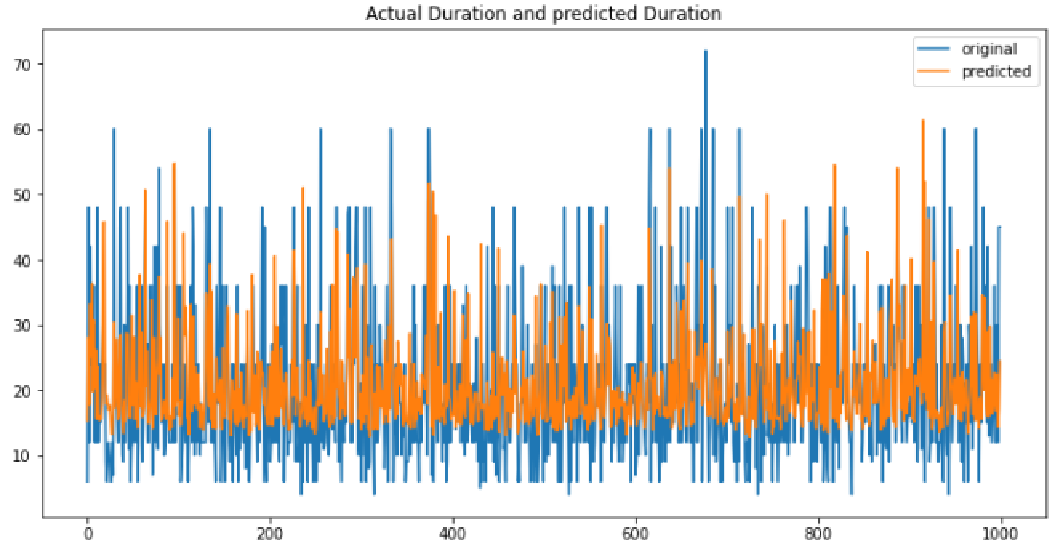


Figure 10 –Actual Duration and predict Duration

2.2.4 Model derivation

Through the previous detection and analysis, it can be seen that the data set can use the linear regression model. In order to better understand the model, we make a simple linear regression model.

After make the linear_model , the Intercept and Slope functions of the model are called, and the slope and intercept of the equation are obtained as 0.00266995, 12.168902198241387.

It can be seen that linear regression equation is y=0.00266995*x+ 12.168902198241387.

When building the linear regression model, we used the method of least squares.The least squares method, also known as the least squares method, is a mathematical optimization modeling method. It finds the best functional match for the data by minimizing the sum of squared errors.

Results of fitting a set of data points using a quadratic function

The unknown data can be easily obtained by using the least squares method, and the sum of squares of errors between the obtained data and the actual data can be minimized. The formula for the least squares method is

$$\sum_{i=1}^{n} (y_i - \overline{y_i})^2$$

The "least squares method" is a standard method for obtaining approximate solutions by regression analysis for a system of linear equations, that is, a system of equations with more equations than unknowns. In this whole solution, the least squares method is calculated to minimize the sum of the residual sums of squares in the result of each equation.

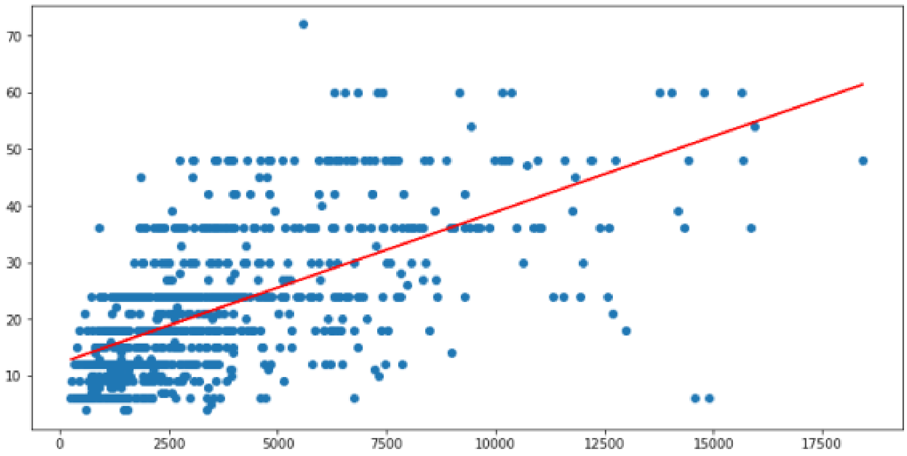


Figure 11 –Linear Regression Equation with True Values

By observing the values and the formula we derived, our dataset can be predicted using a linear regression model. In the next part, the main thing is to write the code, implement the model and predict.

3.Model implementation

In the implementation model part, we use the framework of flask. flask is a popular web framework. Its WSGI toolkit uses Werkzeug and its template engine uses Jinja2. Flask uses BSD authorization. Flask is a very popular web framework that uses the Python programming language to implement related functions.

Flask is a very popular web framework that uses the Python programming language to implement related functions. It is called microframework, and users can choose various databases according to their needs. Flask itself does not provide form verification function, and can be freely configured during project implementation, thereby providing database abstraction layer basic components for application development, and supporting functions such as form data validity verification, file upload processing, user authentication, and database integration.

Flask mainly includes two core function libraries, Werkzeug and Jinja2, which are responsible for business processing and security functions respectively. These basic functions provide rich basic components for the development process of web projects. The Werkzeug library is very powerful and has relatively complete functions. It supports URL routing request integration, and can respond to access requests from multiple users at one time; supports cookie and session management, establishes long-term connection relationships through identity cache data, and improves user access speed; supports interactive Javascript Debugging to improve user experience; it can handle basic HTTP transactions and quickly respond to access requests pushed by clients. Jinja2 library supports automatic HTML transfer function, which can well control script attacks by external hackers. The system runs very fast, and the page loading process will compile the source code into Python bytecode, so as to realize the efficient operation of the template; the template inheritance mechanism can modify and maintain the template content, and provide corresponding templates for users with different needs.

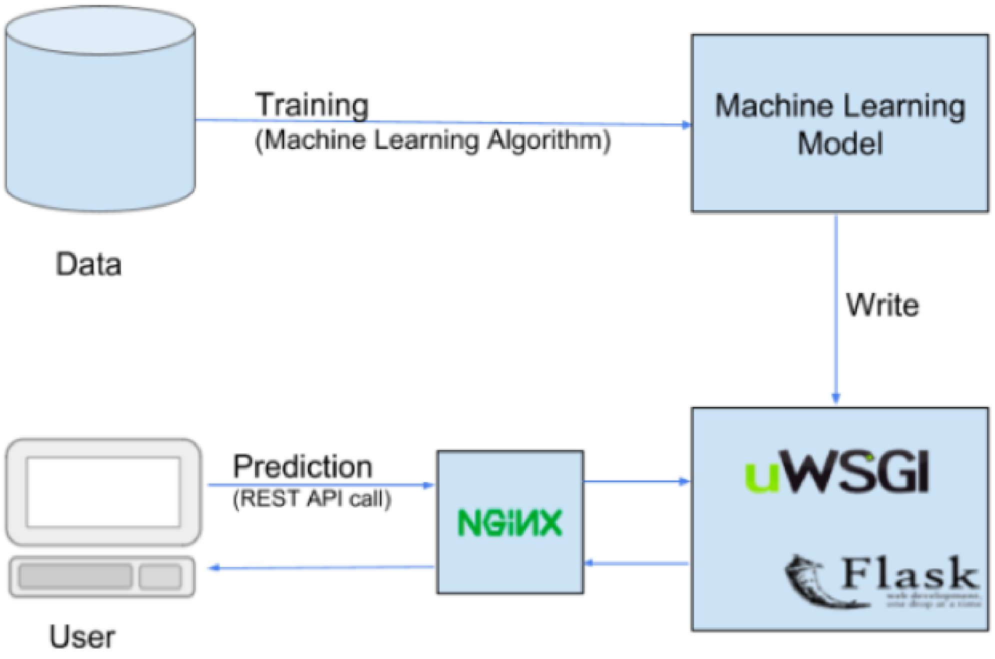


Figure 12– Model deployment process

Deploying a machine learning model by using the API is essentially to decouple the model layer and the application layer. The usual practice is to serialize and save the model training results at the model layer, and directly call the model file at the application layer for prediction.

3.1Model layer

First reading the processed csv file.Then splitting Train and Test Set.Then train model and predict by the model.Finally save the model and save it as 'lr.pkl'.

Model performance is not runtime or execution performance, but how accurate the model is at making predictions. For different datasets, we can choose several models for comparison to obtain the final model. In our experiments, we concluded that linear regression was the best model relative to our chosen dataset, and the predictions from the model yielded more accurate results.

3.2Application layer

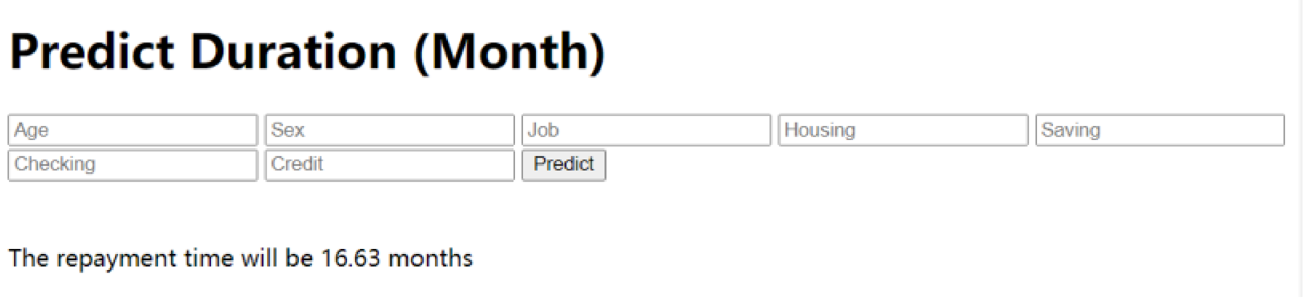
The application layer is divided into two parts of code files: app.py involves the back-end interface, and index.html involves the front-end page display.The role of the application layer is to convert the input high-level language into data, and its main task is to complete specific applications through applications between applications.

In the index.html file, by designing the html file, add the feature values required by the prediction model and the Duration to be predicted in the body.

In the app.py file, by reading the model. Redirecting the API to the home page index.html. Then Redirecting the API to predict the result (Duration). Finally Starting the flask server.

3.3Result

By running the file, then go to host http://127.0.0.1:5000. Enter the known eigenvalues, and finally get the repayment time we predict.



Predict Duration (Month)

The repayment time will be 16.63 months

Figure 13– prediction interface

After getting the results, I gave further thought to the progress of the trial process. In order to be able to be used. We need to read data from the database. The ability to collect, process, and analyze large amounts of data, enabling good inferences about new knowledge in business and areas of scientific knowledge.When the value of the data is relatively accurate and the amount of data is not large,We can use relational databases such as mysql to store data. When there are many databases, we can store data through mongodb or HDFS.

Then, we also need to design a py file to regularly read data from the database to train the linear regression model and count the accuracy of the previous model.

Finally, in the flask part, because the experiments are mainly carried out in the local environment, we can also purchase a server and deploy flask to the server, so that users can better use the linear regression model to predict.

Conclusion

Through the comparison of several regression models, it can be seen that the linear regression model is very easy to understand, and the results have good interpretability, which is conducive to decision analysis. For data with strong correlation coefficient, I can draw a very clear conclusion. At the same time, through the study of linear regression equation and least squares method, I have a deeper understanding of the previous data analysis methods.

In the experimental part, through the hierarchical design of different modules, the web page about the repayment month prediction is also obtained. I hope that in the future, models such as linear regression can be better used in finance and other fields.

References

1. Девятых Д.В., Гергет О.М., Берестнева О.Г. НЕЙРОДИНАМИЧЕСКАЯ ДИАГНОСТИКА НАРУШЕНИЙ ДЫХАНИЯ ВО ВРЕМЯ СНА. Прикаспийский журнал: управление и высокие технологии. 2014. № 4 (28). С. 144-156

2.R. P. Bunker, M . A. Naeem, w . Zhang, .Improving a Credit Scoring M ode1 by Incorporating Bank Statement Derived Features", 0ctober 2016.

3.F. Louzada, A. A. Gui1herme, B. Fernandes, "C1assicication methods app1ied to credit scoring: Systematic review and overa1 1 comparison", February 2016. -8

4.J. Hariharakrishnan, S. Mohanava1 1i, Srividya, K.B. Sundhara Kumar, "Survey oc Pre-processing Techniques cor M ining Big Data", Internationa1 Concerence on Computer, Communication and Signa1 Processing (ICCCSP), 2017.

5.A. Blanco, R. Pino-Mejías, J. Lara, S. Rayo, "Credit scoring models for the microfinance industry using neural networks

4.FINANCIAL MANAGEMENT

Introduction

The main purpose of this section is to evaluate the development prospects and plan the financial and commercial value of the final product presented as part of the research work. Commercial value is determined not only by the presence of higher technical characteristics over competitive developments, but also by how quickly the developer can answer the following questions - will the product be in demand on the market, what will be its price, what is the budget for scientific research, how long will it take to promote developed product to market.

This section provides for consideration of the following tasks:

- Evaluation of the commercial potential of the development.
- Planning of research work;
- Calculation of the research budget;
- Determination of resource, financial, budgetary efficiency of the study.

4.1.Competitiveness analysis

To analyze the consumers of the research results, it is necessary to consider the target market and carry out its segmentation. The analysis of competitive technical solutions from the standpoint of resource efficiency and resource saving allows us to assess the comparative effectiveness of scientific development and determine directions for its future improvement.

Target Market - Since my dissertation mainly uses computers, the target market is some internet companies and banks.

Since the article is mainly about Internet companies, we analyze the amount of resources and types of resources of each Internet company.

	Type of Internet resource			
	rporate website	Internet catalog	Online store	Informational portal

Company size	Large	▨		▒	
	Medium	▥			▥
	small		▨	▨	▒

Table 6. Segmentation map of the development services market

Internet resources:

▨ Firm A ▥ Firm B ▒ Firm C

This example segmentation map shows which niches in the Internet resource development services market are not occupied by competitors or where the level of competition is low. As a rule, two or three segments are chosen, to which the maximum efforts and resources of the enterprise are directed. As a rule, select segments with similar characteristics that will form the target market.

The result of segmentation should be:
selection of the segments on which the enterprise intends to focus;
1. determination of the main segments of this market;
2.selection of the segments on which the enterprise intends to focus;
identifying market segments that are attractive for the enterprise in the future.

4.1.1Analysis of competitive technical solutions from the position

Since the market is constantly changing, it is necessary to systematically conduct a detailed analysis of competitive developments present on the market. Such analysis helps to make adjustments to scientific research in order to successfully compete with their competitors. It is important to realistically evaluate the strengths and weaknesses of competitors' developments. To this end, the method used in the article is compared with the traditional method.

The use of scorecards makes it easy to analyze competing technology solutions from a resource efficiency and resource conservation perspective. This is necessary to assess the comparative validity of scientific developments and to identify directions for their future improvements.

According to the selected comparison object, comprehensively consider the technical and economic characteristics of its development, creation and operation, and select the criteria for comparative evaluation of resource efficiency and resource conservation.The most competitive developments for autonomous resonant inverters (F) are: machine learning methods (K1), traditional analytical methods (K2).

Analysis of competitive technical solutions is determined by the formula:

$$K = \sum B_i \times \text{Б}_i \tag{6.1}$$

where K is the competitiveness of a scientific development or a competitor;

B_i – is the weight of the indicator ;

Б_i – score of the i-th indicator.

Table 7. Evaluation card for comparison of competitive technical solutions

Evaluation criteria ***example***	**Criterion weight**	**Points**			**Competitiveness** Taking into account weight coefficients		
		P_f	P_{i1}	P_{i2}	C_f	C_{i1}	C_{i2}
1	2	3	4	5	6	7	8
Technical criteria for evaluating resource efficiency							
1. Energy efficiency	0,2	5	3	4	0,5	0,3	0,4
2. Reliability	0,15	5	5	3	0,7	0,7	0,42
3. Safety	0.25	4	5	3	0,5	0,3	0,3
4. Functional capacity	0.1	5	3	3	0,5	0,3	0,3
Economic criteria for performance evaluation							
1. Development cost	0,12	4	5	3	0,48	0,6	0,36
2. Market penetration rate	0,1	4	5	3	0,4	0,5	0,3
3. Expected lifecycle	0,08	5	4	4	0,4	0,32	0,32
Total	**1**	32	34	28	3,48	3,02	2,4

4.1.2 Machine learning methods

Artificial intelligence and machine learning are ubiquitous in almost every industry, from businesses doing business to online gaming. The impact of artificial intelligence on the business world may be greater than the impact on people's daily lives.

Tech giants like Meta and Google have invested significant techniques in their products. This is just the beginning, in the next few years, people will see how artificial intelligence and machine learning will gradually infiltrate various industries.

Regardless of the size and type of industry, customer service will always be an important part of any business. Machine learning is expected to completely change customer service or support in the next few years. AI-enabled systems are expected to have sentiment analysis technology to help respond better to customer concerns.

4.1.3 Traditional analytical methods

With the traditional analysis method, mainly by hiring a large number of employees, excel analysis is used. At the same time, compared with machine learning, excel has the following disadvantages.

1.Excel provides limited security, it can only restrict user access and modify rights, but cannot manage roles for users and cannot restrict access to data at the row level.

2.At the same time, Excel's cross-platform is relatively bad, because Excel can only cross two platforms of PC and Jmac, and most database products can be run on any platform by installing the client, but if you use Excel.

Therefore, the projects under development have many unique characteristics and are low cost, so they are promising.

4.2.SWOT analysis

Complex analysis solution with the greatest competitiveness is carried out with the method of the SWOT analysis: Strengths, Weaknesses, Opportunities and Threats. The analysis has several stages. The first stage consists of describing the strengths and weaknesses of the project, identifying opportunities and threats to the project that have emerged or may appear in its external environment. The second stage consists of identifying the compatibility of the strengths and weaknesses of the project with the external environmental conditions. This compatibility or incompatibility should help to identify what strategic changes are needed.

Table 8.-SWOT analysis

	Strengths: S1. can save time S2. save money S3. adapt the development of the times	**Weaknesses:** W1. professional staff required W2. requires regular maintenance W3. user privacy legality
Opportunities: O1. get more users O2. know the user ahead of time O3. Advantages when competing with peers	Strategy which based on strengths and opportunities: Get more high-quality customers in a short period of time	Strategy which based on weaknesses and opportunities: After the development is completed, a large number of customers can be obtained
Threats: T1.user does not cooperate T2. peer competition	Strategy which based on strengths and threats: Better algorithms need to be developed	Strategy which based on weaknesses and threats: Better use of basic properties

4.2.1Project Initiation

The initiation process group consists of processes that are performed to define a new project or a new phase of an existing one. In the initiation processes, the initial purpose and content are determined and the initial financial resources are fixed. The internal and external stakeholders of the project who will interact and influence the overall result of the research project are determined.

Table 9. Stakeholders of the project

Project stakeholders	Stakeholder expectations
Internet company	Prediction of repayment time by linear regression model

Table 10. Purpose and results of the project

Purpose of project:	Get the best model for predicting the repayment time is obtained

Expected results of the project:	Achieving more than 80% accuracy
Criteria for acceptance of the project result:	The accuracy rate exceeds 80%
Requirements for the project result:	Can accurately judge customers for the company
Requirements for the project result:	Get more users
	Get more users
	Reduce costs for businesses
	Good assessment of the quality of users

4.2.2The organizational structure of the project

It is necessary to solve the some questions: who will be part of the working group of this project, determine the role of each participant in this project, and prescribe the functions of the participants and their number of labor hours in the project.

Table 11. Structure of the project

№	Participant	Role in the project	Functions	Labor time, hours
1	professor	Provide the framework of the overall model, and run ideas	Experiment with layouts	272hours
2	student	Train machine learning models and draw relevant conclusions	completed experiment	224hours

4.2.3 Project limitations

Project limitations are all factors that can be as a restriction on the degree of freedom of the project team members.

Table 12. Project limitations

Factors	**Limitations / Assumptions**
3.1. Project's budget	300000 руб
3.1.1. Source of financing	Internet company
3.2. Project timeline:	2months
3.2.1. Date of approval of plan of project	2022.3.20
3.2.2. Completion date	2022.5.20

4.2.4 Project Schedule

As part of planning a science project, you need to build a project timeline and a Gantt Chart.

Table 13. Project Schedule

Job title	Duration, working days	Start date	Date of completion	Participants
General Technical supervision	8 days	2022.3.20	2022.3.30	Supervisor
Research and analysis of literature	7 days	2022.3.31	2022.4.10	Supervisor
Clean data and build machine learning models	15 days	2022.4.11	2022.4.30	Supervisor / Student

Build a credit score card based on the machine learning model	4 days	2022.5.1	2022.5.11	Supervisor / Student
Preparing of dissertation	7 days	2022.5.12	2022.5.20	Student

4.3.Gantt chart and budget of scientific research

A Gantt chart, or harmonogram, is a type of bar chart that illustrates a project schedule. This chart lists the tasks to be performed on the vertical axis, and time intervals on the horizontal axis. The width of the horizontal bars in the graph shows the duration of each activity.

Table 14. A Gantt chart

№	Activities	Participants	T_c, days	Duration of the project											
				February			March			April			May		
				1	2	3	1	2	3	1	2	3	1	2	3
1	Drafting and approval of terms of reference			▨											
2	Selection and study of materials on the topic			▨ ■											
3	Conducting patent research			▨											
4	Choice of research direction				■	■									
5	Scheduling work on the					■	■								

	topic														
6	Building layouts (models) and conducting experiments							■	■						
7	Development of a flowchar								■	■	■				
8	Program development											▨			
9	Drawing up an explanatory note											■	■	■	■

4.3.1Scientific and technical research budget

The amount of costs associated with the implementation of this work is the basis for the formation of the project budget. This budget will be presented as the lower limit of project costs when forming a contract with the customer.

To form the final cost value, all calculated costs for individual items related to the manager and the student are summed.

In the process of budgeting, the following grouping of costs by items is used:

-Material costs of scientific and technical research;

-costs of special equipment for scientific work (Depreciation of equipment used for design);

-basic salary;

-additional salary;

-labor tax;

-overhead.

4.3.2 Calculation of material costs

The calculation of material costs is carried out according to the formula:

$$C_m = (1 + k_T) \cdot \sum_{i=1}^{m} P_i \cdot N_{consi}$$

where m – the number of types of material resources consumed in the performance of scientific research;

N_{consi} – the amount of material resources of the i-th species planned to be used when performing scientific research (units, kg, m, m^2, etc.);

P_i – the acquisition price of a unit of the i-th type of material resources consumed (rub./units, rub./kg, rub./m, rub./m^2, etc.);

k_T – coefficient taking into account transportation costs.

Prices for material resources can be set according to data posted on relevant websites on the Internet by manufacturers (or supplier organizations).

Table 15. Material costs

Name	Unit	Amount	Price per unit, rub.	Material costs, rub.
Papers	1	100	100	100
Software copyright fee	1	1	1900	1900
Total				2000

4.3.3 Costs of special equipment

This point includes the costs associated with the acquirement of special equipment (instruments, stands, devices and mechanisms) necessary to carry out work on a specific topic.

Table 16 a. Costs of special equipment (+software)

№	equipment identification	Quantity of equipment	Price per unit, rub.	Total cost of equipment, rub.

1.	cpu	1	10000	10000

4.4Assessment of resource, financial and economic efficiency of STR

4.4.1Basic salary

This point includes the basic salary of participants directly involved in the implementation of work on this research. The value of salary costs is determined based on the labor intensity of the work performed and the current salary system

The basic salary (S_b) is calculated according to the formula:

$$S_b = S_a \cdot T_w ,$$

where Sb – basic salary per participant;

T_w – the duration of the work performed by the scientific and technical worker, working days;

S_d - the average daily salary of an participant, rub.

The average daily salary is calculated by the formula:

$$S_d = \frac{S_m \cdot M}{F_v} ,$$

Where S_m – monthly salary of an participant, rub .;

M – the number of months of work without leave during the year: at holiday in 48 days, M = 11.2 months, 6 day per week;

F_v – valid annual fund of working time of scientific and technical personnel (251 days).

Table 17. The valid annual fund of working time

Working time indicators	
Calendar number of days	365
The number of non-working days	
- weekend	52
- holidays	14
Loss of working time	48

- vacation - isolation period - sick absence	
The valid annual fund of working time	251

Monthly salary is calculated by formula:

$$S_{month} = S_{base} \cdot (k_{premium} + k_{bonus}) \cdot k_{reg}\,, \qquad (x)$$

where S_{base} – base salary, rubles;

$k_{premium}$ – premium rate;

k_{bonus} – bonus rate;

k_{reg} – regional rate.

Table 18. Calculation of the base salaries

Performers	S_{base}, rubles	$k_{premium}$	k_{bonus}	k_{reg}	S_{month}, rub.	W_d, rub.	T_p, work days (from table 7)	W_{base}, rub.
Supervisor	37700			1,3	49010	1633.7	34	58470
Student	19200				24960	832	28	23296

4.4.2Additional salary

This point includes the amount of payments stipulated by the legislation on labor, for example, payment of regular and additional holidays; payment of time associated with state and public duties; payment for work experience, etc.

Additional salaries are calculated on the basis of 10-15% of the base salary of workers:

$$W_{add} = k_{extra} \cdot W_{base}\,, \qquad (x)$$

where W_{add} – additional salary, rubles;

k_{extra} – additional salary coefficient (10%);

W_{base} – base salary, rubles.

4.4.3Labor tax

Tax to extra-budgetary funds are compulsory according to the norms established by the legislation of the Russian Federation to the state social insurance (SIF), pension fund (PF) and medical insurance (FCMIF) from the costs of workers.

Payment to extra-budgetary funds is determined of the formula:

$$P_{social} = k_b \cdot (W_{base} + W_{add}) \qquad \text{(x)}$$

where k_b – coefficient of deductions for labor tax.

In accordance with the Federal law of July 24, 2009 No. 212-FL, the amount of insurance contributions is set at 30%. Institutions conducting educational and scientific activities have rate - 27.1%.

Table 19. Labor tax

	Project leader	**Engineer**
Coefficient of deductions	27.1%	
Salary (basic and additional), rubles	58470	23296
Labor tax, rubles	15845	6313

4.4.4 Overhead costs

Overhead costs include other management and maintenance costs that can be allocated directly to the project. In addition, this includes expenses for the maintenance, operation and repair of equipment, production tools and equipment, buildings, structures, etc.

Overhead costs account from 30% to 90% of the amount of base and additional salary of employees.

Overhead is calculated according to the formula:

$$C_{ov} = k_{ov} \cdot (W_{base} + W_{add}) \qquad (x)$$

where kov – overhead rate.

Table 20. Overhead cost

	Project leader	Engineer
Overhead rate	30%	
Salary, rubles	58470	23296
Overhead, rubles	17541	6989

Formation of budget costs

The calculated cost of research is the basis for budgeting project costs.

Determining the budget for the scientific research is given in the table 15.

Table 21. Items expenses grouping

Name	Cost, rubles
1. Material costs	2000
2. Equipment costs	10000
3. Basic salary	81766
4. Additional salary	0
5. Labor tax	22158
6. Overhead	24530
7. Other direct costs	500
Total planned costs	**140954**

4.5Evaluation of the comparative effectiveness of the project

Determination of efficiency is based on the calculation of the integral indicator of the effectiveness of scientific research. Its finding is associated with the definition of two weighted average values: financial efficiency and resource efficiency.

The integral indicator of the financial efficiency of a scientific study is obtained in the course of estimating the budget for the costs of three (or more) variants of the execution of a scientific study. For this, the largest integral indicator of the implementation of the technical problem is taken as the calculation base (as the denominator), with which the financial values for all the options are correlated.

The integral financial measure of development is defined as:

$$I_f^d = \frac{C_i}{C_{max}} \quad \text{(x)}$$

,

where I_f^d – integral financial measure of development;

C_i – the cost of the i-th version;

$C_{\max}$ – the maximum cost of execution of a research project (including analogues).

The obtained value of the integral financial measure of development reflects the corresponding numerical increase in the budget of development costs in times (the value is greater than one), or the corresponding numerical reduction in the cost of development in times (the value is less than one, but greater than zero).

Since the development has one performance, then I_f^d = 1.

The integral indicator of the resource efficiency of the variants of the research object can be determined as follows:

$$I_m^a = \sum_{i=1}^{n} a_i b_i^a \qquad I_m^p = \sum_{i=1}^{n} a_i b_i^p$$

where I_m – integral indicator of resource efficiency for the i-th version of the development;

a_i – the weighting factor of the i-th version of the development;

b_i^a , b_i^p – score rating of the i-th version of the development, is established by an expert on the selected rating scale;

n – number of comparison parameters.

The calculation of the integral indicator of resource efficiency is presented in the form of table 17.

Table 22 – Evaluation of the performance of the project

Criteria	Weight criterion	Points
1. Energy efficiency	0.12	13
2. Reliability	0.20	12
3. Safety	0.10	13
4. Functional capacity	0.18	13
Economic criteria for performance evaluation		
1. Development cost	0.14	12
2. Market penetration rate	0.18	13
3. Expected lifecycle	0.08	9
Total	1	85

Thus, the effectiveness of the development is presented in table 18.

Table 23 – Efficiency of development

№	Indicators	Points
1	Integral financial measure of development	12
2	Integral indicator of resource efficiency of development	15
3	Integral indicator of the development efficiency	13

Comparison of the values of integral performance indicators allows us to understand and choose a more effective solution to the technical problem from the standpoint of financial and resource efficiency.

Conclusion

Computational results show that the machine learning approach is competitive in the market and not inferior to listed competitors in terms of performance.

Therefore, it can be concluded that the development is sufficiently resource efficient.

On the basis of the analysis and research carried out, it is possible to draw a selection regarding the implementation of the product market development, the market segment for the development implementation.

5.SOCIAL RESPONSIBILITY

The impact of hazardous and harmful production factors on a person can be weakened or eliminated by the normal organization of workplaces, improvement of technological processes, the use of collective and (or) individual protective equipment.

When performing work, most of the time is done in a computer classroom. It is mainly done with a laptop and edited with pycharm.

5.1Legal and organizational issues of occupational safety

In my paper, I mainly predict the repayment time based on the loan amount, age and other information, mainly using the linear regression model and flask framework in machine learning to predict.Loan granting is still one of the main services provided by banks. Therefore, credit risk management is crucial for banking and its importance is increasing over time. Credit risk appears when a wrong decision on loan granting is taken. A large number of such a decisions leads to increase in a number of defaulters causing bankruptcy.

The content of the experiment is a classroom with a computer, and the research method is mainly machine learning. The working equipment is mainly a computer.

5.1.1 Analysis of working conditions in the workplace

In accordance with the characteristics of the working area and GOST 12.0.003-2015 “Hazardous and harmful production factors. Classification”, a list of harmful and dangerous factors for the employee involved in the development of documents has been compiled. The list of factors and documents regulating them is presented in Table 1.

Table 24 - Identified harmful factors

Harmful factors	Regulations
Exceeding the noise level	GOST 12.1.003-2014 SSBT. Noise. General safety

	requirements, GOST 30691 (ISO 4871:1996) Machine noise. Declaration and control of noise performance values)
Insufficient illumination of the working area	artificial lighting. Updated edition of SNiP 23-05-95*;
Temperature in the work area	SP 60.13330.2020 Heating, ventilation and air conditioning SNiP 41-01-2003
Physical overload	GOST 9241-4-2009 GOST 9241-4-2007

5.1.2Analysis of Noise level analysis

Sudden high-intensity noises, even short-term ones (explosions, impacts, etc.), can cause both acute neurosensory effects (dizziness, tinnitus, hearing loss) and physical damage (rupture of the eardrum with bleeding, damage to the middle ear and snails).

The permissible noise level is limited by GOST 12.1.003-83 and SanPiN 2.2.4 / 2.1.8.10-32-2002. The maximum sound level of constant noise at workplaces does not exceed 80 dBA. The maximum sound level of continuous noise in our workplace is up to 75 dBA, the main noise comes from the production and processing of the lathe. GOST 12.1.003-2014 SSBT. Noise. General safety requirements, (GOST 30691 (ISO 4871:1996) Noise of machines. Statement and control of noise performance values.

Noise from aerodynamic noise sources can be reduced by using anti-vibration pads installed between the base of the machine, the instrument and the supporting surface. Rubber, felt, cork, shock absorbers are used as gaskets. Add protective equipment (I added these green parts of the protective equipment).

5.1.3Analysis of work area lighting

If the lighting is insufficient, it will cause great harm to the staff's eyesight and mental health. I know this very well.

The rooms have both natural and artificial lighting. Due to the fact that the work of the operator for servicing the radio control panel corresponds to the category of visual work III b, the following requirements for the workplace, my workplace, should be observed: E = 300 lux.

Natural lighting is provided through light openings that provide the required natural light coefficient (KEO) of at least 1.2%.As light sources in artificial lighting, fluorescent lamps of the LB type are mainly used.

5.1.4 Analysis of In-lab temperature

This set of rules applies to the design of systems for internal heat and cold supply, heating, ventilation and air conditioning in buildings under construction, reconstructed or overhauled, public buildings with a height of not more than 50 m and residential buildings with a height of not more than 75 m, including multifunctional buildings and buildings single function, our small factory is in this range.

Working area: A space of a certain volume in a room in which people are provided and requirements for the parameters of the air environment are set.SP 60.13330.2020 Heating, ventilation and air conditioning SNiP 41-01-2003 (as amended).

Direct Cooling Scheme: A refrigeration scheme in which the air of the conditioned room is cooled in a heat exchanger by the working medium (refrigerant) of the refrigeration machine, this solution is used in our small factory workshop.

Operated (working) area: A space of a certain volume in a room in which people are provided and requirements for the parameters of the air environment are set.

5.1.5Analysis of Mental Health Factors

If students use the computer for a long time to do experiments, it will cause physical overload and emotional overload.According to the relevant regulations of GOST 9241-4-2009, we can make the seating and layout of the office more reasonable and reduce the risk of physical injury to students.

At the same time, according to GOST 9241-4-2007 ergonomic requirements for the use of visual display terminals (VDT) for office work, we should also limit the working hours of students to prevent excessive working hours and harm to mental health.

5.2 Environment safety

The proper disposal or recycling of hazardous computer components is a global issue. Most computers and peripherals use and contain at least some materials that can be considered toxic to the environment. Because the experiment mainly uses computers, this part mainly analyzes the treatment of electronic waste and microclimate indicators.

The safety of loading and unloading operations must be ensured by: the choice of methods for the production of work, lifting and transport equipment and technological equipment; the preparation and organization of work sites; the use of protective equipment for workers; the conduct of a medical examination of persons admitted to work and their training.

5.2.1Analysis of Electronic wastes

Electronic wastes generated from any modern establishments. They may be described as discarded electrical or electronic devices. Some electronic scrap components, such as CRTs, may contain contaminants such as Pb, Cd, Be or brominated flame retardants.

Non-hazardous/solid waste is all waste which has not been classified as hazardous: paper, plastics, glass, metal and beverage cans, organic waste etc. While not hazardous, solid waste can have serious environmental and health impact if left uncollected and untreated.

The recycling of electronic waste has become the focus of attention in recent years, and how to properly dispose of electronic waste has become an important part of global environmental protection.

The recovery process of metals from electronic waste is relatively complicated. Usually, metals and impurities are separated by high temperature, and then various metals are extracted through several corresponding processing processes. Precious metals such as copper, gold, silver, platinum, and palladium in electronic waste are generally recovered by converter processing. Through consulting the information, we learned that the current mainstream methods of processing electronic waste are mainly melting and refining.

5.2.2Analysis of Batteries from portable computer systems

According to GOST R 52105-2003. Resources saving. Waste treatment.Classification and treatment methods of the mercury containing waste. Basic principles.

In the recycling process of computer batteries, the quality of the electronic components in stock (such as the appearance, labels, outer packaging, etc.) of the electronic components is first tested, and the electronic components in the factory that have been tested can be sold for the second time. Other types are made of metal raw materials to prevent environmental pollution.

5.2.3Analysis of Hazardous Pollutants

Hazardous waste is waste that has been identified as potentially causing harm to the environment and human health and therefore needs special, separate treatment and handling. Chemical and physical characteristics determine the exact collection and recycling process. Flammability, corrosiveness, toxicity, ecotoxicity and explosiveness are the main characteristics of hazardous waste. Liquid, gaseous and powder waste need special treatment by default to avoid the dispersal of the waste.

Because all living things need water to survive. Water pollution has increased significantly over the decades and has now developed into a serious worldwide problem. The presence and persistence of harmful pollutants such as dyes, pharmaceuticals and personal care products, heavy metals, fertilizers and pesticides and their transformation products are serious environmental and health concerns. Various methods have been tried to purify and maintain water quality. To prevent dangerous pollutants from entering the water and atmosphere, we need to control this.

5.2.4Analysis of Waste Electronic Components Recycling

Almost 99% of the components of a computer can be recycled. Recycling can avoid serious toxins, chemicals and heavy metals from going to landfills and polluting the environment. Being a significant part of waste electrical and electronic equipment (WEEE), computer waste is gaining more attention due to its tremendous generation and toxic environmental concerns.

The complex and diverse material content in the computer makes them ideal for recycling. At the same time, presence of hazardous contents such as flame retardants and heavy metals makes obstructions in recycling procedure. The recycling method will vary for each material and component of e-waste.

Electronic waste is the fastest growing solid waste and one of the most difficult to dispose of solid waste. At the same time, it contains huge social wealth and resources. Solving the problems of its disposal and recycling is necessary for the sustainable development of Shehu. , is one of the core scientific and technological issues in the "production, consumption, scrap" industrial chain. Solving the problem of waste at the source is of reference and demonstration significance for all electronic waste problems including waste electronic products. Electronic waste contains a large amount of non-ferrous metals such as copper, aluminum, lead, and zinc, and precious metals such as gold and silver. Electronic waste is called "urban mines", and its development cost is much lower than that of primary resources in mines, and the energy required for development is less than that of primary resources in mines. The shortage of non-ferrous metal resources and serious environmental pollution are of great help to human development.

5.3.Emergency safety

An mergency poses an immediate risk of significant harm to health, life, property or the environment. Preparing for emergencies is an important part of workplace health and safety program.Because it is done in the laboratory, in the emergency safety part, I mainly analyze the fire and snow.

The proper disposal or recycling of hazardous computer components is a global issue. Most computers and peripherals use and contain at least some materials that can be considered toxic to the environment. This section

describes tools and procedures that help identify these materials and the steps for the proper handling and disposal of the materials.

Computers and peripherals contain materials that can be harmful to the environment. Hazardous materials are sometimes called toxic waste. These materials can contain high concentrations of heavy metals such as cadmium, lead, or mercury. There are regulations for the disposal of hazardous materials. Contact the local recycling or waste removal authorities for information about disposal procedures and services.

The category of the room for electrical safety, according to the PUE, is a room without increased danger, all electrical installations are used according to the requirement for the PUE.

On electrical safety, in accordance with the rules on labor protection during the operation of electrical installations, Employees are required to be trained in safe methods and techniques for performing work in electrical installations, and have protective tools.

According to SanPiN 1.2.3685-21, the workplace environment was determined, which ensures safety for workers.

According to SP 12.13130.2009, the categories of premises for explosion and fire hazard were determined, which is G moderate fire hazard due to the danger of gases and chips in the workpieces.

Determined the production of steel is a moderate negative impact on the environment, to objects of category II.

5.3.1Analysis of fire safety

According to GOST R 12.3.047-98 SSBT " Fire safety of technological processes. General requirements. Control methods" analysis and assessment of fire hazard of production facilities (technological processes) is carried out on the basis of their risk assessment.

Among all kinds of disasters, fire is one of the main disasters that most frequently and commonly threatens public safety and social development. Fire not only destroys material property, causes chaos in social order and destroys ecological balance, but also directly or indirectly endangers life.

The values of permissible fire hazard parameters should be such as to exclude the loss of life and limit the spread of the accident beyond the considered technological process to other facilities, including hazardous industries.

5.3.2Analysis of severe frost in winter

Snow disasters are less harmful than major meteorological disasters such as typhoons, floods, and droughts, and geological disasters such as earthquakes, but they cannot be ignored.

When a snow disaster occurs, it will mainly cause disasters in the following aspects: hinder the safety of lifeline projects such as transportation, communication, and transmission lines;

Although the experiment is indoors, it is mainly done in the computer classroom. If the severe frost in winter causes the short circuit of the wires, it may cause a great delay to the progress of the experiment.

Development of preventive measures to prevent emergencies;

- Vodokanal: it is necessary to ensure the supply of drinking and technical water to workers if it is not possible to interrupt the technological cycle of manufacturing parts. Also in the workshop it is recommended to have a supply of drinking water at the rate of 2 l / person. in shift.

- Heating main: provide room heaters powered by the electrical network, as well as PPE (warm clothes, gloves, hats).

-Electrical networks: should be provided with a generator (gasoline or diesel) that can produce the required power.

2) Technogenic - espionage, sabotage;

Emergencies resulting from sabotage are occurring more and more frequently. In most cases, such threats turn out to be false, however, work in this case still does not stop. Based on the Federal Law of July 22, 2008 N 123-FZ (as amended on April 30, 2021) "Technical Regulations on Fire Safety Requirements" https:/ /docs.cntd.ru/document/902111644.

5.3.3 Development of preventive measures to prevent emergencies

To prevent the possibility of sabotage, the enterprise must be equipped with a video surveillance system, round-the-clock security, access control system, reliable communication system, as well as exclude the dissemination of information about the security system of the facility, the location of premises and equipment in the premises, signaling devices, their installation locations and number. Officials conduct training every six months to formulate emergency procedures.

It is also necessary to provide emergency exits for personnel. The number of evacuation exits from the building on each floor is at least two. The width of the evacuation exit (gate) depends on the total number of people evacuated through the exit, but the width is not less than 0.8 m. The height of the passage on the evacuation routes is not less than 2 m.

An emergency poses an immediate risk of significant harm to health, life, property or the environment. Preparing for emergencies is an important part of workplace health and safety program.

For occupational and process safety as well as health and environmental protection and corporate security, comprehensive preventive measures must be taken. It is very important to analyze accidents, incidents and their causes in detail and make hazard analyses and the risk minimization measures.

All employers are obligated to ensure staff training emergency procedures in workplace. This may include what to do in case of a fire, earthquake, or other emergency; identifying locations of emergency exits;

And processes to follow to evacuate the building in the case of an emergency. These procedures are site specific and should be a part of the training for all new employees. In addition, regular drills or reviews of procedures are important to ensure that if an actual emergency occurs, everyone is able to react accordingly and safely.

Conclusion

This is an ethical principle, which consists in the fact that in order to implement a public duty in the decision-making process, it is necessary to take into account not only the interests of individuals or organizations making these decisions, but also the interests, values and goals of broad social groups and society as a whole.

Reference for social responsibility

1.GOST 12.1.003-2014 SSBT. Noise. General safety requirements,

2.GOST 30691 (ISO 4871:1996) Machine noise.

3.artificial lighting. Updated edition of SNiP 23-05-95*;

4.SP 60.13330.2020 Heating, ventilation and air conditioning SNiP 41-01-2003

5.GOST 9241-4-2009 and GOST 9241-4-2007

6.SanPiN 1.2.3685-21 and SP 12.13130.2009

7GOST R 12.3.047-98 SSBT "Fire safety of technological processes.Fire Safety Analysis

Applications

1.clean.py

```
import numpy as np
import pandas as pd
df = pd.read_csv('german_credit_data.csv')
sex_change = {'male':1, 'female':2}
Housing_change = {'free':1, 'rent':2,'own':3}
Saving_accounts_change = {'little':1,'moderate':2,'rich':3,'quite rich':3}
Checking_account_change = {'little':1,'moderate':2,'rich':3}

df.Sex = df.Sex.map(sex_change)
df.Housing = df.Housing.map(Housing_change)
df['Saving accounts'] = df['Saving accounts'].map(Saving_accounts_change)
df['Checking account'] = df['Checking account'].map(Saving_accounts_change)
df['Duration']=df['Duration']
del df['Unnamed: 0']
del df['Purpose']
# print(df)
df.to_csv('data1.csv',index=False)
```

2.compare.ipynb

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import train_test_split
df = pd.read_csv('data1.csv')
X = df.drop(['Duration'],axis=1).values
y = df['Duration'].values
X_train, X_test, y_train, y_test = train_test_split(X, y.astype('int'), test_size=0.3)
sns.displot(df['Duration'] , bins=30 , kde=True )
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestRegressor
def evaluate_model(y,X):
    r2= r2_score(y,y_pred)
    MAE = mean_absolute_error(y,y_pred)
    MSE = mean_squared_error(y,y_pred)
    print("r2 : %.4g" % r2)
    print("MAE : %.4g" % MAE)
print("MSE : %.4g" % MSE)
model = RandomForestRegressor(
        min_samples_split=100, min_samples_leaf=20, max_depth=8,
max_features='sqrt',
    random_state=10)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
evaluate_model(y_test, y_pred)
param_test1 = {'n_estimators': range(10, 71, 10)}
model = GridSearchCV(estimator=RandomForestRegressor(
    min_samples_split=100, min_samples_leaf=20, max_depth=8, max_features='sqrt',
    random_state=10), param_grid=param_test1, cv=10
  )
model.fit(X_train, y_train)
model.best_params_
{'n_estimators': 40}
param_test2 = {
    'max_depth': range(3, 14, 2),
    'min_samples_split': range(50, 201, 20)
}
model = GridSearchCV(estimator=RandomForestRegressor(
    n_estimators=40, min_samples_leaf=20, max_features='sqrt', oob_score=True,
    random_state=10), param_grid=param_test2, cv=10
  )
model.fit(X_train, y_train)
GridSearchCV(cv=10,
            estimator=RandomForestRegressor(max_features='sqrt',
                                            min_samples_leaf=20,
                                            n_estimators=40, oob_score=True,
                                            random_state=10),
            param_grid={'max_depth': range(3, 14, 2),
                        'min_samples_split': range(50, 201, 20)})


model.best_params_
'max_depth': 11, 'min_samples_split': 50}
param_test3 = {
    'min_samples_split': range(10, 90, 20),
    'min_samples_leaf': range(10, 60, 10),
}
model = GridSearchCV(estimator=RandomForestRegressor(
    n_estimators=40, max_depth=11, max_features='sqrt', oob_score=True,
    random_state=10), param_grid=param_test3, cv=10
  )
model.fit(X_train, y_train)
GridSearchCV(cv=10,
            estimator=RandomForestRegressor(max_depth=11, max_features='sqrt',
                                            n_estimators=40, oob_score=True,
                                            random_state=10),
          param_grid={'min_samples_leaf': range(10, 60, 10),
                    'min_samples_split': range(10, 90, 20)})
print(model.best_params_)
{'min_samples_leaf': 10, 'min_samples_split': 30}
param_test4 = {
      'max_features': range(3, 9, 2),
}
model = GridSearchCV(estimator=RandomForestRegressor(
      n_estimators=40, max_depth=11, min_samples_split=30,
min_samples_leaf=10, oob_score=True,
      random_state=10), param_grid=param_test4, cv=10
   )
model.fit(X_train, y_train)
print(model.best_params_)
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
params = {
      'n_estimators':40,
      'max_depth': 11,
      'min_samples_split': 30,
        'min_samples_leaf': 10,
        'max_features': 5
    }
rf = RandomForestRegressor(**params)
rf.fit(X_train, y_train)
y_pred = rf.predict(X_test)
evaluate_model(y_test, y_pred)
r2 : 0.448
MAE : 7.071
MSE : 87.65
round(rf.score(X_test,y_test)*100,2)
44.8
cols =['Age', 'Sex', 'Job', 'Housing', 'Saving accounts', 'Checking account','Credit amount']


plt.figure(figsize=(15, 5))
plt.bar(range(len(cols)), rf.fit(X_train, y_train).feature_importances_)
plt.xticks(range(len(cols)), cols, rotation=-45, fontsize=14)
plt.title('Feature importance', fontsize=14)
plt.show()
from sklearn.neural_network import MLPRegressor
mlpr = MLPRegressor(hidden_layer_sizes=(500,300),
                    activation='relu',
                   solver='adam',
                   max_iter=100,
                   random_state=123,
#                   early_stopping=True,
#                   validation_fraction=0.2,
#                   tol=1e-8,
                  )
mlpr.fit(X_train,y_train)
mlpr.fit(X_train, y_train)
y_pred = mlpr.predict(X_test)
evaluate_model(y_test, y_pred)
mlpr = MLPRegressor(hidden_layer_sizes=(500,),
                   activation='relu',
                   solver='adam',
                   max_iter=100,
                   random_state=123,
#                   early_stopping=True,
#                   validation_fraction=0.2,
#                   tol=1e-8,
                  )

mlpr.fit(X_train,y_train)
mlpr.score(X_test, y_test)
mlpr = MLPRegressor(hidden_layer_sizes=(1000,500),
                    activation='relu',
                    solver='adam',
                    max_iter=100,
                    random_state=123,
#                      early_stopping=True, ##
#                      validation_fraction=0.2, ##
#                      tol=1e-8,
                   )

mlpr.fit(X_train,y_train)
mlpr.score(X_test, y_test)
plt.figure()
plt.plot(mlpr.loss_curve_)
plt.xlabel("iters")
plt.ylabel(mlpr.loss)
plt.show()
mlpr.fit(X_train, y_train)
y_pred = mlpr.predict(X_test)
evaluate_model(y_test, y_pred)
mlpr.fit(X_train, y_train)
```

```
y_pred = mlpr.predict(X_test)
evaluate_model(y_test, y_pred)
r2 : 0.4033
MAE : 7.42
MSE : 94.75
from sklearn.linear_model import LinearRegression
lr= LinearRegression()
lr.fit(X_train, y_train)
y_pred = lr.predict(X_test)
evaluate_model(y_test, y_pred)
r2 : 0.4472
MAE : 7.209
MSE : 87.78
print(lr.score(X_test,y_test))
0.4472002098763278
```

1. makemodel.py

```
import pickle
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
df = pd.read_csv('data1.csv')
X = df.drop(['Duration'], axis=1).values
```

```
y = df['Duration'].values
X_train, X_test, y_train, y_test = train_test_split(X, y.astype('int'), test_size=0.3)
lr=LinearRegression()
lr.fit(X,y)
y_pred=lr.predict([[35,1,2,2,1,1,1744]])
# plt.plot(X,y)
# # plt.plot(X_test,y_pred)
# plt.show()
print(y_pred)
filename='lr.sav'
pickle.dump(lr,open(filename,'wb'))
```

3.app.py

```
import numpy as np
from flask import Flask, request, render_template
import pickle
import joblib
# 1. Loading the saved model
# We load the model.pkl file and initialize the flask app.
app = Flask(__name__)
model = joblib.load('lr.pkl')
# 2. Redirecting the API to the home page index.html
@app.route('/')
def home():
    return render_template('index.html')
# 3.Redirecting the API to predict the result (Duration)
@app.route('/predict',methods=['POST'])
def predict():
    int_features = [float(x) for x in request.form.values()]
    final_features = [np.array(int_features)]
    prediction = model.predict(final_features)
    output = round(prediction[0], 2)
    return render_template('index.html', prediction_text='The repayment time will be {} months'.format(output))
# 4. Starting the flask server
# Navigate to URL http://127.0.0.1:5000/ (or) http://localhost:5000
if __name__ == "__main__":
#    app.run(debug=True)
    app.run(host='127.0.0.1', port=8080, debug=True)
```

4.index.html

```
<!DOCTYPE html>
<html >
<!--From https://codepen.io/frytyler/pen/EGdtg-->
<head>
  <meta charset="UTF-8">
  <title>ML API</title>
```

```
 <link href='https://fonts.googleapis.com/css?family=Pacifico' rel='stylesheet'
type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Arimo' rel='stylesheet'
type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Hind:300' rel='stylesheet'
type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Open+Sans+Condensed:300'
rel='stylesheet' type='text/css'>
<link rel="stylesheet" href="{{ url_for('static', filename='css/style.css') }}">


</head>


<body>
 <div class="login">
        <h1>Predict Duration (Month) </h1>


    <!-- Main Input For Receiving Query to our ML -->
   <form action="{{ url_for('predict')}}"method="post">
{#        <input type="text" name="experience-1" placeholder="Experience-1"
required="required" />#}
{#        <input type="text" name="test_score-1" placeholder="Test Score-1"
required="required" />#}
{#        <input type="text" name="interview_score-1" placeholder="Interview
Score-1" required="required" />#}
```

```
     <input type="text" name="Age" placeholder="Age" required="required" />
    <input type="text" name="Sex" placeholder="Sex" required="required" />
         <input type="text" name="Job" placeholder="Job" required="required" />
    <input type="text" name="Housing" placeholder="Housing" required="required" />
    <input type="text" name="Saving" placeholder="Saving" required="required" />
    <input type="text" name="Checking" placeholder="Checking" required="required" />
    <input type="text" name="Credit" placeholder="Credit" required="required" />
    <button type="submit" class="btn btn-primary btn-block btn-large">Predict</button>
  </form>
 <br>
 <br>
 {{ prediction_text }}
</div>
</body>
</html>
```