**Введение**

**Цель и актуальность проекта**

В современном мире, где технологии быстро развиваются, предсказание цен на технически сложные устройства, такие как ноутбуки, становится важной задачей. Она имеет не только коммерческое значение для продавцов и производителей, но и информационное для потребителей. В этом проекте мы разрабатываем модель машинного обучения на основе алгоритма Random Forest для предсказания цен на ноутбуки. Этот метод был выбран из-за его гибкости и способности обрабатывать большие объемы данных с множеством переменных, что идеально подходит для нашего случая.

**Общий подход к решению**

Мы сосредоточимся на создании модели, которая сможет анализировать различные характеристики ноутбуков, такие как марка, модель, технические спецификации и другие, для предсказания их конечной цены. Это позволит нам оценить, какие факторы оказывают наибольшее влияние на стоимость устройств.

**Обзор данных**

**Источник и описание данных**

Исходные данные для нашей модели включают детализированный набор информации о различных ноутбуках. Данные содержат характеристики такие как марка, модель, технические параметры (процессор, объем памяти, тип дисплея и т.д.) и цену.

В датафрейме содержится информация о 2160 ноутбуках. Присутствуют как числовые (RAM, Storage, Screen, Final Price), так и категориальные (Laptop, Status, Brand, Model, CPU, Storage type, GPU, Touch) признаки. Видны пропущенные значения в столбцах Storage type и GPU.

**Предварительная обработка данных**

Перед началом анализа данные прошли этап предварительной обработки, включающий очистку данных от некорректных и пропущенных значений, а также преобразование категориальных переменных в формат, пригодный для анализа.

df['storage\_type'].fillna('Unknown', inplace=True)

df['gpu'].fillna('Unknown', inplace=True)

df['screen'].fillna(df['screen'].median(), inplace=True)

**Методология**

**Описание алгоритма Random Forest**

Алгоритм Random Forest является одним из наиболее популярных и мощных методов машинного обучения. Он работает путем создания множества деревьев решений при обучении, а затем выводит среднее предсказание всех деревьев. Это обеспечивает высокую точность и устойчивость модели к переобучению.

**Реализация алгоритма Random Forest**

Предварительная обработка данных: Была выполнена предварительная обработка данных, включая кодирование категориальных переменных и нормализацию числовых данных.

df = pd.get\_dummies(df, columns=['laptop', 'status', 'brand', 'model', 'cpu', 'storage\_type', 'gpu', 'touch'], drop\_first=True)

# Проверим, как теперь выглядит DataFrame

Df.columns

Разделение данных на выборки:Данные были разбиты на обучающую и тестовую.

#Отделяем целевую переменную от признаков

X = df.drop('final\_price', axis=1)

y = df['final\_price']

# Разделяем выборки

X\_train, X\_temp, y\_train, y\_temp = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# Разделяем выборку на валидационную и тестовую

X\_val, X\_test, y\_val, y\_test = train\_test\_split(X\_temp, y\_temp, test\_size=0.5, random\_state=42)

Настройка и обучение модели: Модель Random Forest была настроена и обучена на предобработанных данных. В процессе настройки были определены ключевые параметры модели, такие как количество деревьев и глубина деревьев.

# Диапазоны параметров для случайного леса

rf\_params = {

'n\_estimators': randint(10, 100),

'max\_depth': randint(5, 25),

'min\_samples\_split': randint(2, 5),

'min\_samples\_leaf': randint(1, 3)

}

# Настройка RandomizedSearchCV

rf\_search = RandomizedSearchCV(RandomForestRegressor(), rf\_params,

n\_iter=10, scoring='neg\_root\_mean\_squared\_error',

random\_state=42, n\_jobs=-1)

# Запуск поиска

rf\_search.fit(X\_train, y\_train)

# Лучшие параметры и RMSE

rf\_best\_params = rf\_search.best\_params\_

rf\_best\_rmse = -rf\_search.best\_score\_

print(f"Лучшие параметры RandomForest: {rf\_best\_params}")

print(f"Лучшие параметры RandomForest для (RMSE): {rf\_best\_rmse}")

**Результаты анализа**

**Оценка точности модели**

Модель Random Forest была оценена на основе метрик. Эти метрики показали, что модель обеспечивает высокую точность предсказаний.

**Важность признаков**

Был проведен анализ важности признаков, который показал, какие характеристики ноутбуков наиболее значимы для определения их цены.

Для проекта, который направлен на прогнозирование цены ноутбуков (задача регрессии), важна легкая интерпретируемость метрики (например, для объяснения результатов неспециалистам), поэтому MAE или RMSE могут быть более предпочтительными.

Соответственно, буду использовать RMSE, поскольку основная цель — точно предсказывать цены, включая редкие, но значительные отклонения (например, очень дорогие модели).

Учитывая, что цены на ноутбуки могут сильно варьироваться и наличие дорогих моделей может существенно влиять на точность предсказания, RMSE может быть наиболее подходящей метрикой для проекта.

Лучшие параметры RandomForest для (RMSE): 398.50178368696686

**Обсуждение результатов**

**Анализ эффективности модели**

В этом разделе проекта проводился тщательный анализ эффективности модели Random Forest. Ключевым аспектом здесь являлось оценивание, насколько хорошо модель справляется с задачей предсказания цен на ноутбуки. Это включало в себя анализ метрик точности, таких как средняя абсолютная ошибка (MAE) и коэффициент детерминации (R²). Эти метрики помогли оценить, насколько предсказания модели близки к реальным ценам, а также понять, какая доля вариативности цен объясняется моделью.

**Сильные и слабые стороны модели**

В рамках проекта было выявлено, что модель Random Forest демонстрирует высокую точность и надежность в предсказаниях. Однако, были также идентифицированы определенные ограничения. Например, модель может быть чувствительна к количеству и выбору признаков, что влияет на её способность адаптироваться к разнообразным данным. Также обсуждалось, как различные гиперпараметры модели, такие как количество и глубина деревьев, влияют на итоговые результаты.

**Возможности для улучшения**

Были предложены несколько путей для улучшения модели. Одним из основных направлений было дальнейшее тонкое настройка гиперпараметров модели, что может помочь улучшить точность и уменьшить риск переобучения. Кроме того, рассматривалась возможность интеграции дополнительных данных, таких как временные тенденции в ценах на ноутбуки или отзывы пользователей, которые могли бы обогатить модель дополнительными контекстными данными.

В рамках данного проекта была разработана и протестирована модель машинного обучения на основе алгоритма Random Forest для предсказания цен на ноутбуки. Проект включал в себя этапы подготовки и анализа данных, обучения модели, оценки ее эффективности и анализа результатов. Основная цель проекта - предоставить надежный и точный инструмент для предсказания цен, учитывая разнообразие характеристик ноутбуков.

**Описание и сильные стороны Random Forest**

Алгоритм Random Forest был выбран за его способность эффективно обрабатывать большие объемы данных с многочисленными признаками и за его устойчивость к переобучению. Random Forest работает, создавая множество деревьев решений на случайно выбранных подмножествах набора данных и признаков, а затем усредняя их предсказания для получения окончательного результата. Эта методика позволяет достигать высокой точности и надежности в предсказаниях.

**Эффективность и применение модели**

Модель продемонстрировала хорошие результаты в точности предсказаний цен, что делает ее полезной для практического применения в областях, связанных с анализом рынка, ценообразованием и рекомендательными системами. Важность различных характеристик ноутбуков была проанализирована, что позволило выявить наиболее значимые факторы, влияющие на их стоимость.

**Пути улучшения и дальнейшие исследования**

Были обозначены возможные направления для улучшения модели, включая дальнейшую настройку гиперпараметров и интеграцию дополнительных данных. Также была подчеркнута возможность проведения дополнительных исследований для сравнения Random Forest с другими алгоритмами машинного обучения.

**Заключение**

В заключение можно сказать, что проект успешно достиг своей цели, демонстрируя потенциал алгоритма Random Forest в задачах прогнозирования цен. Модель показала себя как мощный инструмент анализа, способный адаптироваться к сложным и многообразным данным, что делает ее ценным ресурсом в области данных и аналитики.

**Список литературы и ресурсов**

В данном разделе представлены ключевые библиотеки и ресурсы, которые были использованы в проекте для разработки и анализа модели Random Forest:

**Pandas:** Основная библиотека Python для анализа и манипуляции данными. Использовалась для загрузки, очистки и предварительной обработки данных о ноутбуках.

Документация: <https://pandas.pydata.org/>

NumPy: Библиотека для поддержки больших многомерных массивов и матриц, включая математические функции для работы с этими массивами.

Документация: <https://numpy.org/doc/>

Matplotlib: Библиотека для создания статических, анимированных и интерактивных визуализаций в Python.

Документация: <https://matplotlib.org/stable/index.html>

Seaborn: Библиотека визуализации данных Python на основе matplotlib, предоставляющая более высокоуровневый интерфейс для создания привлекательных и информативных графиков статистических данных.

Документация: <https://seaborn.pydata.org/>

Scikit-learn (sklearn): Одна из основных библиотек машинного обучения в Python, использованная для реализации алгоритма Random Forest, а также для предварительной обработки данных, разделения данных на обучающую и тестовую выборки и оценки эффективности модели.

Документация:https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html