ДИПЛОМНЫЙ ПРОЕКТ

Автор работы:
Димитрий Гуськов

Ментор:
Дмитрий Крылов

ТЕМА

Страхование автомобилей африканской компанией AutoInLand

ОПИСАНИЕ ПРОБЛЕМЫ

В процессе страхования у африканской страховой компании AutoInLand имеется слишком мало точек соприкосновения с клиентами.

Компания считает, что для достижения более высоких стандартов уровня обслуживания им необходимо предвидеть будущие потребности с точки зрения объема запросов на возмещение убытков.

Цель этой работы - разработать прогностическую модель, которая определяет, подаст ли клиент заявку на страхование транспортного средства в ближайшие три месяца.

ПЛАН РАБОТЫ

Данная работа состоит из следующих частей:

Часть 1. Введение*

Часть 2. Построение базовой модели

Часть 3. Анализ данных

Часть 4. Заполнение пустых значений

Часть 5. Генерация признаков

Часть 6. Подготовка данных для построения модели

Часть 7. Эксперимент 1

Часть 8. Эксперимент 2

Часть 9. Эксперимент 3

Часть 10. Эксперимент 4

Часть 11. Восстановление лучшего результата

Часть 12. Кросс валидация

Часть 13. Стороннее решение

ИНФОРМАЦИЯ О ДАННЫХ

12079 НАБЛЮДЕНИЙ

14 ПРИЗНАКОВ

1 ID: номер записи

2 Policy Start Date: Начало действия страхового полиса

3 Policy End Date: Конец действия страхового полиса

4 Gender: Пол

5 Age: Возраст

6 First Transaction Date: День первой транзакции

7 No_Pol: неизвестно

8 Car_Category: Категория авто

9 Subject_Car_Colour: Цвет авто

10 Subject_Car_Make: Марка авто

11 LGA_Name: Название города

12 State: Название штата

13 ProductName: Название продукта (авто)

14 target: целевая переменная

ПОСТРОЕНИЕ БАЗОВЫХ МОДЕЛЕЙ

В качестве базовых моделей были выбраны DecisionTree и CatBoost.

Значительная предобработка не проводилась. Пустые данные были заполнены значением no_value, дубликаты были устранены.

Метрики моделей:

PROBS THRESHOLD = 0.2

	DecisionTree
F SCORE	0.20093
ROC AUC	0.71369

	CatBoost
F SCORE	0.17142
ROC AUC	0.82728

О РЕЗУЛЬТАТАХ АНАЛИЗА ДАННЫХ

1.

Наиболее часто встречаемые значения:

- Начала страховки: 2010 год → 13267 (99% от данных).
- Конца страховки: 2011 год → 13096 (99% то данных).

2.

- Количество мужчин составляет 65 % + от данных. Женщин около 30%.

3.

- В признаке возраста изначально присутствовали отрицательные значения. Они были устранены.

- После обработки разброс составляет от 20 до 60 лет. Медианное значение 41.

4.

- Наиболее 3 популярных цвета: черный (~2000 машин), серебристый (~ 600 машин), серый (~ 575 машин).

5.

- Самые распространенные авто: Toyota (~ 5400 авто), Honda (~ 1100 авто), Lexus (~ 800 атво).

6.

- Первые 3 самых распространенных штата: Lagos (~ 3500), Benue (~ 600), Abuji-Manucipal (~ 200).

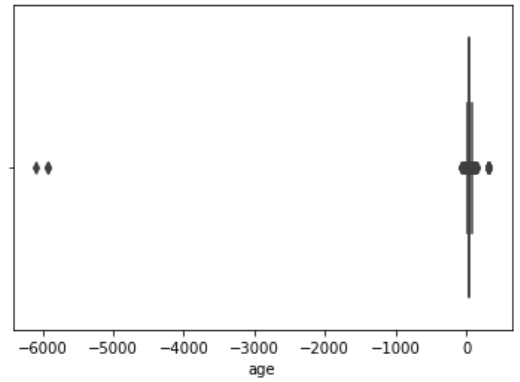
7.

- Самые распространенные тип авто: Car Classis (~ 7000), CarSafe (~ 4100).

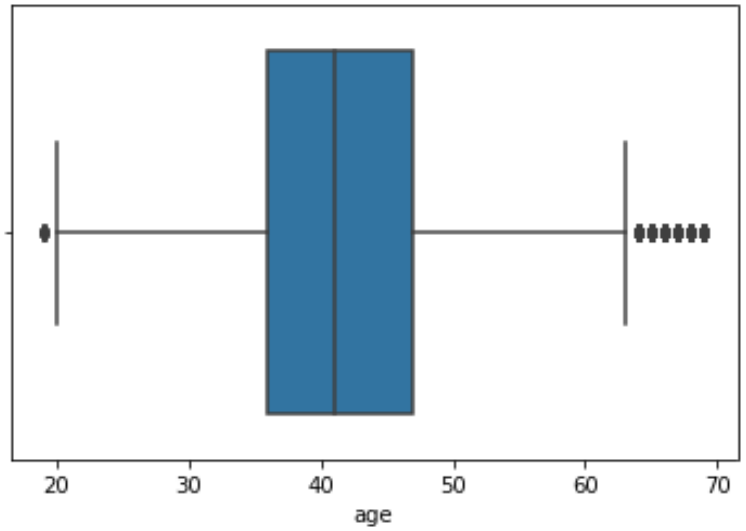
8.

- Распределение target переменной: 1 – 20%, 0 -80%.

```
IQR 15.0

 Uniqie Outliers

array([   79,     2,   120,    81,     1,    11,    78,    12,    73,
          82,     7,    10,    76,     3,    93,    77,    80,     9,
           6,   -76,    -2,    89,    74,    75,    84,    90,   140,
           8,     5,    86,   320,   -12,    83,    85, -6099, -5939,
           4,   -22,   144,   112,     0,    88,   -27,    87,   133,
         102,   -51,   128,   -26,   100], dtype=int64)
```




```
IQR 11.0

 Uniqie Outliers

array([69, 64, 65, 67, 68, 66, 19], dtype=int64)
```



ОБРАБОТКА ДАННЫХ

1. В признаке no_pol взяты только значения от 1 до 5 включительно (из 10 бальной шкалы)
2. В признаке категории авто Saloon изменен на Sedan
3. В признаке цвета авто все оттенки серого заменены на значение Grey
4. Также в признаке цвета значение D. Red исправлено на Dark Red
5. В признаке марок авто значения переписаны в соответствующий тип авто: Truck, Sedan, Ford, Land Rover, Jeep
6. В признаке марок авто также исправлены ошибки в названиях записей и переведены, либо в Sedan, либо в Motorcycle
7. С помощью модели KNN Imputer заполнены пустые значения
8. Созданы мат. признаки A+B, A*B
9. Данные пола закодированы с помощью OneHotEncoder
10. Из дат получены значения годов и длительности страховки

ЭКСПЕРИМЕНТ 1

В данном эксперименте были использованы данные с доп. уникальной обработкой №1.

- MeanTargetEncoding для категориальных переменных.
- KNN Imputer для пустых значений, а также построены 5 моделей.

PROBS THRESHOLD = 0.2

CatBoost

F SCORE	0.39408
ROC AUC	0.81158

Decision Tree

F SCORE	0.19343
ROC AUC	0.59612

RandomForest

F SCORE	0.19953
ROC AUC	0.77178

LGBMClassifier **ЛУЧШИЙ РЕЗУЛЬТАТ**

F SCORE	0.39999
ROC AUC	0.98539

XGBoostClassfier

F SCORE	0.37735
ROC AUC	0.80532

ЭКСПЕРИМЕНТ 2

В данном эксперименте были использованы данные с доп. уникальной обработкой №2.

- Dummies для категориальных переменных.
- KNN Imputer для пустых значений,

а также построена лучшая модель предыдущего эксперимента

PROBS THRESHOLD = 0.2

LGBMClassifier

F SCORE 0.07472

ROC AUC 0.75981

ЭКСПЕРИМЕНТ 3

В данном эксперименте были использованы данные с доп. уникальной обработкой №3.

- MeanTargetEncoding для категориальных переменных.
- MODE для пустых значений,

а также построена лучшая модель предыдущего эксперимента

PROBS THRESHOLD = 0.2

LGBMClassifier

F SCORE	0.07472
ROC AUC	0.75981

ЭКСПЕРИМЕНТ 4

В данном эксперименте были использованы данные с доп. уникальной обработкой №4.

- MeanTargetEncoding для категориальных переменных.
- KNN Imputer для пустых значений,

а также построены 5 моделей.

PROBS THRESHOLD = 0.2

CatBoost

F SCORE	0.18149
ROC AUC	0.80058

Decision Tree

F SCORE	0.17021
ROC AUC	0.57739

RandomForest

F SCORE	0.18390
ROC AUC	0.75295

LGBMClassifier

F SCORE	0.15105
ROC AUC	0.80613

XGBoostClassfier **ЛУЧШИЙ РЕЗУЛЬТАТ**

F SCORE	0.18941
ROC AUC	0.79231

Результаты экспериментов + CrossValidation

Лучшим результатом оказалось использование модели LGBMClassifier и обработка данных с MeanTargetEncoding + KNN Imputer.

PROBS THRESHOLD = 0.2

F SCORE = 0.39999

Была проведена CrossValidation, однако она не дала желаемых результатов.

F SCORE = 0.15789

НАИБОЛЕЕ ВАЖНЫЕ ПРИЗНАКИ

LGBMClassifier с лучшим SCORE

1. LGA_NAME (510) - название города
2. AGE (468) - возраст
3. F_D (414) - первый день страхования
4. SUBJECT_CAR_COLOUR (327) - цвет
5. F_M (322) - первый месяц страхования
6. STATE (312) - штат
7. SUBJECT_CAR_MAKE (258) – марка авто
8. PRODUCT_NAME (172) – название продукта
9. NO_POL (88) - *
10. GENDER_MEAN (69) – ср. знач. возраста

Другие...

AUTO ML

Последним решением было использовать AutoML. Данное решение взято из внешних источников

Загружая финальные предсказания на соревнование модель AUTO ML показала результат в 0.18 по метрике F SCORE.

Предсказания лучшегр решения в ноутбуке с LGBMClassifier продемонстрировали на соревновании значение 0.10 по метрике F SCORE

ИТОГИ

Итоговый Вывод

Лучшую оценку дала модель AutoML

0.186046.
Другие модели показали немного худший результат на соревновании.

Безусловно, этого недостаточно, однако с учетом проделанной работы и опробованных методов следует сделать вывод, что данные являются нерепрезентативными- При построении модели и решении задачи нельзя исключать эмпирические знания, связанные с объектом исследования при сборе данных.

Основная цель состояла в урегулирование страховых исков и того, подаст ли клиент заявку на страхование транспортного средства в ближайшие три месяца.

Из полезных признаков изначально имелось:
Начало и конец страховки, пол, возраст и тип автомобиля. Остальные данные не несут в себе большой пользы.

Для более точного решения задачи необходимо собрать следующие данные:
- **доход** (личный), **доход** (семейный), ***состояние в браке***, ***число людей в семье***, ***наличие детей***, ***кредитная история***, ***длительность работы***, ***финансовая историю клиента*** (здесь будет полезна любая информация), ***стаж вождения*** (одна из обязательных и ключевых метрик, странно, что она отсутствовала).