МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«БРЕСТСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
ФАКУЛЬТЕТ ЭЛЕКТРОННО-ИНФОРМАЦИОННЫХ СИСТЕМ
Кафедра интеллектуальных информационных технологий

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К КУРСОВОМУ ПРОЕКТУ ПО ДИСЦИПЛИНЕ
«Модели решения задач в интеллектуальных системах»
Тема: «Прогнозирование цен на жилье с использованием модели случайного леса»

КП.ИИ-21.210572-40 03-01

Листов: 13

Выполнил:
студент 4-го курса,
ФЭИС,
группы ИИ-21
Худик А.А.
Проверил:
Головко В.А.

Брест 2024

СОДЕРЖАНИЕ

1 ВВЕДЕНИЕ

В современном мире стремительного развития технологий обработки и анализа данных задача прогнозирования цен на жилье становится особенно актуальной. Это открывает новые возможности для использования методов машинного обучения, улучшения качества оценки объектов недвижимости и оптимизации процессов ценообразования. Прогнозирование цен играет ключевую роль для покупателей, продавцов, инвесторов и специалистов в сфере недвижимости, предоставляя им точные и обоснованные данные для принятия решений.

Создание моделей прогнозирования цен на жилье, таких как случайный лес, позволяет учитывать множество факторов, включая площадь, расположение, возраст здания и другие важные параметры. Использование библиотек, таких как TensorFlow, дает возможность разрабатывать сложные алгоритмы, способные анализировать большие объемы данных и обеспечивать высокую точность предсказаний. Такие подходы не только повышают уровень автоматизации процессов, но и способствуют более эффективному управлению недвижимостью.

Кроме того, внедрение моделей машинного обучения в анализ цен на жилье создает новые перспективы для развития интеллектуальных систем. Например, улучшение точности прогноза цен способствует созданию платформ, которые помогают пользователям оценивать рыночную стоимость объектов в реальном времени. Это делает рынок недвижимости более прозрачным и доступным для всех участников.

Применение моделей случайного леса в TensorFlow для прогнозирования цен на жилье является важным шагом к расширению использования искусственного интеллекта в повседневной жизни. Такие технологии позволяют значительно улучшить процесс принятия решений и способствуют созданию инновационных решений в области анализа и обработки данных.

2 ПОСТАНОВКА ЗАДАЧИ

Целью данной работы является разработка системы прогнозирования цен на жилье, основанной на методах машинного обучения. В качестве базовой модели будет использоваться алгоритм случайного леса, реализованный с помощью библиотеки TensorFlow, который будет обучаться на специализированном датасете, содержащем данные о недвижимости, включая такие параметры, как площадь, расположение, возраст здания и другие характеристики.

Задачи, которые необходимо решить в рамках проекта:

1. **подготовка набора данных:**
   Необходимо собрать и подготовить набор данных, содержащий параметры объектов недвижимости и их рыночную стоимость. Данные должны быть очищены от нерелевантной или некорректной информации. Важно учесть разнообразие характеристик, влияющих на стоимость, а также сбалансированность набора данных. Также требуется подготовить данные для последующей подачи в модель, включая нормализацию и обработку пропущенных значений.

2. **выбор архитектуры модели:**
   Проанализировать существующие реализации алгоритмов случайного леса и выбрать подходящую конфигурацию для решения задачи. Это включает в себя настройку таких параметров, как количество деревьев, глубина деревьев и критерии разбиения.

3. **обучение модели:**
   Обучить модель случайного леса на подготовленных данных, используя TensorFlow. Это позволит построить модель, способную учитывать нелинейные зависимости между характеристиками объектов и их ценой.

4. **оценка качества работы модели:**
   После обучения модели необходимо провести её тестирование на независимом наборе данных, используя метрики, такие как MSE (Mean Squared Error) и RMSE (Root Mean Squared Error). Анализ результатов поможет определить точность и надежность прогноза, а также выявить возможные направления для улучшения модели.

3 ВЫБОР И ОПИСАНИЕ ИСПОЛЬЗУЕМЫХ ИНСТРУМЕНТОВ

Для решения задачи прогнозирования цен на жилье с использованием случайного леса в TensorFlow были выбраны следующие инструменты и библиотеки:

- **TensorFlow:** основной фреймворк для разработки и обучения моделей машинного обучения. TensorFlow обеспечивает эффективное выполнение вычислений и предоставляет множество инструментов для создания сложных моделей, включая реализацию алгоритма случайного леса;
- **Seaborn:** библиотека для визуализации данных. Применялась для анализа и визуализации взаимосвязей между характеристиками недвижимости, что помогло выявить ключевые факторы, влияющие на стоимость жилья;
- **Pandas:** библиотека для работы с табличными данными. Использовалась для загрузки и предобработки набора данных, включая очистку, нормализацию и преобразование данных в формат, подходящий для обучения модели;
- **Matplotlib:** библиотека для построения графиков. Использовалась для визуализации распределений данных и анализа результатов предсказаний модели;

Эти библиотеки обеспечили эффективный процесс работы над проектом, включая предобработку данных, обучение модели случайного леса и визуализацию результатов. Такой набор инструментов является оптимальным для решения задачи прогнозирования цен на жилье.

4 ОБУЧЕНИЕ НЕЙРОННОЙ СЕТИ

В данном разделе описаны этапы обучения модели для прогнозирования цен на жилье с использованием случайного леса в TensorFlow, а также процесс подготовки данных.

4.1 Подготовка данных

Подготовка данных — один из ключевых этапов в процессе обучения модели. Для эффективного прогнозирования цен на жилье, на основе исходных данных о недвижимости, были выполнены следующие шаги:

- **скачивание данных:** для работы был использован набор данных, содержащий информацию о недвижимости, включая такие параметры, как площадь, количество комнат, возраст здания, расположение, наличие парковки и другие характеристики. Данные были загружены в формате CSV с помощью библиотеки Pandas. Пример данных:

Id	RoofStyle	YearBuilt	YearRemodAdd	LotFrontage	LotArea	Street	GarageType	GarageYrBlt	Alley
1	Gable	2003	2003	65	8450	Pave	Attchd	2003	NA
2	Gable	1976	1976	80	9600	Pave	Attchd	1976	NA
3	Gable	2001	2002	68	11250	Pave	Attchd	2001	NA
4	Gable	1915	1970	60	9550	Pave	Detchd	1998	NA
5	Gable	2000	2000	84	14260	Pave	Attchd	2000	NA
6	Gable	1993	1995	85	14115	Pave	Attchd	1993	NA
7	Gable	2004	2005	75	10084	Pave	Attchd	2004	NA
8	Gable	1973	1973	NA	10382	Pave	Attchd	1973	NA
9	Gable	1931	1950	51	6120	Pave	Detchd	1931	NA
10	Gable	1939	1950	50	7420	Pave	Attchd	1939	NA

- **очистка данных:** на этапе предварительной обработки были устранены пропущенные значения, некорректные или выбросные данные. Например, строки с некорректными значениями (отрицательная площадь или цена) были удалены, а пропуски заменены на медианные значения по соответствующему признаку;
- **анализ данных:** с использованием библиотек Seaborn и Matplotlib был проведен анализ данных, чтобы определить взаимосвязи между характеристиками объектов и их ценой. Построены диаграммы корреляции, распределения и ящичные диаграммы для выявления ключевых факторов;
- **разделение на тренировочную и тестовую выборки:** данные были разделены на тренировочную и тестовую выборки. Это разделение обеспечило независимость тестовых данных, необходимых для оценки качества модели;

4.2 Выбор архитектуры нейронной сети

Для задачи прогнозирования цен на жилье была выбрана модель случайного леса, реализованная с использованием *TensorFlow Decision Forests*. Этот выбор был обусловлен следующими преимуществами:

- **Обработка табличных данных:** случайный лес является одним из наиболее подходящих алгоритмов для работы с табличными данными, содержащими числовые и категориальные признаки.
- **Интерпретируемость:** благодаря структуре модели можно легко анализировать вклад каждого признака в итоговое предсказание, что особенно важно в задачах прогнозирования цен.
- **Устойчивость к переобучению:** модель случайного леса хорошо справляется с задачами, где признаки могут быть скоррелированы или содержать выбросы.

Характеристики модели

- **Архитектура дерева:** модель состоит из нескольких деревьев решений, объединенных в ансамбль. Каждое дерево строится на случайной подвыборке данных, а разбиения узлов деревьев основаны на случайно выбранных подмножествах признаков. Это позволяет снизить вероятность переобучения и повысить обобщающую способность.
- **Глубина деревьев:** максимальная глубина деревьев была ограничена, чтобы избежать переобучения, но при этом позволить модели улавливать сложные зависимости в данных.
- **Количество деревьев:** было выбрано оптимальное количество деревьев, обеспечивающее баланс между временем обучения и точностью предсказаний.

Функция активации и методы обучения

Хотя модель случайного леса не использует нейронные сети, структура обучения аналогична некоторым аспектам глубокого обучения:

- **Активация:** для преобразования признаков и их влияния на конечное предсказание модель использует нелинейные разбиения в узлах деревьев, аналогичные нелинейностям в нейронных сетях.
- **Оптимизация гиперпараметров:** для достижения максимальной точности была проведена оптимизация гиперпараметров модели (глубина деревьев, количество деревьев, размер подвыборки данных).

Использование случайного леса в сочетании с *TensorFlow Decision Forests* позволяет интегрировать классический подход с преимуществами современного фреймворка машинного обучения, что обеспечивает высокую производительность модели при прогнозировании цен на жилье.

4.3 Обучение нейронной сети

1. **Создание случайных подвыборок (bootstrap)** Для каждого дерева в ансамбле формируется случайная подвыборка данных из обучающей выборки с возвращением. Это означает, что одни и те же данные могут попасть в подвыборку несколько раз, а некоторые данные могут вовсе не быть выбраны. Такой подход снижает корреляцию между деревьями, так как каждое дерево видит немного разные данные.
2. **Выбор случайного набора признаков** На каждом этапе разбиения дерева выбирается случайное подмножество признаков для поиска оптимального условия разделения. Параметр `max_features` управляет размером этого набора:
3. **Построение дерева** Дерево обучается на своей подвыборке, постепенно разделяя данные на узлах:
   (a) **выбор условия разделения:** для каждой вершины рассматриваются все доступные признаки из случайного подмножества. Для каждого признака находятся пороговые значения, минимизирующие ошибку (например, MSE).
   (b) **разделение:** данные разделяются на две группы (левую и правую ветви) в соответствии с выбранным условием. Этот процесс повторяется до достижения заданной глубины дерева (`max_depth`) или пока в листьях не останется минимальное число объектов (`min_samples_leaf`).
4. **Формирование прогнозов на листьях** Когда дерево достигает листа, оно фиксирует прогноз для объектов, попавших в этот лист. Прогнозом может быть среднее значение целевой переменной для объектов в листе.
5. **Повторение процесса для всех деревьев** Каждый этап (создание подвыборки, выбор признаков, построение дерева) повторяется для каждого дерева в ансамбле. В результате получается множество независимых деревьев, обученных на различных подмножествах данных.
6. **Комбинирование прогнозов** После обучения каждого дерева их прогнозы сохраняются для дальнейшего усреднения. Такой подход снижает вероятность переобучения, поскольку каждое дерево вносит свой уникальный вклад в итоговый результат.

5 ТЕСТИРОВАНИЕ НЕЙРОННОЙ СЕТИ

В данном разделе рассматриваются этапы тестирования обученной модели для предсказания цен на недвижимость с использованием алгоритма Random Forest. Основной целью тестирования является оценка корректности работы модели, её точности и эффективности. Тестирование включало следующие этапы:

- **Функциональное тестирование**: проверка работы модели на тестовых данных для анализа её способности правильно предсказывать цену недвижимости. Модель должна демонстрировать способность точно и стабильно работать на различных примерах данных, не участвующих в процессе обучения.
- **Оценка качества модели**: после обучения модель тестируется на отложенной выборке. Оценка качества производится с использованием нескольких метрик, которые позволяют получить полное представление о её эффективности. К основным метрикам можно отнести:
  - **MSE (Mean Squared Error)** — среднеквадратичная ошибка, которая измеряет среднее квадратичное отклонение предсказанных значений от истинных. Это основная метрика для задач регрессии, показывающая, насколько хорошо модель предсказывает целевые значения.
  - **R^2 (коэффициент детерминации)** — метрика, которая показывает, какую часть вариации зависимой переменной объясняет модель. Чем выше значение R^2, тем лучше модель предсказывает данные.

  Также строится график изменений ошибки на обучающих и валидационных данных по мере увеличения числа деревьев в модели, что позволяет оценить динамику улучшения точности.
- **Тестирование производительности**: проверка времени выполнения модели на тестовом наборе данных, а также анализ её эффективности при обработке больших объёмов данных. Оценка производительности важна для понимания, насколько быстро модель может генерировать предсказания на реальных данных в условиях реального времени.

В результате тестирования модель показала хорошую общую точность, особенно при предсказании цен для более распространённых типов недвижимости. Основные ошибки были связаны с домами, которые имели нетипичные характеристики или сильно отличались от представленных в обучающих данных. Для повышения точности в таких случаях планируется дополнительно расширить и улучшить данные, а также провести настройку гиперпараметров модели.

6 ЗАКЛЮЧЕНИЕ

В ходе выполнения курсового проекта была достигнута основная цель — разработана система предсказания цен на жилье с использованием алгоритмов деревьев решений, в частности, модели случайного леса (Random Forest) на базе TensorFlow Decision Forests. Основное внимание было уделено подготовке специализированного датасета, обучению модели и оценке её точности на основе различных метрик.

В процессе работы была реализована модель, которая эффективно предсказывает цену на жилье, используя различные характеристики, такие как площадь, количество комнат, состояние недвижимости и другие. Обучение проводилось на подготовленных данных с использованием метрик средней квадратичной ошибки (MSE), что позволило всесторонне оценить её эффективность. Проведённая настройка гиперпараметров и архитектуры модели позволила достичь высокой точности предсказания.

Использованные методы и технологии продемонстрировали потенциал деревьев решений для решения задач регрессии в области предсказания цен. Разработанная система может быть применена для оценки рыночной стоимости недвижимости, автоматизации процессов анализа рынка жилья и оптимизации принятия решений в сфере недвижимости.

Результаты работы показывают перспективность применения алгоритмов деревьев решений в задачах предсказания и регрессии. Данный проект открывает новые возможности для улучшения точности прогноза в области недвижимости, а также для внедрения таких технологий в реальный бизнес.

7 Список использованной литературы

1. H. W. Tan, W. P. et al. *TensorFlow Decision Forests: A Library for Decision Forests with TensorFlow* [Электронный ресурс]. – Режим доступа: `https://www.tensorflow.org/decision_forests`. – Дата доступа: 19.11.2024.
2. Guo, Y., et al. *The Basics of Decision Trees in Machine Learning* [Электронный ресурс]. – Режим доступа: `https://towardsdatascience.com/the-basics-of-decision-trees-in-machine-learning-38c3e2080f9d`. – Дата доступа: 19.11.2024.
3. Kaggle. *Housing Prices: Advanced Regression Techniques* [Электронный ресурс]. – Режим доступа: `https://www.kaggle.com/c/house-prices-advanced-regression-techniques`. – Дата доступа: 19.11.2024.
4. W. G. Gilks. *Decision Trees for Regression and Classification* [Электронный ресурс]. – Режим доступа: `https://machinelearningmastery.com/decision-trees-for-regression-and-classification/`. – Дата доступа: 17.11.2024.
5. TensorFlow. *TensorFlow Documentation* [Электронный ресурс]. – Режим доступа: `https://www.tensorflow.org/docs`. – Дата доступа: 19.11.2024.