Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 658.1-047.44:004.85

Епимашко
Александр Дмитриевич

АЛГОРИТМЫ АНАЛИЗА И АВТОМАТИЗАЦИИ ОБРАБОТКИ ФИНАНСОВОЙ ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ

на соискание степени магистра технических наук

по специальности: 1-40 80 06 «Искусственный интеллект»

_____

Научный руководитель
Бойко Игорь Михайлович
Кандидат технических наук

_____

Минск 2021

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Магистерская работа не имеет связи с другими научными исследованиями университета, так как связана с фирмой, где работает магистрант, а именно ИВЦ Минфина. Результаты диссертационного исследования были использованы врабочем процессе, при построении различных прогнозов на основе информации хранилища данных.

Цель работы состоит в исследовании и автоматизации обработки различной финансовой информации на предприятии с использованием технологий машинного обучения и последующего принятия решения для получения большей прибыли.Задачи исследования: анализ финансовой информации с использованием технологий машинного обучения; выбор конкретного алгоритма машинного обучения,построение модели и оценка ее качества;разработка алгоритмов оптимизации и автоматизация обработки данных о финансах.

Новизна полученных результатов характеризуется разработкой модели для получения определенных прогнозов в результате анализа финансовой информации. Кроме того, был исследован алгоритм машинного обучения многомерной линейной регрессии и к нему были применены различные усовершенствования, для более быстрой и точной работы алгоритма.

Положения, выносимые на защиту: в результате работу с магистерской диссертацией, был достигнут ряд результатов, в частности были разработаны несколько этапов предобработки изначальных данных, такие как разработка и отбор признаков, была разработана модель анализа финансовой информации для предсказания различных результатов. Это было выполнено путем модернизации существующего алгоритма машинного обучения – многомерной линейной регрессии, а точнее было увеличено скорость работы алгоритма, а также его качество. Кроме того, результаты магистерской работы, могут значительно улучшить обработку финансовой информации, ее анализ и получение дополнительной выгоды в результате более качественного планирования.

Было принято участие в научных конференциях аспирантов, магистрантов и студентов учреждения образования БГУИР, а также, некоторые работы были опубликованы в научном журнале RSGlobal.

Список публикаций:

– «Кто оказал содействие разработке ISO/IEC TR 9126-2:1991».

– «Цифровые системы управления данными и системы поддержки принятия решений»

– «Система контроля доступа на основе биометрических параметров»

Общий объем магистерской диссертации составляет 63 страниц, которые включают в себя 4 главы:«системный анализ и постановка задачи», «подбор и обработка финансовой информации для анализа алгоритмами

машинного обучения», «построение модели и ее обучение для анализа и автоматизации обработки финансовой информации» и «оценка качества созданной модели», куда включены 21 иллюстрация, 13 таблиц, библиографический список из 30 наименований, 1 приложение.

КРАТКОЕ ВВЕДЕНИЕ

Финансы организаций являются частью общей системы финансовых отношений и отражают процесс образования, распределения и использования доходов на предприятиях.Анализ финансовой информации необходим, чтобы принять правильное управленческое решение, которое влечет за собой большую прибыль или возможность сократить убытки. Особенно остро встает вопрос об исследовании и автоматизации обработки очень больших объемов данных. Рассмотрение вопроса об анализе и автоматизации обработки финансовой информации является актуальным.

Машинное обучение используется и в финансовой сфере, однако пока этот процесс на многих фирмах еще только развивается. Многие компании обычно ретроспективны, то есть часто лишь представляют прошлые или настоящие факты без обеспечения достаточного контекста, без объяснения причинно-следственных связей, а также без рекомендаций, какие шаги предпринять. Иными словами, они фиксируют произошедшее, но ничего не предписывают. Это работа различных BI-систем, и в этом отношении рост потенциала компании довольно ограничен, так как формируемые отчеты содержат информацию о прошлом, а не о будущем. Построением различных прогнозных моделей, рекомендаций и планов, для дальнейшего развития предприятия на основе имеющейся информации, занимаются управляющий состав компании, а для построение более лучших прогнозов помогут – алгоритмы машинного обучения.

Именно по этим причинам было принято решение использовать алгоритмы машинного обучения для проведения анализа и автоматизации обработки финансовых данных Министерства Финансов Республики Беларусь. По причине наличия большого количества информации в хранилище данных и строящихся, на основе лежащих там информации, отчётов, анализ с помощью алгоритмов машинного обучения представляется возможным. Это будет необходимо для улучшения отслеживания различных факторов, например, таких как, тенденцию роста или падения доходов, или расходов в определенных областях, предсказание о росте или погашении Государственного долга Республики Беларусь.

В диссертационной работе будут рассмотрены алгоритмы анализа и автоматизации обработки данных о финансах с использованием технологии машинного обучения для успешного функционирования предприятия.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Магистерская диссертация состоит из 4 глав, каждая из которых содержит разделы и подразделы:

– первая глава «Системный анализ и постановка задачи» включает в себя теоретические сведения касательно магистерской диссертации. В главе описывается назначение алгоритмов анализа финансовой информации, зачем это необходимо и какие показатели можно анализировать; описываются различные алгоритмы машинного обучения, их характеристики и краткий анализ, произведена сравнительная характеристика алгоритмов. Кроме того, был произведен анализ применения алгоритмов машинного обучения в других существующих проектах и были поставлены задачи, которые необходимо решить в магистерской диссертации.

– вторая глава «Подбор и обработка финансовой информации для анализа алгоритмами машинного обучения» включает определение и описание различных метрик, как регрессии, так и классификации, для оценки работы алгоритмов работы машинного обучения. Были определены используемые данные и определены признаки, а также описаны действия, предпринятые в предобработке данных, для более корректной работы алгоритмов машинного обучения. Было описано, как формировались и отбрасывались признаки, с помощью различных корреляций, для более качественной работы алгоритма машинного обучения.

– третья глава «Построение модели и ее обучение для анализа и автоматизации обработки финансовой информации» включает процесс исследования алгоритмов машинного обучения и выбора конкретного из них для поставленной задачи. Также были описаны подходы, которые были предприняты для построения модели машинного обучения для анализа и автоматизации обработки финансовой информации, а также подходы, предпринятые к ее обучению, для получения определенных прогнозов, на основе построенной и обученной модели и данных, которые будут выбраны для обучения.

– четвертая глава «Оценка качества созданной модели» включает оценку работы алгоритма машинного обучения с помощью определенной метрики, а также описание выбора определённой метрики для оценки работы алгоритма. Кроме того, описано, как выбирались обучающие и валидационные выборки.

Были описаны выводы, которые были получены в результате написания магистерской диссертации, а также были описаны использованные источники, которые были использованы для написания данной диссерации.

ЗАКЛЮЧЕНИЕ

В ходе работы с магистерской диссертацией были достигнуты следующие результаты:

– были исследованы различные источники информации по предметной области магистерской диссертации, в частности, это касается анализа финансовой информации, различных алгоритмов машинного обучения и методам оценки их качества работы;

– были определены используемые для дальнейшего анализа данные и на основе их были выделены объекты и их признаки. Данные объекты прошли необходимую предобработку для дальнейшего анализа их определенной технологией машинного обучения. На этапе предобработки данных были выполнены следующие действия: приведение всех булевых значений, в том числе представленных в текстовой форме («0» или «1»), в числовую форму (0,1); приведение к общему формату всех числовых значений вплоть до четырех знаков после запятой; подключение обработки дубликатов, противоречивых и фиктивных значений; восстановление или заполнение пропусков в булевых данных; устранение проблем с неопределенными типами данных и отсутствующими значениями.

– был произведен отбор признаков для устранения проблемы коллинеарности. Это было выполнено для числовых признаков с использованием коэффициента корреляции Пирсона, где значения, которые не входили в интервал [-0.8,0.8] были отброшены и для бинарных признаков с использованием коэффициента корреляции Мэтьюса, где значения, которые не входили в интервал [-0.95,0.95] также отбрасывались для дальнейшей более точной работы алгоритма машинного обучения.

– была разработана модель анализа и обработки финансовой информации для предсказания различных результатов. Быливыполненыусовершенствования существующего алгоритма машинного обучения – многомерной линейной регрессии, а точнее была увеличена скорость работы алгоритма, а также его качество. Это было достигнуто путем изменения оригинальной формулы алгоритма линейной регрессии путем сингулярного разложения исходных матриц-признаков, а также путем применения других различных дополнительных преобразований.

– с помощью метрик качества была произведена оценка работы разработанной модели. Для метрик регрессии былавыбрана метрика средней абсолютной ошибки (MAE), а для метрик классификации – F-мера. К исходным данным была применена кросс-валидация, для того, чтобы избежать проблемы переобучения алгоритма машинного обучения.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

Было принято участие в научных конференциях аспирантов, магистрантов и студентов учреждения образования БГУИР, а также, некоторые работы были опубликованы в научном журнале RSGlobal.

Список публикаций:

– «Кто оказал содействие разработке ISO/IEC TR 9126-2:1991».

– «Цифровые системы управления данными и системы поддержки принятия решений»

– «Система контроля доступа на основе биометрических параметров»