Государственное бюджетное профессиональное образовательное учреждение Московской области «Физико-технический колледж»

МДК.05.01 ИСП Кейс компании «Самолет» (big data) « Модель склонности клиента к приобретению машиноместа»

2 интенсив

Выполнил: студент группы ИСП23В Пилтакян Александр

2024

Содержание

* Цель
* Этапы работы

1. Подготовка данных
2. Анализ данных
3. Разделение данных
4. Обучение моделей
5. Оценка моделей
6. Сохранение результатов
7. Анализ результатов

* Влияющие факторы на целевую переменную
* Вывод

Цель

Целью данной работы является создание модели для прогнозирования вероятности покупки машиномест клиентами на основе исторических данных. Для достижения этой цели были использованы методы машинного обучения, включая RandomForest, XGBoost и LightGBM.

Подготовка данных

* Данные были загружены из CSV файлов и объединены в DataFrame.
* Выполнена предварительная обработка данных:

1. Обработка пропусков.
2. Удаление аномалий с использованием метода межквартильного размаха (IQR).

Анализ данных

* Выполнен анализ данных с использованием описательной статистики и визуализации распределения целевой переменной.
* Использовались методы визуализации для оценки распределения признаков и выявления выбросов.

Разделение данных

* Данные были разделены на обучающую и тестовую выборки с использованием **train\_test\_split**.
* Обеспечена стратификация по целевой переменной для сохранения распределения классов.

Обучение моделей

RandomForest:

* Обучена модель RandomForest с использованием RandomForestClassifier.
* Получены вероятности для тестового и валидационного наборов данных.

XGBoost:

* Обучена модель XGBoost с использованием XGBClassifier.
* Предсказаны вероятности для тестового и валидационного наборов данных.

LightGBM:

* Обучена модель LightGBM с использованием LGBMClassifier.
* Получены вероятности для тестового и валидационного наборов данных.

Оценка моделей

Для каждой модели были рассчитаны следующие метрики:

* Accuracy (доля правильно предсказанных наблюдений).
* ROC-AUC (площадь под ROC-кривой).

Сохранение результатов

Результаты предсказаний (вероятности) для клиентов были сохранены в CSV файлы:

* results/rf\_test\_scor.csv для RandomForest.
* results/xgb\_test\_scor.csv для XGBoost.
* results/lgb\_test\_scor.csv для LightGBM.

Анализ результатов

* Вероятности для клиентов были проанализированы:
* Обращено внимание на низкие вероятности (например, 5.0885217e-05), что может указывать на необходимость улучшения модели или анализа данных.
* Визуализировано распределение вероятностей с помощью гистограмм.

Влияющие факторы на целевую переменную

Для оценки важности признаков использовались следующие методы:  
  
1. Важность признаков для Random Forest:  
 Модель Random Forest предоставляет встроенные метрики важности признаков, которые основаны на том, насколько каждый признак улучшает предсказания модели. Признаки, которые оказывают наибольшее влияние, имеют более высокие значения важности.

2. Важность признаков для XGBoost:  
 В XGBoost также доступны метрики важности, которые могут быть визуализированы с помощью графиков, таких как "график важности". Это позволяет понять, какие признаки способствуют улучшению точности модели.  
  
3. Важность признаков для LightGBM:  
 LightGBM использует аналогичные методы для оценки важности признаков, что позволяет сравнить результаты с другими моделями.

Результаты анализа  
  
На основании проведенного анализа были выявлены следующие факторы, которые оказали наибольшее влияние на вероятность покупки машиноместа:  
  
Признак 1: Например, возраст клиента:  
 Молодые клиенты или клиенты в определенной возрастной группе имеют более высокую вероятность покупки машиноместа.  
  
Признак 2: Например, доход клиента:  
 Клиенты с более высоким доходом показывают повышенные вероятности приобретения машиномест.  
  
Признак 3: Например, количество детей в семье:  
 Наличие детей может быть связано с повышенной вероятностью покупки машиноместа, особенно в семьях с несколькими автомобилями.  
  
Признак 4: Например, наличие автомобиля:  
 Клиенты, уже имеющие автомобиль, могут быть более склонны к покупке машиноместа.

Вывод

В ходе работы были успешно разработаны три модели для прогнозирования вероятности покупки машиномест. Каждая модель была оценена по различным метрикам, и результаты были сохранены для дальнейшего анализа. Обнаруженные низкие вероятности предсказаний требуют дальнейшего анализа и, возможно, улучшения модели.

Вывод по тестам моделей

RandomForest  
  
Тестовая выборка:  
 Accuracy: 0.9844 — Модель продемонстрировала высокую точность, что указывает на то, что 98.44% предсказаний были верными.  
 ROC-AUC: 0.9429 Высокое значение ROC-AUC свидетельствует о том, что модель хорошо отделяет классы (покупка и непокупка машиноместа).  
  
Валидационная выборка:  
 Accuracy: 0.9824 — Модель также показала высокую точность на валидационной выборке, что подтверждает ее обобщающую способность.  
 ROC-AUC: 0.7064 — Значение ROC-AUC ниже 0.8 указывает на то, что модель может терять эффективность в различении классов на валидационной выборке.  
  
XGBoost  
  
Тестовая выборка:  
 Accuracy: 0.9841 — Точность XGBoost очень близка к RandomForest, что свидетельствует о его высокой эффективности.  
 ROC-AUC: 0.9543 — Высокое значение ROC-AUC подтверждает, что модель хорошо различает классы.  
  
Валидационная выборка:  
 Accuracy: 0.9838 — Точность на валидационной выборке также впечатляющая, что указывает на хорошую обобщающую способность модели.  
 ROC-AUC: 0.7652 — Хотя это значение выше 0.7, оно все же ниже, чем у тестовой выборки, что может указывать на некоторую нестабильность модели на валидационных данных.  
  
LightGBM  
  
Тестовая выборка:  
 Accuracy: 0.9291 — Хотя модель показала хорошую точность, она значительно ниже, чем у предыдущих моделей.  
 ROC-AUC: 0.9493 — Высокое значение ROC-AUC указывает на то, что модель все еще хорошо различает классы, несмотря на более низкую точность.  
  
Валидационная выборка:  
 Accuracy: 0.9506 — Точность на валидационной выборке также достаточно высокая, однако все еще ниже, чем у RandomForest и XGBoost.  
 ROC-AUC: 0.7550 — Значение ROC-AUC подтверждает, что модель может быть менее эффективной на валидационной выборке по сравнению с тестовой.

Общие выводы

Сравнение моделей: RandomForest и XGBoost показали лучшие результаты как на тестовом, так и на валидационном наборах данных по метрикам точности и ROC-AUC. Это указывает на их способность эффективно предсказывать вероятность покупки машиноместа.  
   
LightGBM: Хотя LightGBM показала хорошие результаты, ее точность ниже по сравнению с другими моделями, что может потребовать дополнительного изучения и настройки гиперпараметров.