В рамках данного проекта была поставлена задача построить модель с предельно большим значением f1-меры для прогноза оттока клиентов из банка на основе исторических данных о поведении клиентов и расторжении договоров с банком.

Для построения требуемой модели предварительно была проведена подготовка данных.

Далее был рассмотрен баланс классов целевого признака. Было выявлено, что положительный класс составляет лишь 20% от выборки, то есть наблюдается сильный дисбаланс классов.

В исследовательских целях были обучены модели на несбалансированной выборке и рассчитаны их метрики. В качестве алгоритмов обучения были использованы Decision Tree,Random Forest и Logistic Regression. Поскольку обучение моделей осуществлялось на несбалансированной выборке, для повышения качества моделей для каждого алгоритма был проведен подбор гиперпараметров, и была определена модель с наиболее высоким показателем f1-меры.

На следующем этапе исследования рассматривался вопрос дисбаланса классов с целью повышения метрик качества моделей. Для борьбы с дисбалансом рассматривались несколько подходов:

* Взвешивание классов;
* Увеличение выборки (Upsampling);
* Уменьшение выборки (Downsampling).

По каждой получившейся сбалансированной по данным подходам обучающей выборке были рассчитаны метрики качества f1-мера и AUC-ROC, построен график ROC-кривая, при условии, что f1-мера >= 0.59.

Модель с наилучшим показателем качества по метрике f1-мера была использована для тестирования.

На заключительном этапе было проведено обучение модели, показавшей наилучший результат, на "обучающей + валидационной" выборке; и далее была протестирована модель на тестовой выборке. По результатам работы было достигнуто требуемое качество, и, следовательно, модель может использоваться для прогноза оттока клиентов из банка. ​