ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук

Образовательная программа «Экономика и анализ данных»

Отчет о командном программном проекте

на тему **Рекомендательные системы для B2B** (промежуточный, этап 1)

Выполнен студентами:

Группы БЭАД221, 3 курса	Рыльцева Полина Алексеевна
Группы БЭАД222, 3 курса	Зайцева Арина Романовна
Группы БЭАД222, 3 курса	Валиуллина Рената Руслановна

Проверен руководителем проекта:

Мунерман Илья Викторович

Директор исследовательского центра

ИА Интерфакс

Москва 2025

Содержание

1 Аннотация

Проект направлен на разработку рекомендательной системы для B2B-сегмента, строящей прогноз компаний-победителей государственных и коммерческих тендеров. Основная идея заключается в анализе исторических данных, чтобы генерировать списки компаний из числа потенциальных участников, например для целевых рассылок. В рамках проекта предполагается тестировать различные подходы: от классических методов машинного обучения до, возможно, нейросетевых архитектур и графовых алгоритмов, моделирующих связи между заказчиками и поставщиками. Для каждого решения необходимо провести сравнительный анализ с отбором оптимальной модели на основе различных метрик.

Особое внимание предполагается уделить данным, используемым моделью — условий тендеров, профилей компаний, результатов прошлых тендеров. Это может позволить выявлять не только схожесть между компаниями, но и факторы, повышающие шансы победы. Результаты работы моделей планируется визуализировать интерактивно: отображать результаты модели и метрики качества.

Итоговый результат будет содержать лучшую модель/модели — от предобработки данных до визуализации полученного решения. Наш проект способствует повышению эффективности взаимодействия между заказчиками и поставщиками в рамках тендерных процессов, снижая временные и ресурсные затраты на поиск и привлечение релевантных участников. Он также способствует развитию цифровизации в сфере госзакупок и коммерческих закупок, делая их более прозрачными и доступными.

2 Ключевые слова

Рекомендательные системы, машинное обучение, коллаборативная фильтрация, графовые алгоритмы, B2B-тендеры, сравнительный анализ моделей, визуализация данных

3 Введение

В сфере государственных и коммерческих закупок процесс поиска и привлечения подходящих участников тендеров может быть достаточно сложным. Заказчики анализируют огромные объемы данных, что требует значительных временных затрат. С другой стороны, поставщики могут не иметь доступа к релевантным для них тендерам, что снижает их шансы на участие и победу. Эти проблемы подчеркивают необходимость создания решений, которые могли бы упростить процесс взаимодействия между заказчиками и поставщиками, сделав его более эффективным.

Наш проект направлен на разработку рекомендательной системы, которая способна прогнозировать компании-победители тендеров на основе анализа исторических данных. В отличие от классических подходов, наш проект фокусируется на комплексном анализе данных, включая условия тендеров, профили компаний, результаты прошлых закупок, и на широком спектре моделей для выявления лучшего подхода.

Выполнение проекта включает в себя следующие этапы:

1. **Сбор и верификация данных**

   На этом этапе происходит сбор данных из системы Маркер Интерфакс, с которыми работает наша команда, а также предобработка, очистка данных и их анализ.

2. **Feature engineering**

   Здесь создаются и отбираются признаки, которые будут использоваться для обучения моделей. Это включает в себя генерацию новых переменных, например, таких как частота участия компании в тендерах, средняя цена предложений и другие метрики, которые могут повлиять на результат. Кроме того, для использовая различных методов машинного обучения и иных может потребоваться выполнение этого этапа несколько раз.

3. **Моделирование**

   На этом шаге тестируются различные алгоритмы машинного обучения, начиная с классических методов, таких как логистическая регрессия и метод опорных векторов (SVM), и заканчивая более сложными подходами, например, графовыми алгоритмами. Каждая модель оценивается по соответствующим для ее спецификации метрикам, например, precision, recall, F1-score, AUC-ROC. Также возможным шагом здесь может быть и ансамблирование моделей, дающих хорошие результаты.

4. **Получение оценок и интерпретация**

После обучения моделей проводится анализ их результатов, главным образом используя метрики качества. Это в том числе включает в себя интерпретацию важности признаков, выявление ключевых факторов, влияющих на победу в тендерах, а также выявление взаимосвязей между моделями.

5. **Тестирование**

   На этом этапе результаты лучшей модели/моделей необходимо протестировать на тестовых данных, чтобы убедиться в их устойчивости и способности обобщать новые данные.

6. **Создание визуализации**

   Результаты работы моделей из предыдущего шага визуализируются в виде интерактивных графиков или дашбордов, где можно увидеть ключевые метрики, возможно, значимость признаков и прочее. Предполагается добавление фильтров для более адаптивного представления результатов.

3.1 Распределение ролей в команде

Распределение задач в команде организовано следующим образом: для того, чтобы не было последовательного выполнения этапов участниками проекта, предлагается разделить каждый шаг на отдельные роли для членов команды.

В целом, моделирование задачи можно разделить на 3 блока:

1. использование классического машинного обучения (например, логистическая регрессия, дерево решений, и прочее): Зайцева Арина Романовна

2. использование классического ML и более сложные модели (например, графовые алгоритмы и нейросети): Рыльцева Полина Алексеевна

3. ансамблирование моделей, полученных на первых двух блоках: Валиуллина Рената Руслановна

Блок 2 также включает в себя классический ML, поскольку присутствует широкий класс моделей, которые можно протестировать в этой области. Кроме того, участник, работающий над определенным классом моделей также самостоятельно готовит для нее данные, тестирует модель и интерпретирует результаты. Таким образом, для каждого блока моделей ответственный за них выполняет шаги 1 - 5 выполнения проекта соответственно.

Касательно визуализации, участие участников в ее выполнении полностью зависит от итогового ее вида, что не представляется возможным определить на данном этапе выполнения проекта. Для выполнения этого шага может потребоваться от 1 до 3 участников, поэтому команда приняла решение не сецифировать ее реализацию на текущий момент.

4 Анализ существующих решений

Перед началом работы наша команда изучила, существуют ли аналогичные решения для поставленной задачи.

Было обнаружено, что ключевым сервисом для тендерных специалистов в России на сегодняшний день является Маркер Интерфакс [1]. Сервис имеет удобную платформу, на которой можно мониторить закупки, анализировать деятельность участников рынка, подгатавливать аналитические отчеты.

Данный сервис дает всю необходимую информацию для заказчиков и поставщиков. Однако основным интрументом для отбора нужных тендеров ялвяется фильтрация (по региону, типу тендера, цене, отрасли и т.д.). Несмотря на то, что Маркер Интерфакс предоставляет много возможностей для мониторинга рынка, важно понимать, что для извлечения наибольшей пользы от специалиста требуется иногда самомстоятельно проанализировать большое количество информации, если он хочет минимизировать риски и эффективно находить возможности.

Для заказчиков на платформе нет инструмента, который под их конкретный запрос сможет найти тех поставщиков, которые могут с высокой вероятностью успешно выполнить заказ. Этой задачей и решила заняться наша команда.

Мы также нашли еще одну платформу, которая собирает информацию о закупках - Официальный сайт единой информационной системы в сфере закупок в информационно-телекоммуникационной сети Интернет (Официальный сайт ЕИС - [2]). Он предназначен для обеспечения свободного и безвозмездного доступа к полной и достоверной информация о контрактной системе в сфере закупок и закупках товаров, работ, услуг отдельными видами юридических лиц, а также для формирования, обработки и хранения такой информации. Порядок размещения информации и ее содержание регламентируется 44-ФЗ и 223-ФЗ.

На Официальном сайте ЕИС пользователи могут найти информацию о проводимых закупках, реестр контрактов, планы закупок, нормативно-правовые акты, реестр недобросовестных поставщиков, реестр жалоб и еще много полезных для работы ресурсов.

Однако, так же как и в Маркере Интерфакса, нет возможности оперативно, без долгих тщательных поисков и обработки данных, найти надежных поставщиков, которые смогут с высокой вероятностью успешно выполнять заказ.

[1] Маркер Интерфакс: https://marker-interfax.ru
[2] Официальный сайт ЕИС: https://zakupki.gov.ru

5 Обзор литературы

Для построения рекомендатыельных систем необходимо провести анализ как структурированных данных (характеристики компаний, параметры тендеров), так и временных паттернов (история участий, динамика активности). В нашем проекте мы будем использовать базу данных Интерфакс, откуда мы будем брать информацию о компаниях, их финансовых показателях, участии в тендерах и других релевантных данных. Рассмотрим ключевые методы, их применение в существующих исследованиях и адаптируем их идею к нашей задаче прогнозирования победителей для рассылки приглашений при таргетировании компаний

5.1 Метод опорных векторов (SVM)

В статье "Data Mining to Identify Anomalies in Public Procurement Rating Parameters"[5] SVM используется для выявления аномалий в параметрах оценки тендерных заявок. Авторы поставили перед собой цель обнаружить случаи возможной коррупции, когда поставщик с завышенной ценой выигрывает тендер благодаря манипуляциям с параметрами оценки.

Сначала данные о тендерах кластеризуются с помощью алгоритма k-means, затем анализируются с помощью Self-Organizing Maps (SOM), чтобы выявить кластеры с потенциальными аномалиями. Тендеры из "нормальных" кластеров (цена играет значительную роль) помечаются как "не аномалии" , а тендеры из кластера "Null Economic Offer"(вес цены очень маленький) используются как неразмеченные данные. Обучается модель SVM с использованием размеченных и неразмеченных данных. Модель находит границу между "нормальными" тендерами и потенциальными аномалиями в пространстве признаков, сформированном с помощью метода главных компонент.

Для решения поставленной нами задачи SVM можно использовать для предсказания, станет ли конкретная компания победителем тендера. Дополнительно можно использовать кластеризацию (k-means, SOM или другие алгоритмы) для разбиения компаний или тендеров на группы, а затем обучать отдельные модели SVM для каждой группы, это способствовало бы повышению точности предсказаний, если в разных группах действуют разные факторы.

5.2 XGBoost

Статья "Machine Learning in Procurement with a View to Equity"[2] демонстрирует примеры использования различных методов для анализа тендоров и компаний XGBoost — это мощный алгоритм градиентного бустинга. Он строит ансамбль деревьев решений последо-

вательно, где каждое последующее дерево пытается исправить ошибки предыдущих. Это достигается путем оптимизации функции потерь на каждом шаге. Авторы решают задачу предсказания наличия изменений в контрактах, заключенных по результатам тендеров. Это бинарная классификация (изменение есть/изменения нет). Использовались данные из базы ANAC за 2016-2019 годы, содержащие информацию о тендерах и awarded contracts. Данные были сбалансированы по целевому классу (наличие/отсутствие variation) путем выбора равного количества примеров из каждого класса. Модель оценивалась с помощью cross-validation по метрикам accuracy, precision, recall и F1-score. XGBoost показал наилучшие результаты

XGBoost можно использовать также для анализа важности признаков для понимания, какие факторы больше всего влияют на возникновение изменений, которые в свою очередь могут привести к изменению результатов участия компаний в тендере.

5.3 Логистическая регрессия (LogReg)

Логистическая регрессия — линейная модель для решения задач бинарной классификации. Она предсказывает вероятность принадлежности объекта к одному из двух классов. В статье [2] логистическая регрессия используется в качестве baseline-модели. Её преимущество — простота и быстрота обучения. Логистическую регрессию можно было бы использовать для оценки влияния отдельных признаков на вероятность возникновения изменений. Коэффициенты модели показывают направление и силу связи между признаками и целевой переменной.

5.4 Random forest

В статье "Bidders Recommender for Public Procurement Auctions Using Machine Learning"[3] используется Random Forest(ансамблевый метод), основанным на деревьях решений.

Модель Random Forest обучается на данных о тендерах и компаниях. Для построения модели используются различные признаки тендера и компаний. Целевая переменная - CIF (идентификатор) компании-победителя. Для нового тендера Random Forest предсказывает наиболее вероятного победителя. Каждое дерево в Random Forest делает свое предсказание, и окончательный результат определяется путем голосования среди всех деревьев. После того, как Random Forest предсказал победителя, алгоритм ищет в базе данных компании, которые похожи на предсказанного победителя по определенным критериям (размер компании, финансовые показатели, коды деятельности, географическое положение).

5.5 Decision Tree

В статье "Machine Learning in Procurement with a View to Equity"[2] дерево решений используется как baseline-модель, она позволяет оценить эффективность более сложных алгоритмов, таких как XGBoost и Random Forest, в задаче предсказания изменений в контрактах

Дерево решений - относительно простой и понятный алгоритм. Его структура в виде дерева позволяет легко визуализировать процесс принятия решений и выявить важные признаки, влияющие на наличие изменений. Это особенно полезно для baseline-модели, так как позволяет быстро оценить, какие факторы в принципе могут быть связаны с изменениями в контрактах.

Также сравнивая результаты дерева решений с результатами более сложных моделей, можно сделать вывод о том, насколько XGBoost и Random Forest превосходят простой подход по точности предсказаний, что говорит о возможной целесообразности использовать ансамблевого метода для нашей задачи.

5.6 Naive Bayes

Наивный Байес — вероятностный классификатор, основанный на теореме Байеса: он предполагает, что признаки независимы, что упрощает вычисления. Наивный Байес используется как еще один baseline-алгоритм, однако в статье [2] он показал наихудшую точность среди рассмотренных моделей, что может быть связано с нарушением предположения о независимости признаков, что требует дополнительной проверки и использования других методов.

Наивный Байес может быть эффективен для задач с большим количеством категориальных признаков и ограниченным объемом данных, помимо этого он относительно устойчив к выбросам.

5.7 Коллаборативная фильтрация (CF)

В статье "A Hybrid Recommender System for Sequential Recommendation: Combining Similarity Models With Markov Chains"[6] авторы предлагают гибридный подход, который комбинирует модели, основанные на сходстве с цепями Маркова, они в свою очередь могут включать в себя элементы коллаборативной фильтрации.

Коллаборативная фильтрация основана на следующей идее: если два пользователя имеют схожие предпочтения в прошлом, то они, вероятно, будут иметь схожие предпочте-

ния и в будущем. Алгоритм коллаборативной фильтрации анализирует данные о прошлых взаимодействиях пользователей с товарами (например, рейтинги, покупки, просмотры) и находит пользователей со схожими предпочтениями. Затем, на основе предпочтений этих "похожих"пользователей, алгоритм генерирует рекомендации для целевого пользователя. В задаче рекомендации компаний для участия в тендерах коллаборативная фильтрация может быть применена следующим образом: тендеры выступают в роли "пользователей" , а компании — в роли "товаров". Сходство между тендерами определяется на основе различных факторов. Идея заключается в том, что если два тендера похожи, то компании, которые участвовали в одном из них, с большой вероятностью будут заинтересованы и в другом. Таким образом, для данного тендера алгоритм коллаборативной фильтрации порекомендует компании, которые участвовали в похожих тендерах.

5.8 Графовые модели

В статье "Data Quality Barriers for Transparency in Public Procurement"[4] графовые методы используются для представления и анализа данных о государственных закупках в Словении. Данные из разных источников интегрируются в граф знаний (Knowledge Graph), что позволяет выявлять аномалии и паттерны, которые могут говорить о мошенничестве или неконкурентных рынках.

Данные о закупках, компаниях и расходах представляются в виде графа, где узлы — это сущности (тендеры, компании, организации), а ребра — связи между ними (например, компания участвовала в тендере, организация является заказчиком). Финансовые транзакции между государственными организациями и компаниями также моделируются в виде динамического графа, где узлы - организации и компании, а ребра - транзакции между ними. Этот граф анализируется с помощью различных методов обнаружения аномалий. Граф Знаний и граф транзакций используются для выявления аномалий, таких как компании с малым числом сотрудников, выигрывающие крупные тендеры, или слишком высокие расходы в определенных областях.

Для нашей задачи можно было бы построить граф знаний для моделирования связей между тендерами, компаниями и заказчиками. Это позволило бы анализировать структуру рынка и выявлять скрытые связи. Также можно использовать алгоритмы на графах (например, PageRank, Node2Vec) для ранжирования компаний по их релевантности к определенному тендеру или для рекомендации потенциальных партнеров для совместного участия в тендерах.

5.9 Нейронные сети

В статье "Public Procurement Contracts Futurity: Using of Artificial Intelligence in a Tender Process"[1] нейронные сети упоминаются в контексте искусственного интеллекта и машинного обучения как один из инструментов, которые могут быть применены в государственных закупках. В статье отмечается, что ИИ, включая нейронные сети, может автоматизировать задачи, которые традиционно выполнялись людьми, например, оценка тендерных заявок. Авторы подчеркивают, что нейронные сети эффективны для работы с большими объемами данных, которые генерируются в процессе государственных закупок, также нейронные сети могут быть интегрированы в системы электронных закупок для повышения их эффективности и прозрачности информации.

Для нашей задачи можно было бы обучить нейронную сеть для прогнозирования, какая компания с наибольшей вероятностью выиграет тендер. В качестве входных данных можно использовать информацию о компании, о самом тендере, а также данные о предыдущих тендерах и их победителях. Также можно использовать нейронные сети (например, рекуррентные сети (RNN или трансформеры) для анализа текста тендеров и извлечения из него полезных признаков (например, тематика, сложность, требования). Это позволит создать более точные модели для предсказания победителей и таргетирований компаний для рассылки.

5.10 Метрики качества

В результате нашей работы мы хотим прийти к наиболее точному таргетирвоанию компаний для рассылки с предложением участия в тендере, поэтому нам важна не только общая точность, но и охват потенциально интересных компаний, а также минимизация ошибочных рекомендаций, которые могут привести к лишним затратам на рассылку и снижению её эффективности.

a) Precision: доля релевантных рекомендаций (компаний, которые действительно выиграли бы тендер или приняли бы участие) среди топ k рекомендованных компаний для тендера. Эта метрика важна, так как подразумевается ограниченное количество компаний, которым будет производиться рассылка

b) Recall: доля релевантных компаний, которые попали в топ k рекомендаций, относительно общего числа всех релевантных компаний для каждого тендера, будет показывать насколько полно мы охватим потенциальных победителей

c) Average precision: усредненное значение precision по всем тендерам, это более ста-

бильная метрика, чем precision для отдельных тендеров.

d) Mean Average Precision: среднее значение average precision по всем тендерам, метрика для оценки общей производительности системы.

e) NDCG (Normalized Discounted Cumulative Gain): учитывает позицию релевантных рекомендаций в списке - чем выше позиция релевантной рекомендации, тем больший вклад она вносит в значение метрики. Использование данной метрики может быть целесообразным, так как важно, чтобы наиболее вероятные победители были в начале списка.

f)AUC: показывает способность модели различать победителей и не победителей. Полезно для оценки качества модели в целом, но не учитывает ранжирование.

g) F1-score: гармоническое среднее между precision и recall

h) ROI (Return on Investment): отношение прибыли от проведенных тендеров к затратам на рассылку, показывает экономическую эффективность рекомендательной системы.

i) Coverage: доля компаний, которым система может дать рекомендации. Данный показатель может быть важен для оценки, насколько полно система охватывает всех потенциальных участников тендеров.

5.11 Заключение

Проведенный обзор литературы демонстрирует широкий спектр методов машинного обучения, применяемых для анализа государственных и коммерческих закупок, каждый из которых вносит уникальный вклад в решение ключевой задачи проекта: прогнозирование победителей, а также иных задач этой области - рекомендация участников и оценка рисков, обнаружение аномалий.

С точки зрения рекомендательных систем и анализа связей,Random Forest и коллаборативная фильтрация продемонстрировали потенциал в рекомендации компаний-победителей, учитывая как признаки тендеров, так и исторические паттерны участия. Графовые модели, включая Knowledge Graphs, открывают возможности для выявления скрытых связей и структурных свойств рынка. Нейронные сети предлагают перспективы для обработки текстовых данных тендеров и прогнозирования победителей на основе комплексных признаков.

Если говорить про метрики и экономическую эффективность, Precision, Recall, F1-score критичны для минимизации ложных рекомендаций и максимизации охвата целевых компаний.В то же время, NDCG и MAP акцентируют внимание на ранжировании, что важно для приоритизации рассылок. ROI и Coverage связывают результаты с экономической эффективностью, обеспечивая баланс между затратами на рассылку и прибылью.

6 Дальнейший план работы

6.1 Данные и модели

В качестве следующих шагов работы с данными будет проведен окончательный отбор необходимых для работы признаков и feature engineering.

В рамках проекта будет произведена большая работа по моделированию рекомендательной системы. Планируется создать несколько моделей из широкого спектра - от логистической регресси до нейросетей. Будут протестированы различные варианты и отобран лучший по результатам тестирования. Отметим, что в список этих вариантов войдут также ансамбли моделей.

6.2 Визуализация

Заключительным этапом станет визуализация. В качестве наиболее практичных и простых для восприятия форматов можно рассмотреть **Power BI** и **Tableau**, а также создание интерактивных дашбордов. Один из них будет использован для демонстрации полученных результатов работы.

Список литературы

[1] Karem Aboelazm и Khalid Dganni. “Public procurement contracts futurity: Using of artificial intelligence in a tender process”. В: *Corporate Law Governance Review* 7 (янв. 2025), с. 60—72. DOI: 10.22495/clgrv7i1p6.

[2] Ishrat Fatima, Roberto Nai и Rosa Meo. “Machine Learning in Procurement with a View to Equity”. В: янв. 2025. DOI: 10.5772/intechopen.1008730.

[3] Manuel J. García Rodríguez, Vicente Montequín, Francisco Ortega-Fernández и Joaquín Balsera. “Bidders Recommender for Public Procurement Auctions Using Machine Learning: Data Analysis, Algorithm, and Case Study with Tenders from Spain”. В: *Complexity* 2020 (нояб. 2020). DOI: 10.1155/2020/8858258.

[4] Ahmet Soylu, Oscar Corcho, Brian Elvesæter, Carlos Badenes-Olmedo, Francisco Yedro, Matej Kovacic, Matej Posinkovic, Mitja Medvešček, Ian Makgill, Chris Taggart, Elena Simperl, Till Lech и Dumitru Roman. “Data Quality Barriers for Transparency in Public Procurement”. В: *Information* 13 (февр. 2022), с. 99. DOI: 10.3390/info13020099.

[5] Yeferson Torres Berrú и Vivian Batista. “Data Mining to Identify Anomalies in Public Procurement Rating Parameters”. В: *Electronics* 10 (нояб. 2021), с. 2873. DOI: 10.3390/electronics10222873.

[6] Yeongwook Yang, Hong Jun Jang и Byoungwook Kim. “A Hybrid Recommender System for Sequential Recommendation: Combining Similarity Models With Markov Chains”. В: *IEEE Access* 8 (2020), с. 190136—190146. URL: https://api.semanticscholar.org/CorpusID:226229866.