РАЗРАБОТКА WEB-ПРИЛОЖЕНИЯ ДЛЯ ВЫБОРА ОПТИМАЛЬНОГО МЕТОДА ПРОГНОЗНОГО АНАЛИЗА ПЕРСПЕКТИВНЫХ НЕФТЯНЫХ СКВАЖИН.

Филипас И.А.
НИ ТПУ, ИШИТР, 8ПМ1И, email: iaf15@tpu.ru

Введение

Целью работы является написание web-приложения для упрощения работы с дата-сетами и выбора перспективных нефтяных скважин, а также в дальнейшем применения приведенных технологий в других индустриях.

Web-приложение позволяет пользователю загрузить свой дата-сет в формате .csv для его последующей обработки и составления прогнозных моделей.

На данном этапе реализована загрузка дата-сета, обработка мультиколлинеарности, а также построение статистики ROC\AUC кривых для сравнения точности методов предложенных моделей [1]. В качестве прогнозной модели была выбрана логистическая регрессия.

Основная часть

Основная задача создаваемого приложения заключается в том, чтобы пользователи, которые хотят применять технологии больших данных могли использовать их с большим удобством, а также меньшими затратами для анализа и выбора моделей.

Реализуемое приложение позволит пользователю загрузить дата-сет, который будет обработан программой, а затем, с помощью встроенных моделей, будет представлен краткий анализ, какие методы были использованы и какой из них более подходит под его задачи.

Для реализации этого подхода необходимо выбрать несколько методов анализа данных. Под анализом данных в данной работе подразумевается различные настройки логистической регрессии из библиотеки [2] sklearn и использования методологии подготовки данных [3].

На данный момент реализовано 4 метода прогнозного анализа, которые подходят для различных целей (пример работы веб-приложения представлен на рисунке 1):

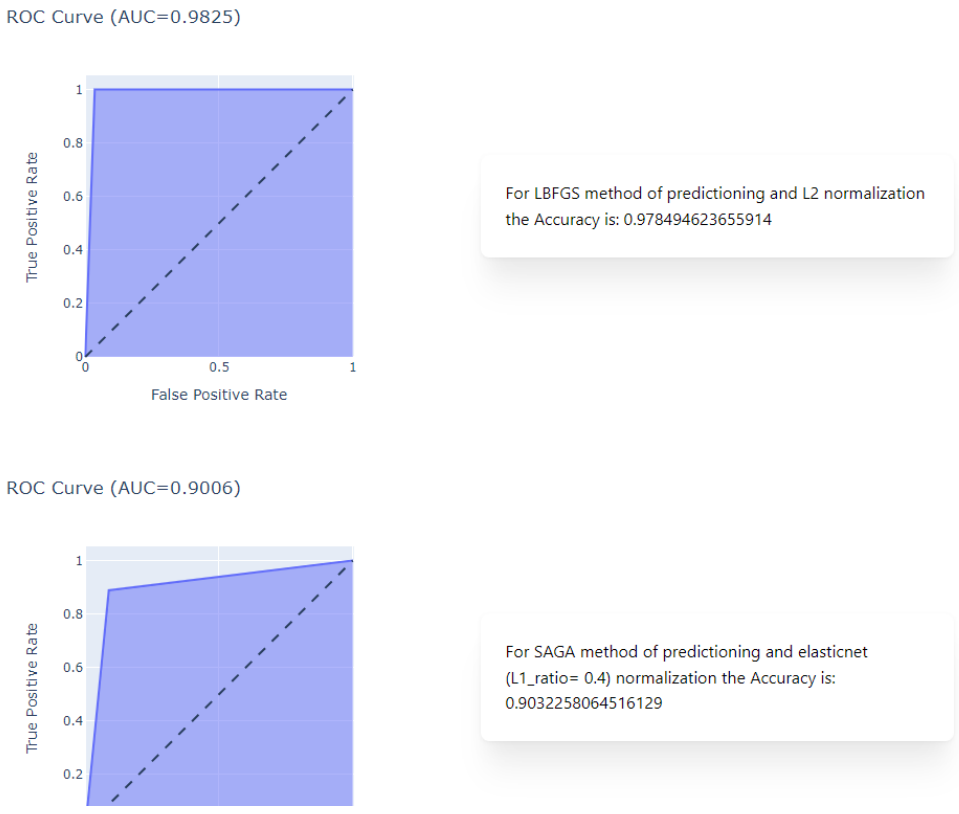

Рис. 1. Пример работы Web-приложения

1) базовый метод, который будет использован без каких-либо дополнительных настроек;
2) метод для больших объемов данных;
3) метод для одной целевой переменной;
4) метод для одной целевой переменной и большого объема данных.

В данной модели реализовано устранение мультиколлинеарности входных параметров, посредством сравнения анализа тепловой карты коллинеарности для выбранного дата-сета и удаления значений, которые могут повлиять на работу программы. Пример работы функции представлен на рисунках 2 и 3. В дальнейшем планируется унифицировать данную функцию, чтобы позволить пользователям выбирать уровень, до которого они хотят убирать переменные от, условного, “Very strict” (очень строгого) до “Very soft” (очень мягкого).

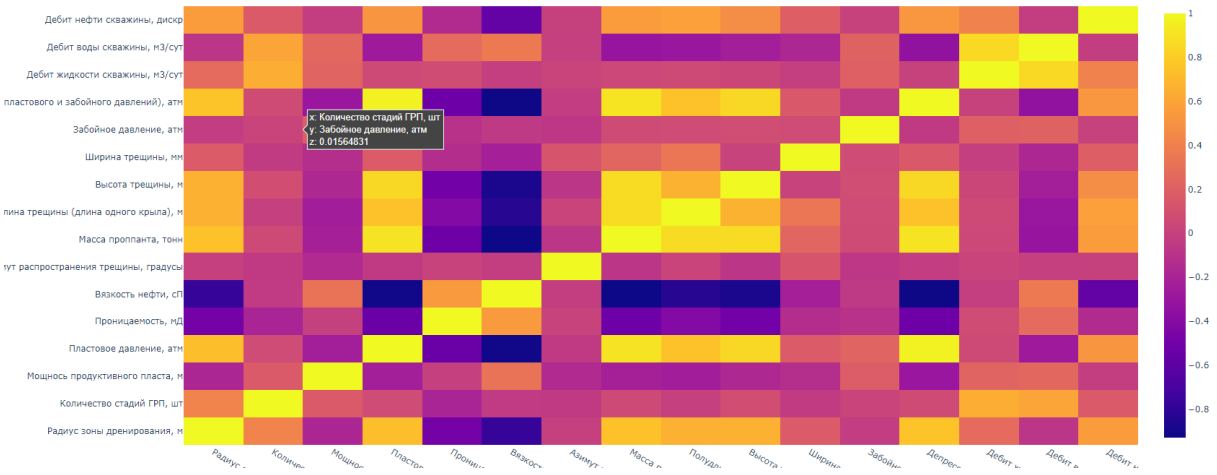

Рис. 2. Пример тепловой карты дата-сета до обработки

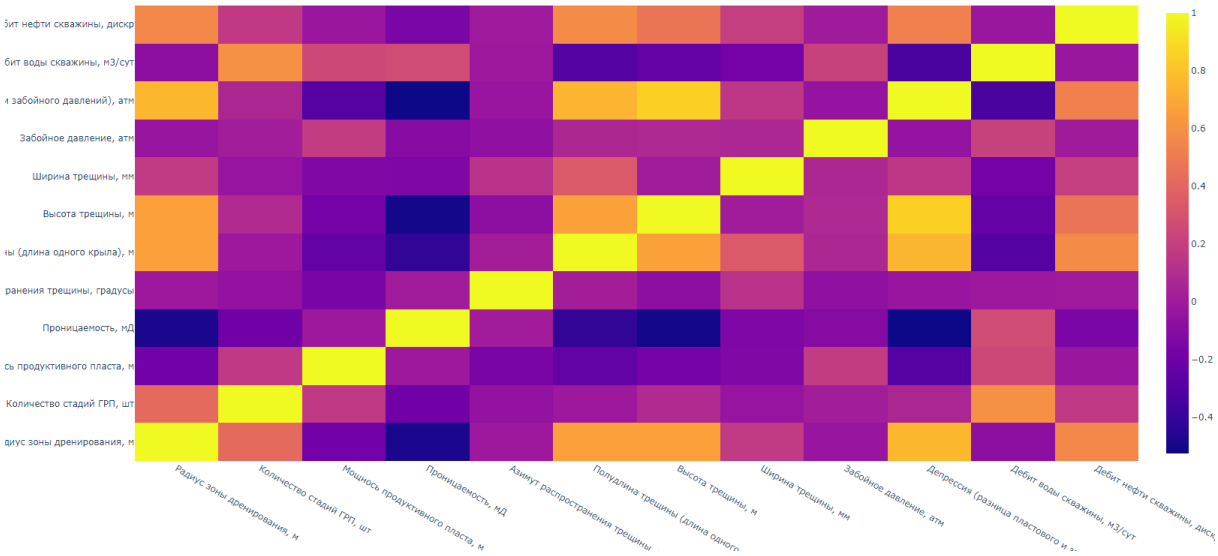

Рис. 3. Пример тепловой карты дата-сета после обработки

В дальнейшем планируется реализовать отдельные вкладки приложений для детального анализа эффективности методов. На вкладке для каждого метода будет реализован график, а также вся необходимая статистическая информация. Дополнительно на этой вкладке планируется реализовать ввод данных, для подстановки их в модель и расчёта рекомендаций.

Примерный путь пользователя данной системы заключен в следующих шагах:

1) загрузка дата-сета для обработки;
2) выбор метода, который пользователь посчитает более правильным, основываясь на значениях, предоставленных ему программой [4];

3) переход на вкладку с методом и просмотр статистических данных, графиков функций и прогнозной моделью;
4) ввод данных, для подстановки в модель и получения практических рекомендаций.

Заключение

На данном этапе реализован конкретный дата-сет, с обработкой мультиколлениарности, а также построения статистики ROC\AUC кривых для сравнения точности методов предложенных моделей [1]. В качестве прогнозной модели была выбрана логистическая регрессия.

В дальнейшем планируется усовершенствовать данный проект, добавить в него дополнительные функции, такие как:

1) обработка любого дата-сета;
2) выбор пользователем целевой функции, которую программа превратит в дискретную, если она таковой не является;
3) добавление страниц под каждый метод логистической регрессии.

Список использованных источников

1. Метрики в задачах машинного обучения. [Электронный ресурс]. – URL: https://habr.com/ru/company/ods/blog/328372/ (дата обращения 13.02.2023).
2. sklearn.linear_model.LogisticRegression. [Электронный ресурс]. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (дата обращения 03.02.2023).
3. Губин Е. И. Методика подготовки больших данных для прогнозного анализа. «Наука и бизнес: пути развития». Выпуск № 3(105). 2020, 2020. – [С. 33-35].
4. Metrics and scoring: quantifying the quality of predictions. [Электронный ресурс]. – URL: https://scikit-learn.org/stable/modules/model_evaluation.html (дата обращения 11.02.2023).