Перевышина Татьяна, 2019

GitHub: https://github.com/prvshna/kaggle

Kaggle: https://www.kaggle.com/c/orders-prediction/overview

Kaggle[1] «Burned orders prediction»

предсказание отмены заказа в «Яндекс.Такси»

Целью данного Kaggle-чемпионата являлось предсказание отмены «Яндекс.Такси». Этот показатель интересен компании, так как оказывает непосредственное влияние на прибыль. Объектом исследования являлся рынок такси, а предметом характеристики заказа, влияющие на факт его отмены. В данном проекте мной было поставлено три задачи:

- анализ основных статистических трендов работы «Яндекс.Такси»;
- формирование новых признаков для модели, основанных на анализе трендов;
- обучение модели, предсказывающей отмену заказа.

Выборка исследования состояла из 1 793 300 наблюдений. Обучающая часть выборки состояла из 80% заказов (1 450 000), а тестовая соответственно из 20% (343 300). Это разделение было сделано для того, чтобы избежать переобучения. Базовый набор данных содержал информацию о дате и времени заказа, классе автомобиля, широте и долготе места вызова такси. При этом базовый набор определял отмену заказа с точностью не более 60%.

В рамках этого проекта были исследованы следующие гипотезы:

1. Местоположение заказа оказывает влияние на вероятность отмены такси.
2. Близость заказа к аэропорту или вокзалу оказывает влияние на вероятность отмены такси.
3. День недели/время суток заказа оказывает влияние на вероятность отмены такси.
4. Класс автомобиля оказывает влияние на вероятность отмены такси
5. Выявление «ложных заказов» в выборке оказывает влияние на точность определения отмены заказа.

На основе данных гипотез были сформированы признаки для предсказания на обучающей, а затем на тестовой выборке. Нужно отметить, что многочисленность выборки позволяла вводить большое число категориальных переменных, без влияния на устойчивость результатов, что я и осуществила в рамках проекта.

[1] Краудсорсинг платформа для людей, которым интересен анализ данных и машинное обучение

Местоположение

Исследования Яндекс показали, что большинство заказов делается из центра города.[2] В связи с этим был сформирован фактор местоположения, основанный на долготе и широте заказа. Координаты были спроецированы на трехмерную плоскость через синус и косинус.

```
train['cossin'] =
np.cos(train['lon'])*np.sin(train['lat'])
train['sincos'] =
np.sin(train['lon'])*np.cos(train['lat'])
train['cos'] = np.cos(train['lat'])
```

Рисунок 1 – Проекция координат на трехмерную плоскость

Такой способ представления местоположения возможен благодаря сферической форме Земли. В результате данного преобразования ROC AUC составил 61%, соответственно рост по сравнению с базовым набором признаков составил 1%

Близость к аэропорту или вокзалу

Особая специфика у заказов, совершаемых из аэропортов и вокзалов. Исходя их того, что эти места имеют несколько повышенный спрос на такси для них были созданы отдельные категориальные переменные, основанные также на широте и долготе.[3] Понятие «близость» предполагает вызов такси не более чем за 400 метров от аэропорта или вокзала. Для данного признака не была использована трехмерная система координат в связи с тем, что искривление земной поверхности незначимо, на таких маленьких расстояниях. Добавление фактора в модель улучшило качество предсказания на 2% по сравнению с базовыми признаками.

День недели/время суток

Статистика показывает, что день недели, время суток оказывает влияние на количество заказов (см. Рисунок 2), именно поэтому было интересно выяснить влияют ли эти показатели на количество «сгоревших заказов».

Рисунок 2 – Количество заказов в процентах по часам и дням недели. Источник: Яндекс Исследования, 2013

[2,3] Яндекс Исследования: [Электронный ресурс] URL: https://yandex.ru/company/researches/2012/ya_taxi_map (Дата обращения: 24.07.2019)

На основе базовых данных было сформован один бинарный и три категориальных признака: будни/выходные, время суток (утро, день, вечер, ночь), часы пик (с 8:00 до 10:00 и с 18:00 до 20:00) и сгруппированное время. ROC-AUC, после добавления данных признаков, возрос на 2,5% по сравнению с базовым набором данных.

Класс автомобиля

Распределение заказов по тарифам не является равномерным. Так большинство клиентов пользуется вариантом «эконом». На основе этого была выдвинута гипотеза о том, что класс может влиять на вероятность отмены заказа. Предоставленная выборка содержала информацию о четырех типах тарифов: «эконом», «комфорт», «бизнес» и не указан.

Рисунок 3 - Распределение заказов по тарифам.
Источник: Яндекс Исследования, 2012

Для того чтобы проанализировать выдвинутую мной гипотезу, для каждого из тарифов была создана дамми-переменная, показывающая принадлежность к тому или иному классу. После добавления в модель данных признаков точность возросла на 0,5%.

«Ложные заказы»

На рынке такси существует жесткая конкуренция. По данным Аналитического центра при Правительстве Российской Федерации существует три крупных стейкхолдера: Яндекс.Такси (Uber), Gett и Maxim. Иногда эти компании играют не по правилам и делают массовые «ложные вызовы» конкурентам. При анализе предложенной базы данных выяснилось, что существуют взаимосвязь между количеством заказов на один промежуток времени и на одну геоточку по переменной «driver found». Именно для таких дублирующихся заказов была введена дамми-переменная «fraud». Добавление данного признака в модель повысило качество на выборке более чем на 8%.

Обучение модели происходило с помощью CatBoost[4], программной библиотеки разработанной компанией Яндекс, по сути, представляющей собой модифицированный

[4] Яндекс Технологии: [Электронный ресурс] URL: https://yandex.ru/dev/catboost/ (Дата обращения:24.07.2019)

метод градиентного бустинга. Одно из преимуществ данного метода заключается в том, что он имеет возможность работать не только с числовыми, но и с категориальными признаками (важное преимущество, учитывая специфику нашей модели), что приводит к меньшему искажению работы. Пример сравнения точности CatBoost и логит-регрессии по ROC-AUC представлено на Рисунках 2-3.

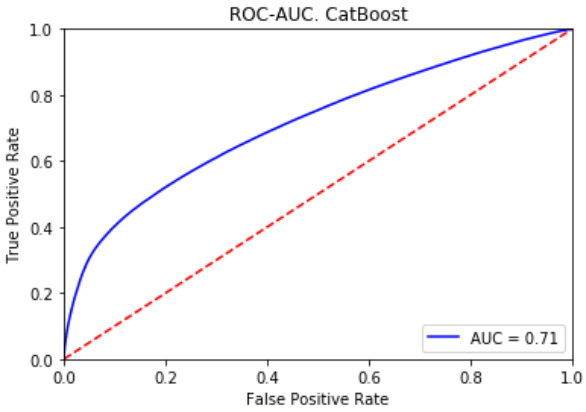

Рисунок 4 – ROC-AUC модели обученный с помощью CatBoost

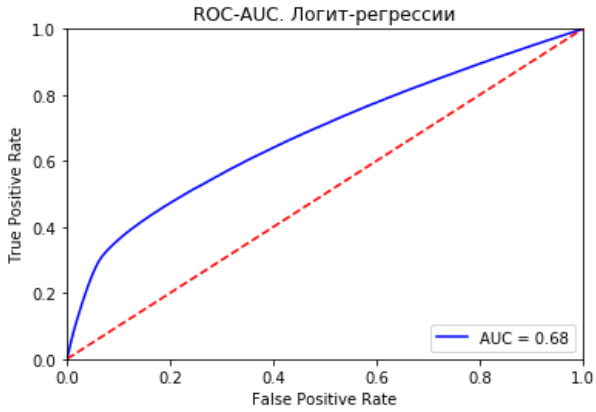

Рисунок 5 – ROC-AUC логит-модели

В рамках данного проекта я сумела подтвердить все предложенные мной гипотезы. Каждый признак увеличивал точность предсказания от 0,5% до 8%. Модель, обученная с помощью CatBoost на обучающей выборки, определяет вероятность отмены заказа с точностью 71%, что в целом больше базового значения на 11%. На тестовой выборке этот показатель несколько ниже и равен 67%. Этот результат позволил мне занять 2 место в Kaggle-чемпионате[5], с отставанием менее чем в 1% от лидера.

[5] Kaggle: [Электронный ресурс] URL: https://www.kaggle.com/c/orders-prediction/leaderboard (Дата обращения:24.07.2019)

Список источников

1. Аналитический центр при Правительстве Российской Федерации: [Электронный ресурс]. «Исследование рынка такси. 2018» URL: http://ac.gov.ru/files/content/15801/issledovanie-taksi-2018-pdf.pdf (Дата обращения:25.07.2019)
2. Яндекс Исследования: [Электронный ресурс]. «Откуда вызывают такси. 250 мест, откуда пользователи Яндекс.Такси чаще всего заказывают машины. По данным Яндекс.Такси, ноябрь-март 2012» URL: https://yandex.ru/company/researches/2012/ya_taxi_map (Дата обращения: 24.07.2019)
3. Яндекс Исследования: [Электронный ресурс]. «Статистика Яндекс.Такси. 2012» URL: https://yandex.ru/company/researches/2012/ya_taxi_pic (Дата обращения: 24.07.2019)
4. Яндекс Исследования: [Электронный ресурс]. «Такси в Москве. Ноябрь 2013». URL: https://yandex.ru/company/researches/2013/ya_moscow_taxi_2013#toc2.2.2 (Дата обращения: 24.07.2019)
5. Яндекс Технологии: [Электронный ресурс]. «CatBoost. Продвинутая библиотека градиентного бустинга на деревьях решений с открытым исходным кодом» URL: https://yandex.ru/dev/catboost/ (Дата обращения: 24.07.2019)
6. Kaggle: [Электронный ресурс] URL: https://www.kaggle.com/c/orders-prediction/leaderboard (Дата обращения: 24.07.2019)