Решение команды RDX

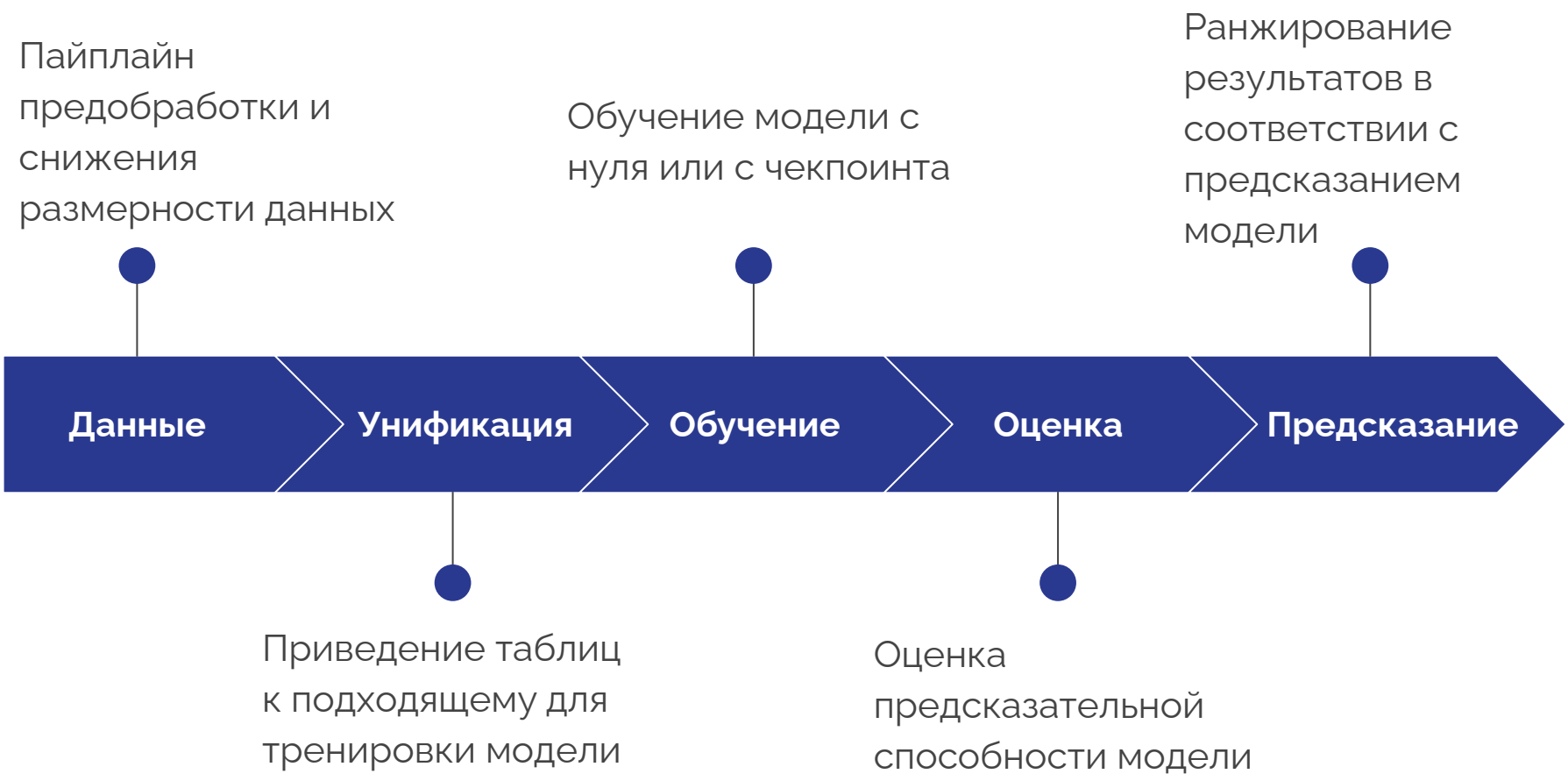
Цифровой прорыв

Победи отток клиентов в сфере быстрого питания

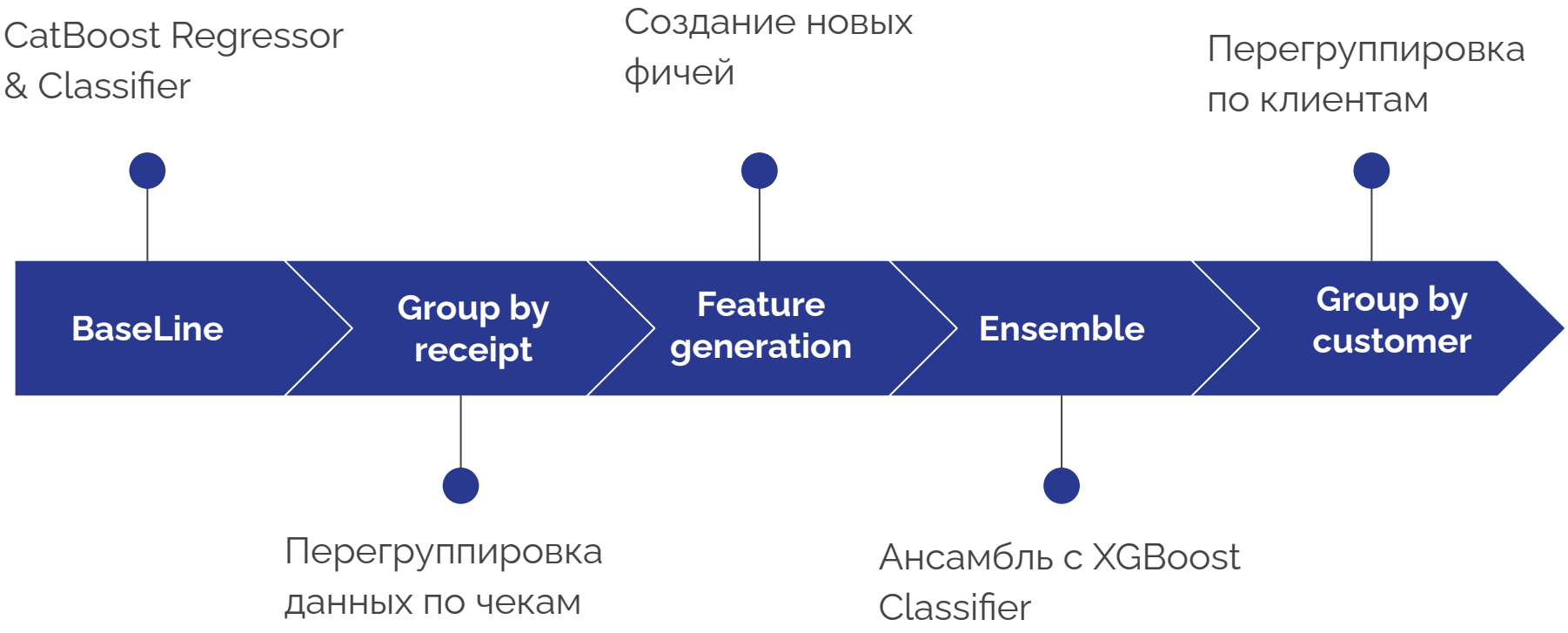
DCS | Digital Consulting Solutions

Суть задачи

Расчет retention для ресторана. Модель будет основана на анализе больших объемов данных и методах машинного обучения, так как уменьшить customer churn является критически важным для успешного развития бизнеса.


Пайплайн предобработки и снижения размерности данных
Обучение модели с нуля или с чекпоинта
Ранжирование результатов в соответствии с предсказанием модели
Данные
Унификация
Обучение
Оценка
Предсказание
Приведение таблиц к подходящему для тренировки модели виду
Оценка предсказательной способности модели

Процесс решения


CatBoost Regressor & Classifier
Создание новых фичей
Перегруппировка по клиентам
BaseLine
Group by receipt
Feature generation
Ensemble
Group by customer
Перегруппировка данных по чекам
Ансамбль с XGBoost Classifier

Работа с данными

- Фильтрация выбросов в переменных
- Заполнение NaN
- Encoding переменных
- Группировка по чекам
- Группировка по клиентам
- Генерация фичей

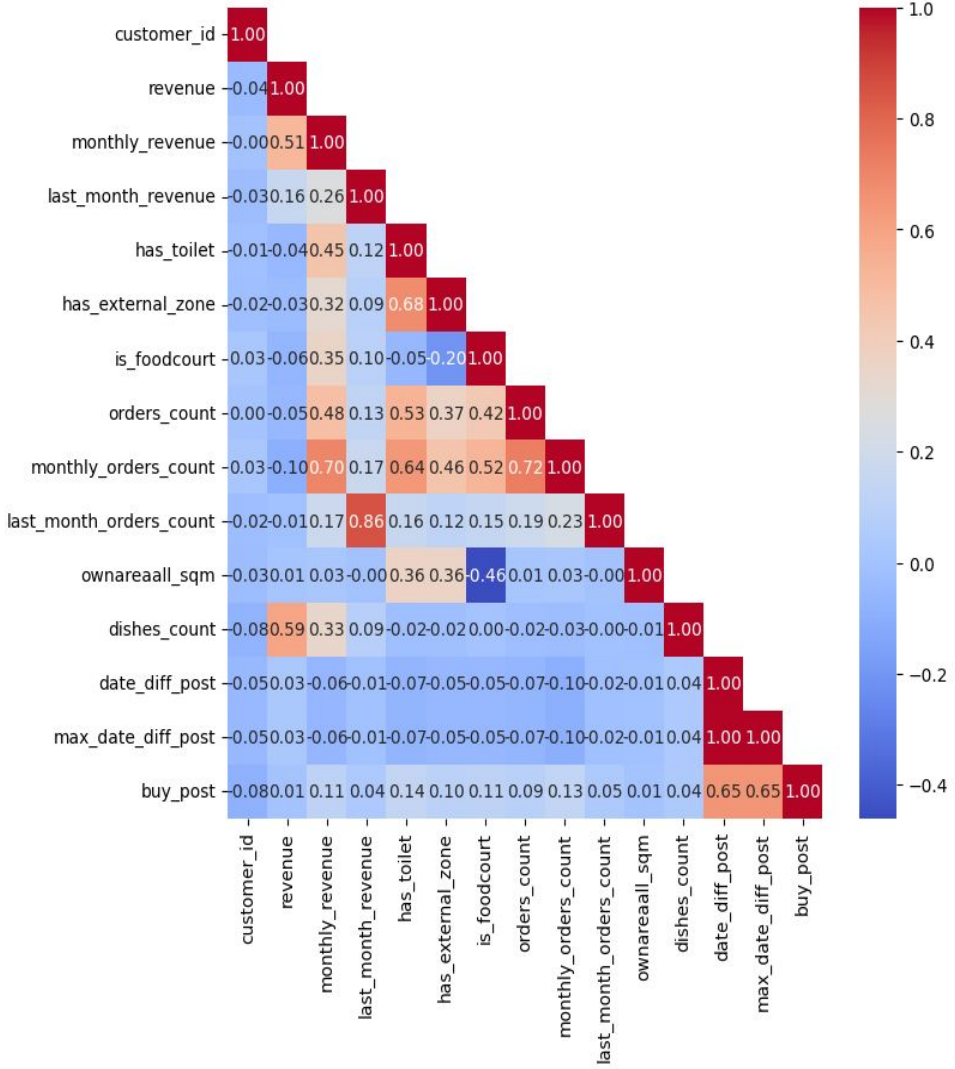
Унификация

Построение прокси фичей

- Количество в заказе
- Средний чек
- Среднее количество заказов
- Среднее количество посещений в месяц

Обработка данных

Матрица корреляций фичей



Решение

CatBoost

XGBoost

Градиентный бустинг деревьев решений с пайплайном предобработки данных, валидацией модели и дообучением с чекпоинта. Легко интерпретируемая модель

Основные метрики

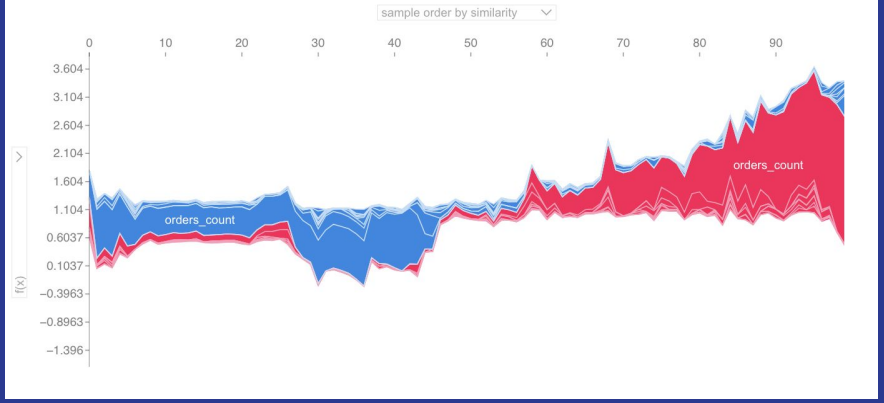
test_size = 0.2

```
print(classification_report(y_test_clf, xgbc.predict(xgbc_test)))
```

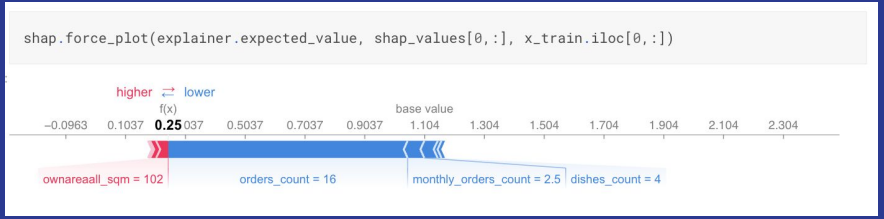
	precision	recall	f1-score	support
0	0.50	0.12	0.19	34719
1	0.74	0.95	0.83	89441
accuracy			0.72	124160
macro avg	0.62	0.54	0.51	124160
weighted avg	0.67	0.72	0.65	124160

Дополнительные метрики

```
shap.force_plot(explainer.expected_value, shap_values[0:100,:], x_train.iloc[0:100,:])
```

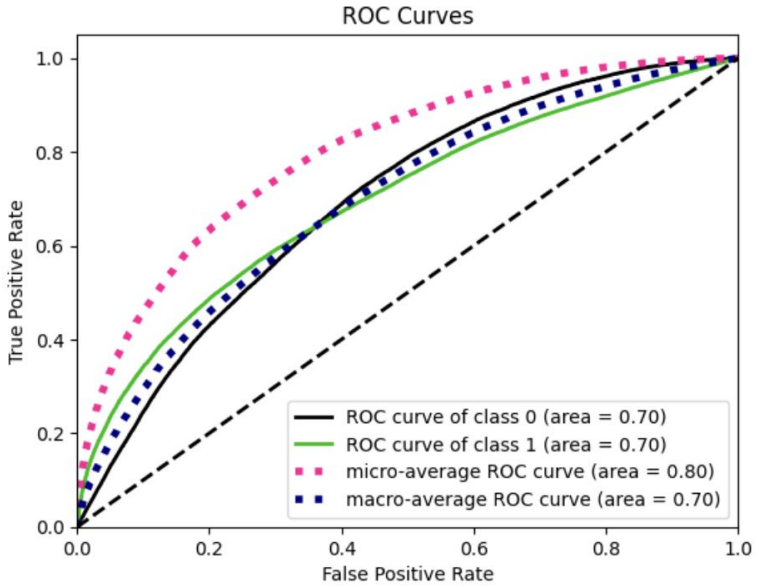



```
shap.force_plot(explainer.expected_value, shap_values[0,:], x_train.iloc[0,:])
```



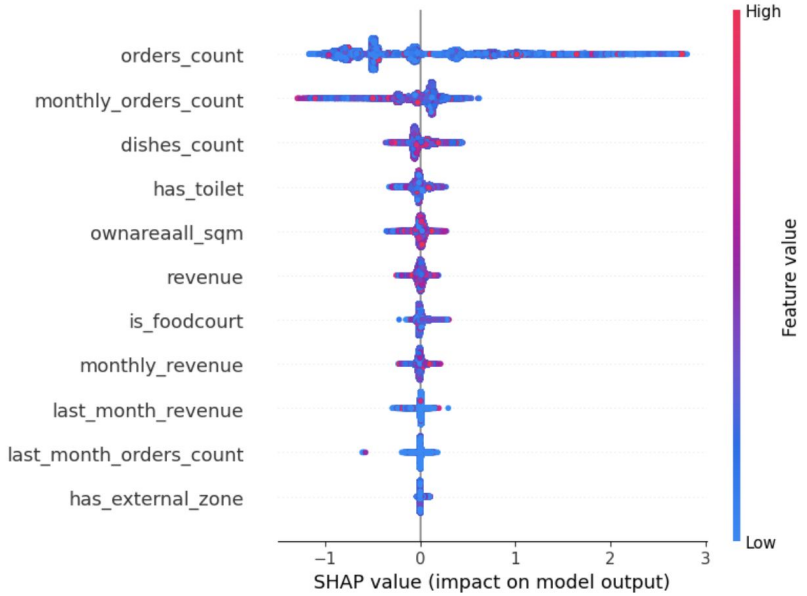
Основные метрики

```
skplt.metrics.plot_roc_curve(y_test_clf, xgbc.predict_proba(xgbc_test))
plt.show()
```



Дополнительные метрики

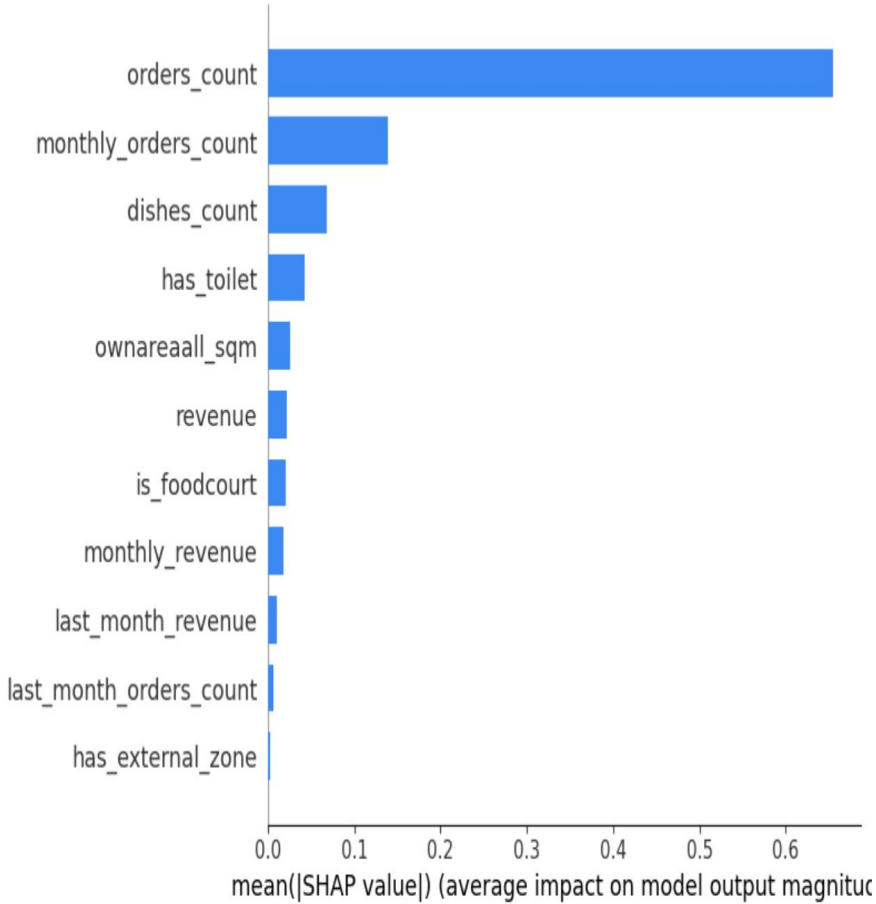
```
shap.summary_plot(shap_values[:100000], x_train[:100000])
```



Результат предсказания

Влияние фичей

```
shap.summary_plot(shap_values, x_train, plot_type="bar")
```



Планируемые доработки

Данные	Модель	Оценка
Перенос обработки данных в *Airflow* и унификация процесса обработки данных, а также дальнейшего сохранения в *feature store*	Вынос работы модели в отдельный сервис, при помощи *MLFlow* и настройка регулярного дообучения и обновления артефактов на основе *S3*	Внедрение карты метрик и динамический сбор пользовательского фидбека с последующей обработкой и передачей в модель

Ссылки на github и скринкаст

Github

ScreenCast

Команда

Дмитрий Жуковский, Data Engineer

Александр Шаталин, Data Scientist

Максим Тер-Мкртчан, ML Engineer