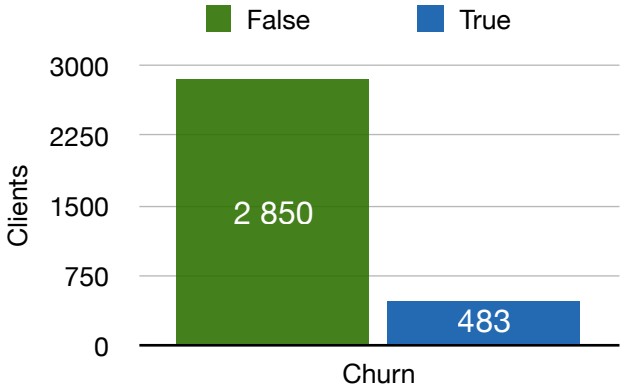
Задача

Необходимо провести анализ данных клиентов телеком-компании, построить модель по оценке вероятности оттока на основе данных и оценить качество модели (классификация на 2 класса).

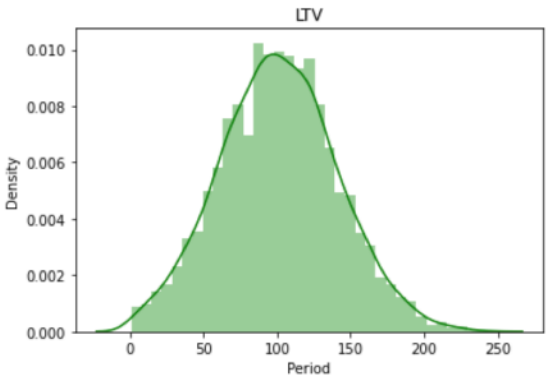
Анализ данных

В данных содержится информация о 3333 клиентах из 51 штата Америки, отток клиентов составляет 17 %.



Согласно данным:

- 50 % являются клиентами 6-11 лет;
- 25 % являются клиентами менее 6 лет;
- 25 % являются клиентами от 11 до 17 лет;

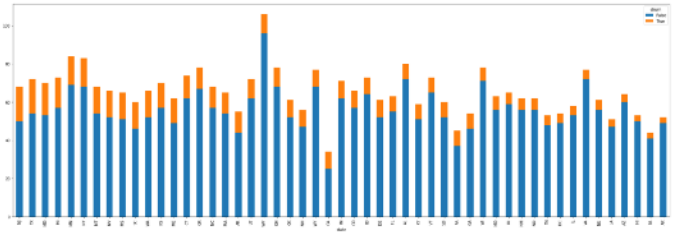


Общая сумма дохода составляет 198 000 $, из них:

- 51 % оплата дневных звонков;
- 28 % оплата вечерних зконков;
- 15 % оплата ночных звонков;
- 5 % оплата международных звонков.

Топ-5 по оттоку клиентов в разрезе штата:

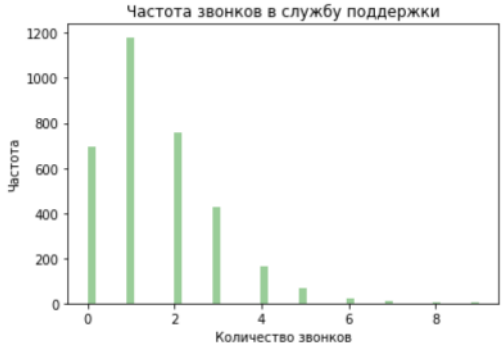
- Нью-Джерси - 26 %;
- Техас - 25 %;
- Мэриленд - 24 %;
- Мичиган - 22 %;
- Миннесота - 18 %.



Тарифная сетка звонков для каждого штата составляет:

Тариф	Ставка, ₡ per minute
Дневной	17
Вечерний	8,5
Ночной	4,5
Международный	27

Согласно данным клиенты редко обращаются в абонентскую службу (в среднем не более 2 звонков за указанный период).



Анализ в разрезе подключенных опций показал:

- 90 % клиентов не подключают пакет международного тарифа;
- 72% клиентов не подключают пакет голосовых сообщений.

Machine learning

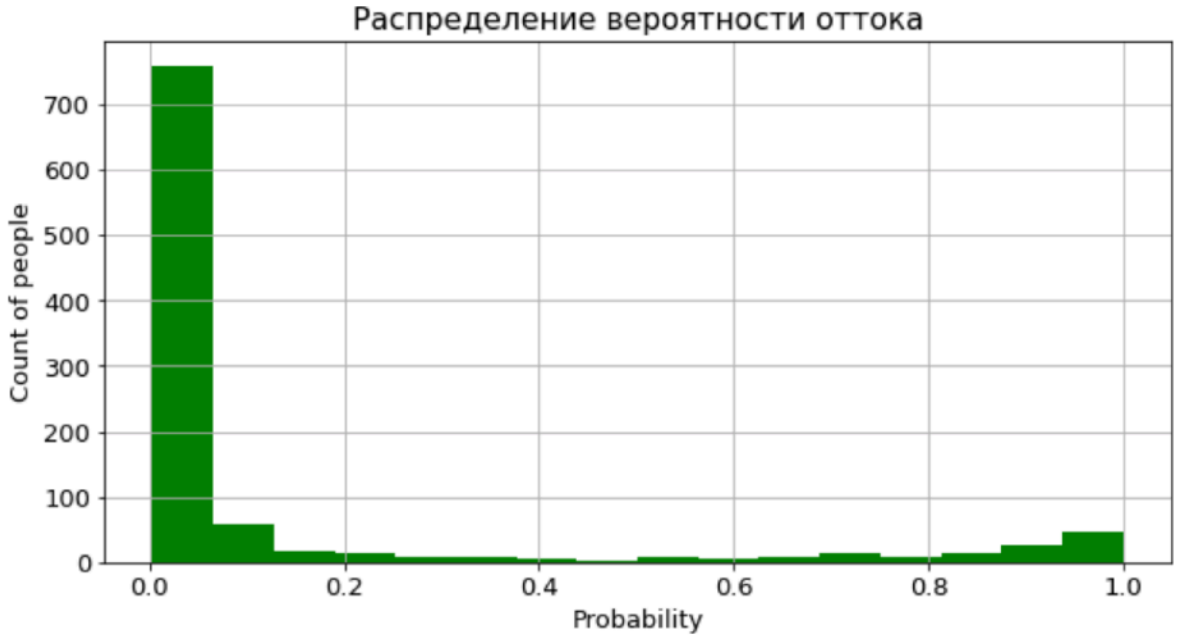
Для построения предиктивной модели данные были разделены на обучающую и валидационную выборки.

Оценка качества проводилась с использованием метрики ROC-AUC score, принимающая значения в диапазоне 0 до 1 (чем выше, тем лучше), т.к. она более устойчива к несбалансированному набору целевой переменной (‘churn’).

Были получены следующие показатели:

Модель	ROC-AUC score
CatBoost	0,95
Random forest	0,93
Decision tree	0,91
Логистическая регрессия (признаки с p-value < 0.05)	0,83
Логистическая регрессия (все признаки)	0,80
K Nearest Neighbor (kNN)	0,64

Для тестовой выборки (30% данных из датасета) была построена гистограмма распределения вероятности оттока клиентов:



Согласно гистограмме распределения можно сделать вывод, что большая часть клиентов имеет низкую вероятность оттока, и 50 человек из рассматриваемой выборки покинут компанию с вероятностью 95%.