СЕКЦИЯ 2. ИНФОРМАЦИОННО-ВЫЧИСЛИТЕЛЬНЫЕ И РОБОТОТЕХНИЧЕСКИЕ СИСТЕМЫ

МЕТОДОЛОГИЯ ПОДГОТОВКИ БОЛЬШИХ ДАННЫХ ДЛЯ ПРОГНОЗОГО АНАЛИЗА

Е.И. Губин
Национальный исследовательский Томский политехнический университет,
Россия, г. Томск, пр. Ленина, 30, 634050
E-mail: gubine@tpu.ru

METHODOLOGY OF BIG DATA PREPARATION FOR PREDICTIVE ANALYSIS

E. I. Gubin
Tomsk Polytechnic University, Russia, Tomsk, Lenin Ave., 30, 634050
E-mail:gubine@tpu.ru

Annotation. *Formation of basic competencies in the preparation of initial data in the field of large amounts of data for predictive analysis*.

Бурное развитие информационных технологий особенно в области больших данных предъявляет повышенные требования к качеству исходных данных. Учитывая, что большая часть реальных данных носит характер слабо структурированных, то вопрос «чистоты» последних носит критический характер. Достаточно сказать, что при тщательной и корректной подготовке исходных данных удается почти на 20% повысить предсказательную силу традиционных прогнозных моделей. В литературе представлены подходы к подготовке исходных данных для дальнейшего предиктивного анализа [1–4], но они не охватывают всего спектра необходимых подготовительных этапов и носят характер некоторой «вкусовщины».

В настоящей работе автор хотел бы представить методику переподготовки исходных данных, включающую обязательные шаги статистического анализа и организации формата данных для корректного предиктивного анализа.

Процесс сбора и подготовки исходных данных, является одним из самых трудоемких и сложных этапов в анализе больших объемов данных, который порой занимает до 80% всего рабочего времени. Использование статистических методик и современного программного обеспечения позволяет значительно сократить временные и финансовые затраты на данном этапе и повысить эффективность и качество конечных результатов.

В работе предлагаются конкретные шаги по формированию основных бизнес целей и первичному анализу исходных данных, который включает проверку качества данных и простейшие статистики, исправление ошибочных и противоречивых данных. Важным этапом является формирование объясняющих переменных и выбора целевой функции.

В режиме подготовки первичных (исходных) данных осуществляется «очищение» данных, анализ «выбросов» и дублирующих строк. Важной составляющей этапа является выявление мультиколлениарности в объясняющих переменных и в случае ее наличия - удаление этих переменных. Масштабирование позволяет преобразовать исходные данные в единый цифровой формат, что значительно повышает точность прогнозных моделей.

В таблице 1 приведены основные этапы подготовки данных для дальнейшего прогнозного анализа и возможные шаги для последующей их корректировки.

В таблице 2 приведен пример выбора целевой функции (GoodBad) для финансовой модели, тренировочной (обучающейся) и тестовой выборок, а также процентное соотношение между ними.

Таблица 1 – Основные этапы подготовки исходных данных

Problems of initial data set/ исходные («грязные») данные	**Format attribute/** формат переменной	**Comment/** комментарий
1.*Missing data*/ Отсутствующие данные	Numeric/числовой, Char/текст	1.Add in (average, median, frequency...) 2.Delete this cases (rows)
2. *Mistakes of data*/ Ошибки в данных	Numeric/числовой, Char/текст	1. Add in (average, median, frequency...) 2. Delete this cases (rows)
3.Outliers of data / Выбросы данных	Numeric/числовой	Delete this cases (rows)
4.Duplicate cases(rows) /Дублирующие наблюдения(строки)	Duplicate ID (observations)	Remove one of the duplicate
5.Multicollinearity in the original data/ Мультиколлинеарность	Linear combination of variables (attributes)	Remove one of the attributes
6.Digitalization of data/ Цифровизация данных	Numeric/числовой, Char/текст	Converting to numeric format

Таблица 2 – Выбор целевой функции, обучающейся и тестовой выборок

Objective function/ Целевая функция(GB)	Binary (0,1)	«GB» = 1 плохой заемщик, «GB» = 0 хороший заемщик
Training samples/ Обучающая выборка	Sampling 70%–80%	Representative relative to the objective function (GB)/ Репрезентативная по GB
Testing samples/ Тестовая выборка	Sampling 30%–20%	Representative relative to the objective function (GB)/ Репрезентативная по GB

В данной работе предложена методика подготовки данных для построения прогнозных моделей классификации. Этапы подготовки данных включают в себя следующие шаги: **1**. проверку исходных данных на ошибки (описки), **2**. на отсутствие данных («missing»), **3**. на выбросы данных («outliers»), **4**. на наличие дублирующих строк (наблюдений), **5.** на проверку исходных объясняющих переменных (атрибутов) на мультиколлинеарность, **6**. трансформация исходных данных в цифровой формат («цифровизация») и **7**. выбор целевой переменной.

Полученная методика реализована в программных пакетах Python, SAS и SAS Enterprise Miner. Сравнение точности результатов, полученных без подготовки данных и с применением предложенной методики подготовки данных показало повышение предсказательной силы прогнозной модели почти на 20%. Наибольшую точность (75%) демонстрирует решение, полученное с помощью SAS Enterprise Miner.

СПИСОК ЛИТЕРАТУРЫ

1. Вершинин А.С., Губин Е.И. Применение инструмента DATA MINING для оценки кредитоспособности заемщика // Информационные технологии в науке, управлении, социальной сфере и медицине: Труды V Междунар. конференции. – Томск, 2018. – Т.2. – С. 18–21.
2. Вершинин А.С., Губин Е.И. Использование инструментов SAS для оценки рисков заемщиков // Молодежь и современные информационные технологии: Труды XVI Междунар. научно-практической конференции студентов, аспирантов и молодых ученых.– Томск, 2018. – С. 379–380.
3. Huang Shan, Gubin E. Data cleaning for data analysis // Молодежь и современные информационные технологии: Труды XVI Междунар. научно-практической конференции студентов, аспирантов и молодых ученых. – Томск, 2018. – С. 387–389.
4. Демченко И.С., Губин Е.И. Modern Big Data preprocessing techniques // Новая наука: история становления, современное состояние, перспективы: Труды Международной научно-практической конференции – Ч.1 – Стерлитамак, 2018г. – С. 4–7.

РАЗВИТИЕ МАГИСТЕРСКОЙ ПРОГРАММЫ «ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ»

В.Г. Спицын, Ю.А. Иванова, А.А. Друки
Национальный исследовательский Томский политехнический университет,
Россия, г. Томск, пр. Ленина, 30, 634050
E-mail: spvg@tpu.ru

DEVELOPMENT OF A MASTER PROGRAM «ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING»

V.G. Spitsyn, Y.A. Ivanova, A.A. Druki
Tomsk Polytechnic University, Russia, Tomsk, Lenin str., 30, 634050
E-mail: spvg@tpu.ru

Annotation. *The development of the master program "Artificial Intelligence and Machine Learning" is described. The foundations of this area were formed during the formation at the Department of Computer Engineering in 2004 of the scientific group of Intellectual Image Processing. In the process of developing a scientific group from 2007 to 2019, 9 candidate dissertations were defended. 4 initiative projects of the group are supported by the RFBR Grants. Currently, the educational process for the master's program is provided by 1 professor, doctor of technical sciences, 2 associate professors, candidate of technical sciences, 3 assistants and 8 graduate students.*

В 2004 г. на кафедре вычислительной техники ТПУ была создана научная группа Интеллектуальной обработки изображений (ИОИ) в составе: профессор Спицын В.Г., аспирант Цой Ю.Р., студенты: Чернявский А.В., Федотов И.В., Белоусов А.А. В 2006 г. предложенный научной группой проект «Разработка технологии автоматизированного улучшения качества цифровых изображений на основе применения эволюционирующей нейронной сети» был поддержан грантом РФФИ № 06-08-00840.

В 2007 г. в диссертационном совете Д 212.269.06 при ТПУ защищена диссертация на соискание ученой степени кандидата технических наук: Цой Ю.Р. «Нейроэволюционный алгоритм и программные средства для обработки изображений» [1]. Руководитель – профессор Спицын В.Г.

В 2009 г. предложенный научной группой проект «Создание программного комплекса автоматизированной обработки изображений и распознавания образов на