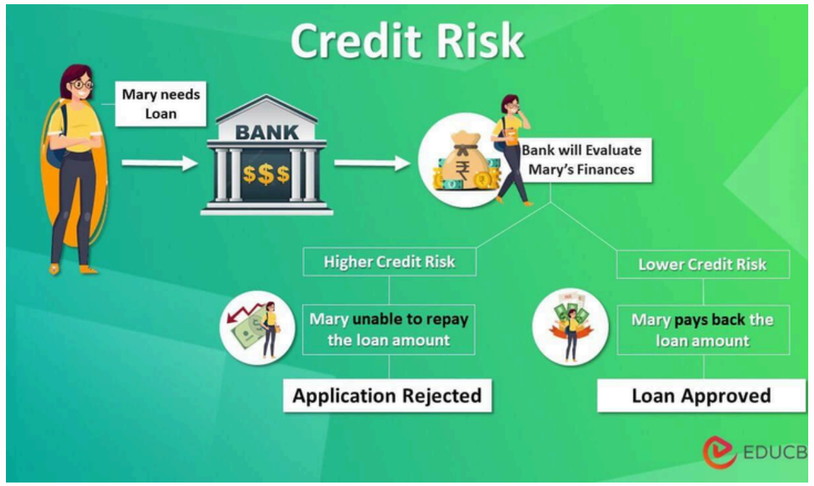
УПРАВЛЕНИЕ КРЕДИТНЫМ РИСКОМ

Создатель: Умиджон Сатторов,
студент платформы Skillbox

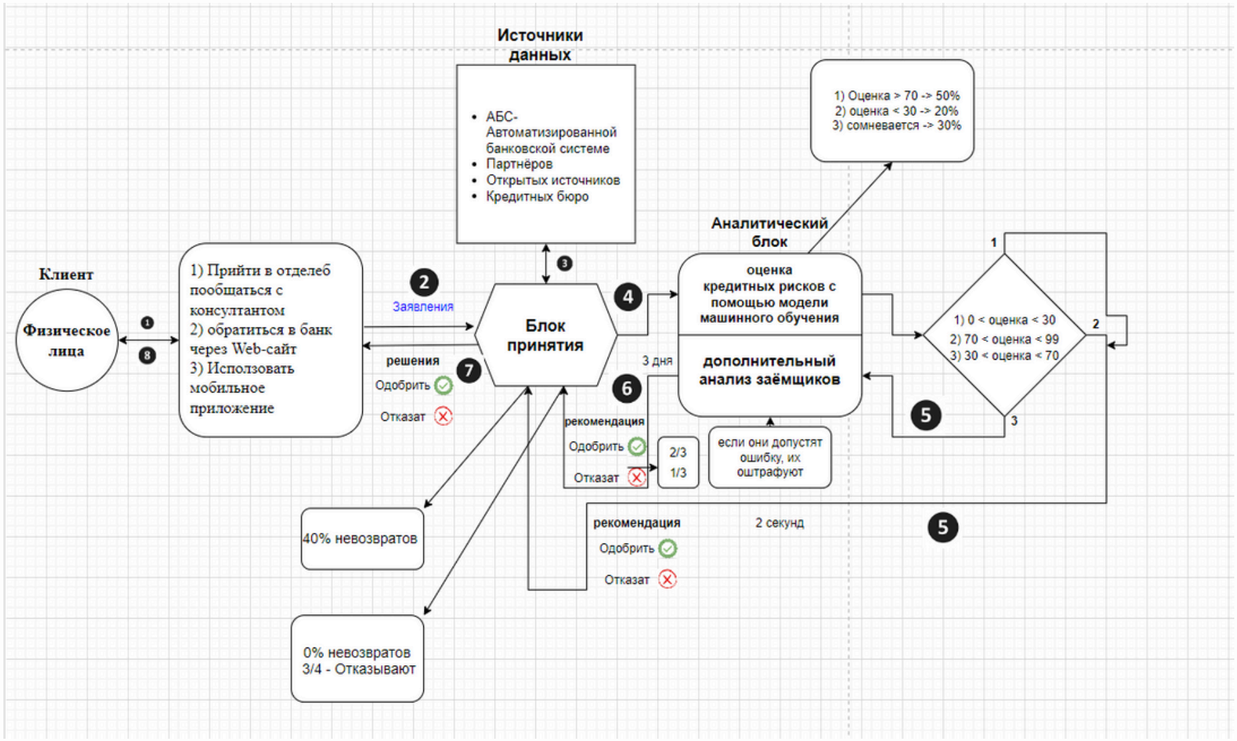
ОГЛАВЛЕНИЕ

BUSINESS ANALYSIS



01 Пользователю банка может потребоваться кредит в банке, но банку необходимо определить, является ли клиент мошенником или нет, чтобы не потерять свои финансовые ресурсы.

02 Практически каждый банк имеет собственную систему управления кредитными рисками для выявления таких рисков. Как бизнес-аналитик, мы должны изучить существующую систему банка, прежде чем углубляться в решение проблемы.



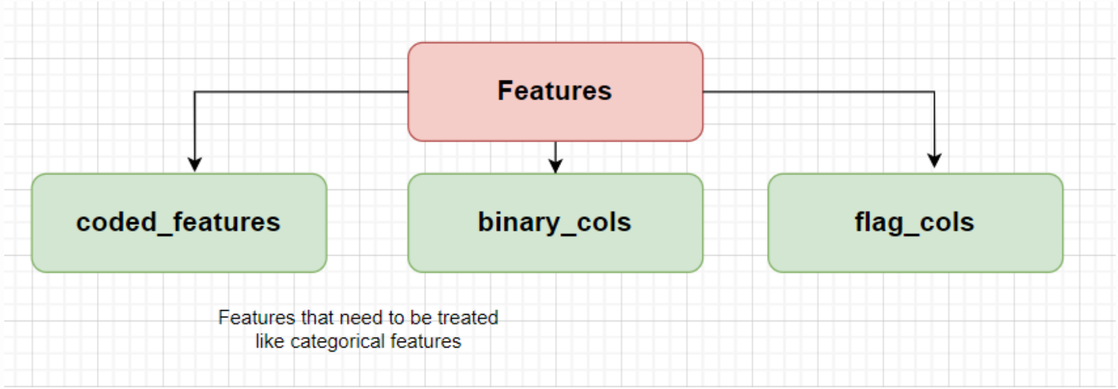
DATA ANALYSIS AND PREPARATION

Data sources and features

Данные содержали 61 различную функцию, все в числовой форме. Но поскольку некоторые столбцы данных были бинаризированы и закодированы, мне пришлось рассматривать их как категориальные столбцы, а не числовые.

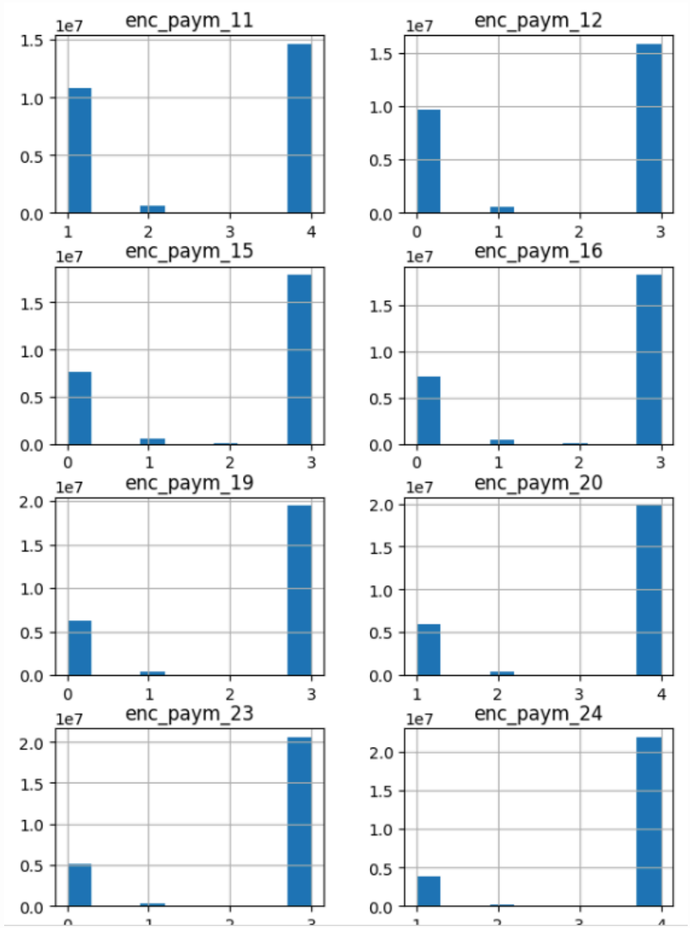
```
1 df.info()
```

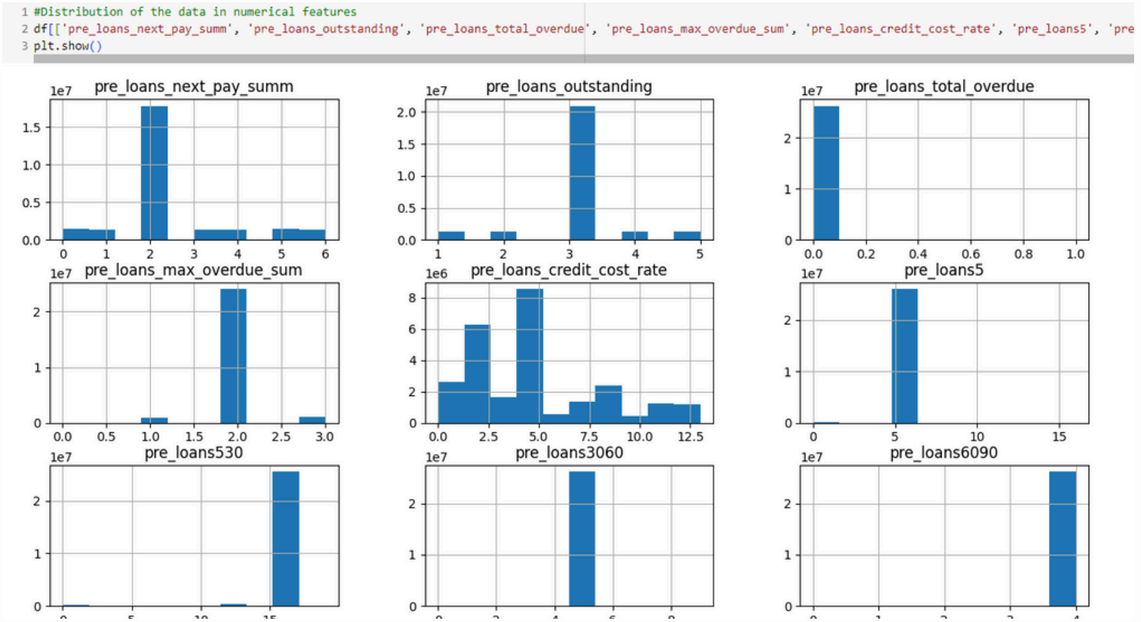
```
<class 'pandas.core.frame.DataFrame'>
Index: 26162717 entries, 0 to 2450629
Data columns (total 61 columns):
 #   Column                          Dtype
---  ------                          -----
 0   id                              int64
 1   rn                              int64
 2   pre_since_opened                int64
 3   pre_since_confirmed             int64
 4   pre_pterm                       int64
 5   pre_fterm                       int64
 6   pre_till_pclose                 int64
 7   pre_till_fclose                 int64
 8   pre_loans_credit_limit          int64
 9   pre_loans_next_pay_summ         int64
 10  pre_loans_outstanding           int64
```



Набор данных содержал много информации о сеансе для одного и того же идентификатора. Поэтому от меня требовалось агрегировать набор данных и хранить одну строку данных для одного идентификатора. Именно поэтому я использовал OneHotEncoder, чтобы не потерять свои данные из-за агрегации.

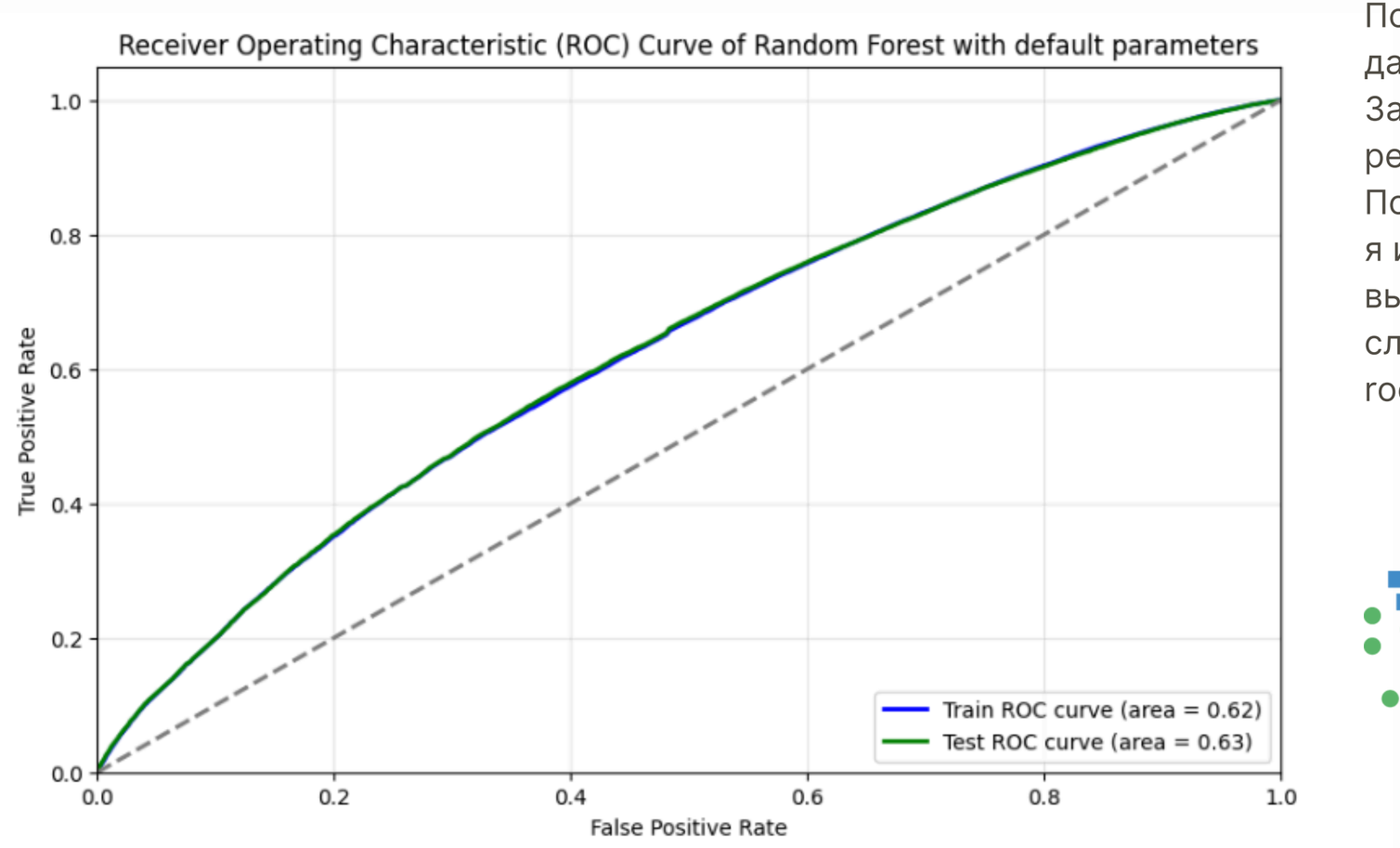
Для предварительной обработки данных я разделил свой набор данных на три отдельные группы, такие как code_features и двоичные столбцы, которые мне пришлось рассматривать как категориальные признаки (а не числовые). После предварительной обработки данных в соответствии с типом объекта.





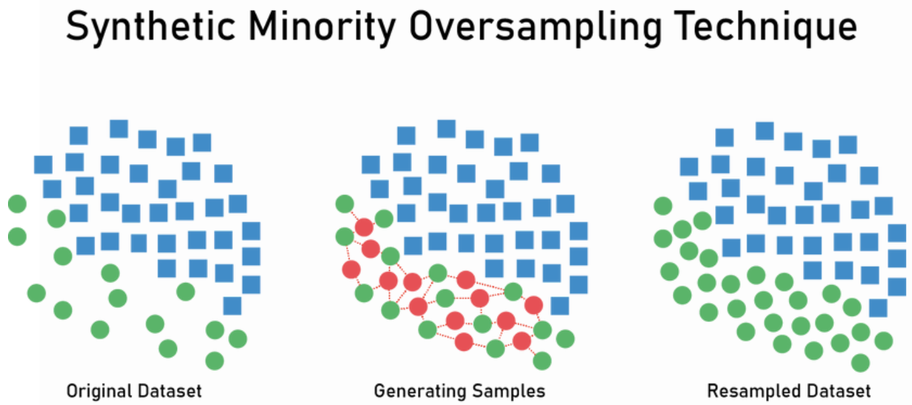
В целях анализа я наблюдал различные особенности данных, такие как размер, распределение набора данных, корреляцию функций данных друг с другом, и проверял, является ли набор данных несбалансированным или нет (наиболее распространенная проблема машинного обучения).

MODEL SELECTION

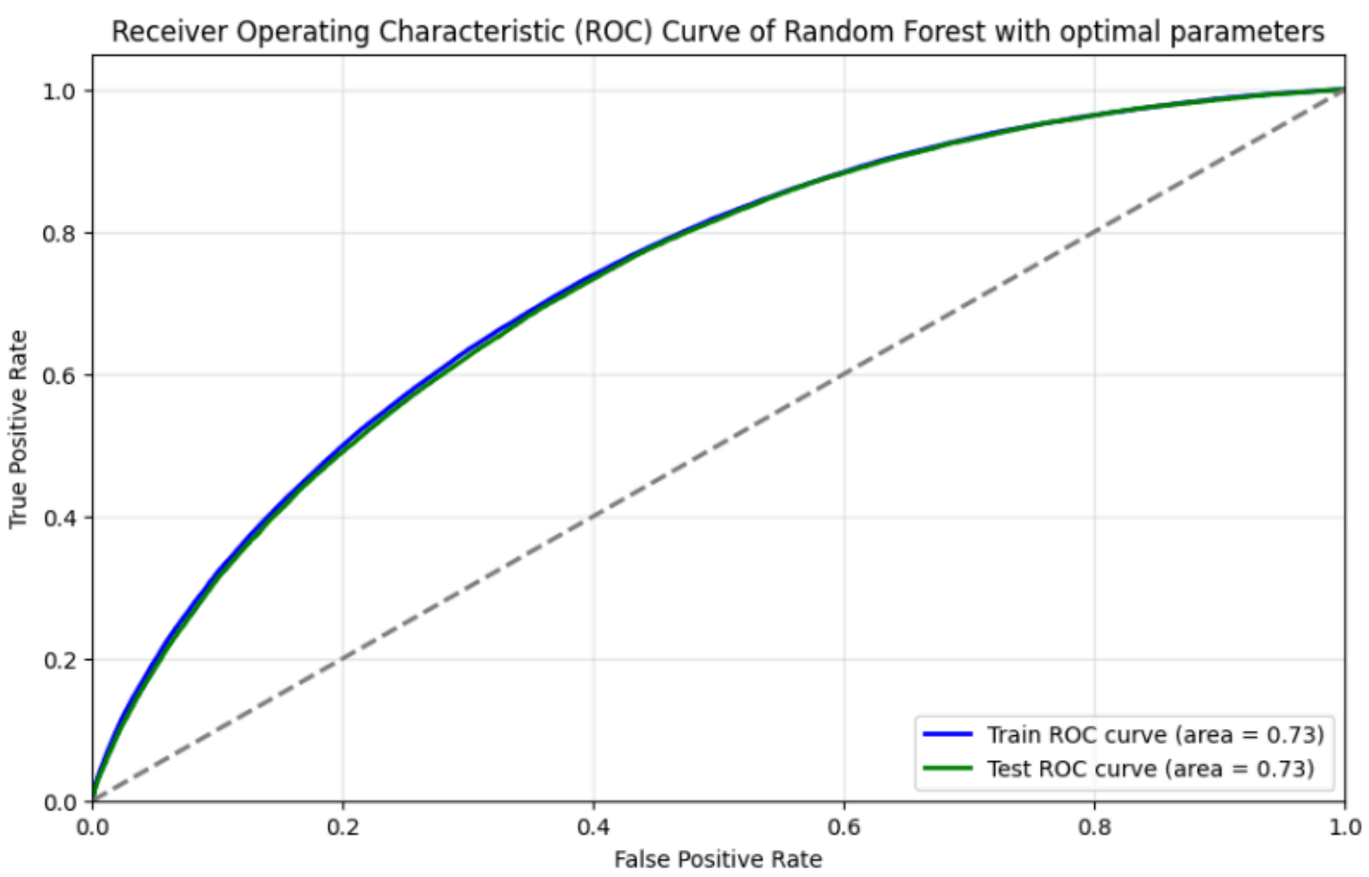


Random Forest classifier

После применения агрегации к моему набору данных у меня осталось 3 миллиона строк данных. Затем я использовал технику SMOTE, чтобы решить проблему несбалансированных данных. После применения SMOTE к моему набору данных я использовал полученные результаты повторной выборки для обучения алгоритму классификатора случайного леса. Он вернул максимум 0,63 roc_score.

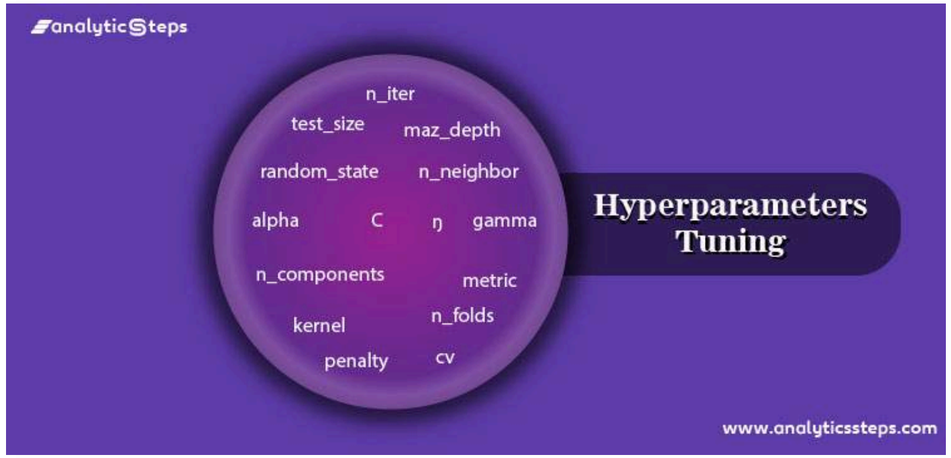


MODEL SELECTION



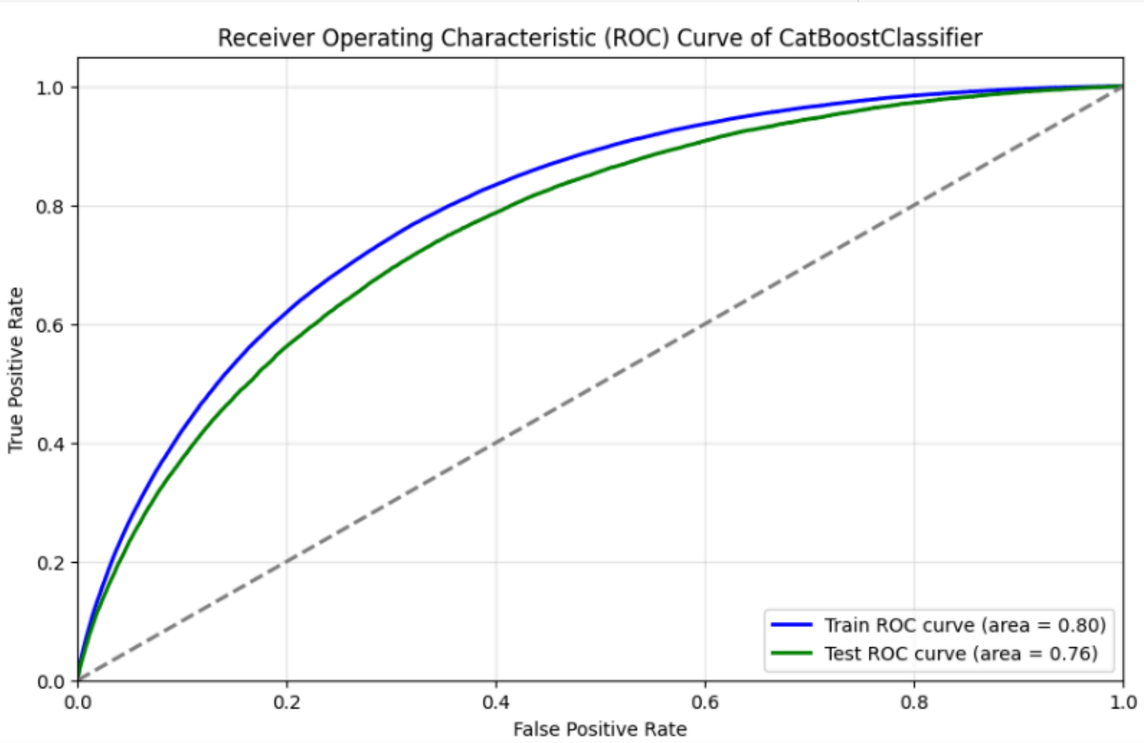
Random Forest classifier

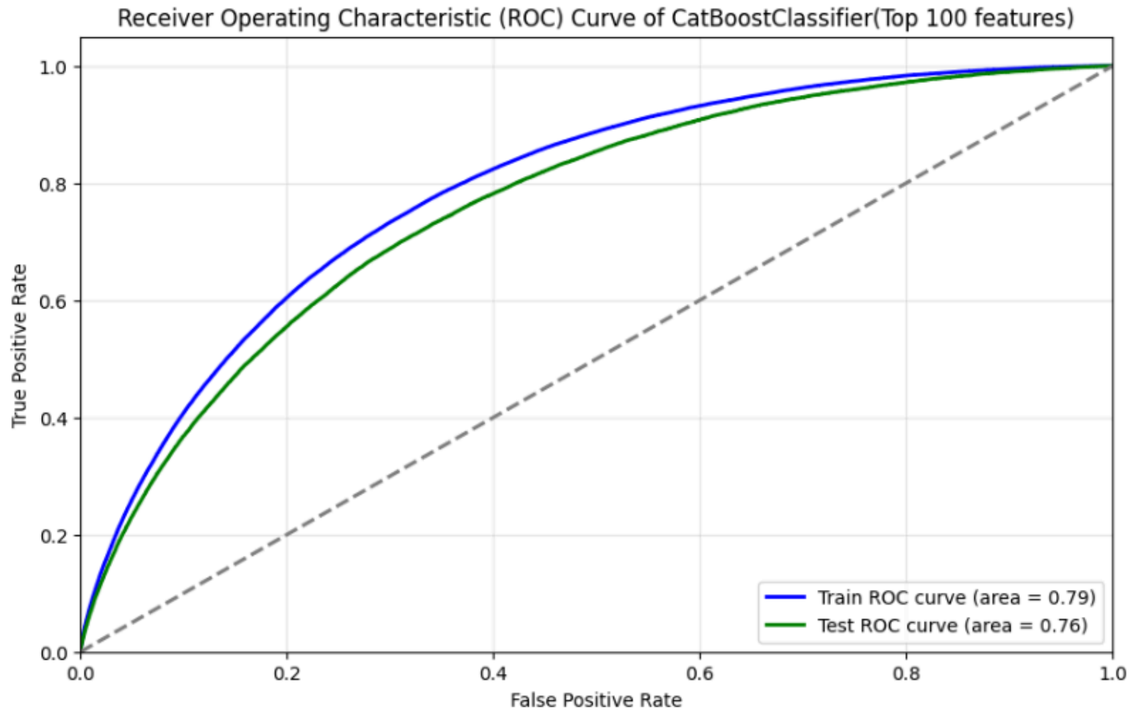
Поскольку моя первоначальная модель не дала наилучшего результата, я решил уточнить ее гиперпараметры. Я думал, что таким образом смогу повысить качество модели случайного леса.



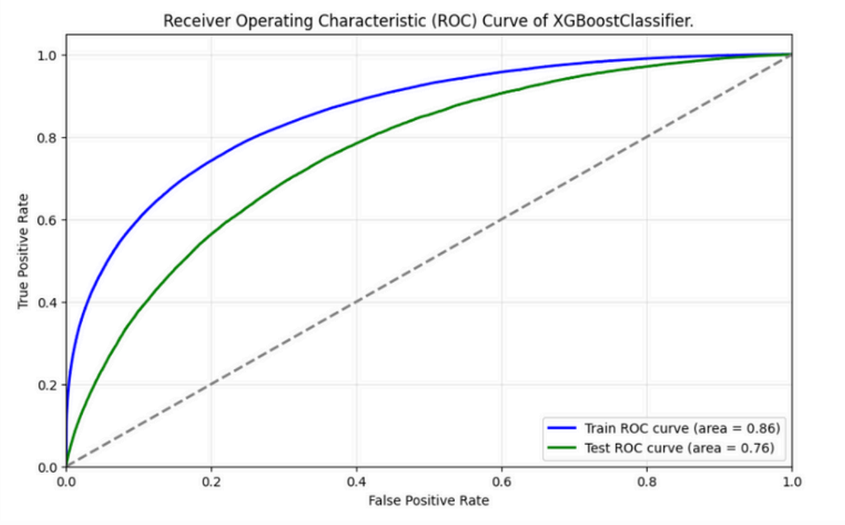
MODEL SELECTION

Поскольку такие модели, как Random Forest Classifier, продемонстрировали выдающиеся результаты при определении того, является ли клиент банка мошенником, я решил создать модель классификации с использованием алгоритмов Gradient Boosting. Среди них алгоритмы CatBoost Classifier показали хорошие результаты и дали наивысший балл roc_auc. Чтобы снизить стоимость вычислений и время, затрачиваемое на прогнозирование, я сохранил только 100 важных функций, позволяющих предсказать, является ли пользователь банка мошенником или нет.

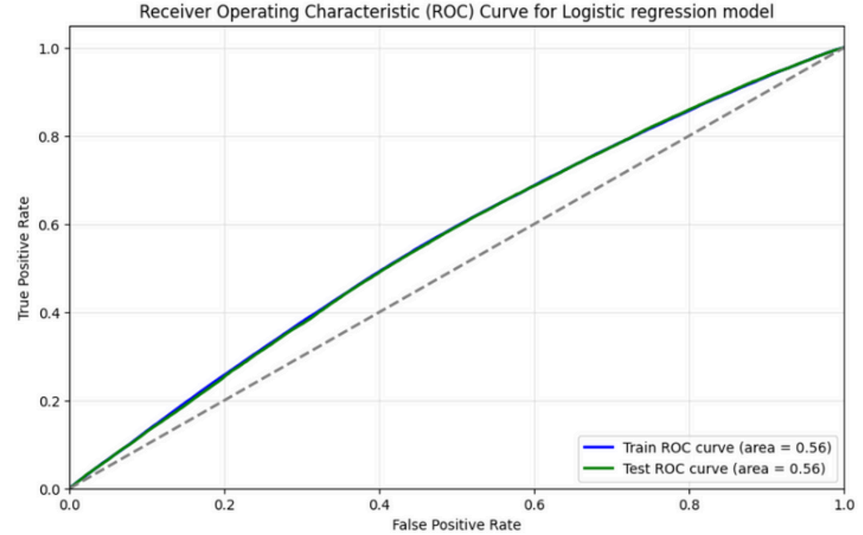




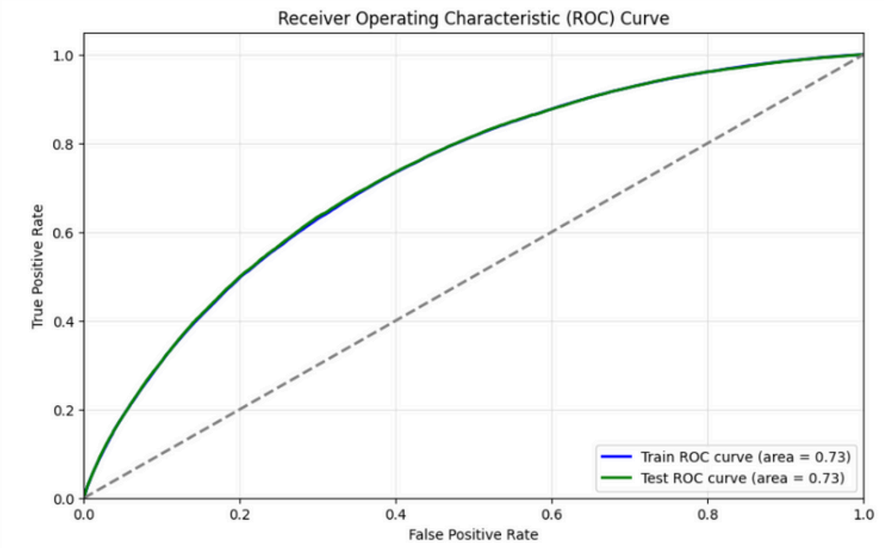
OTHER MODELS RESULTS



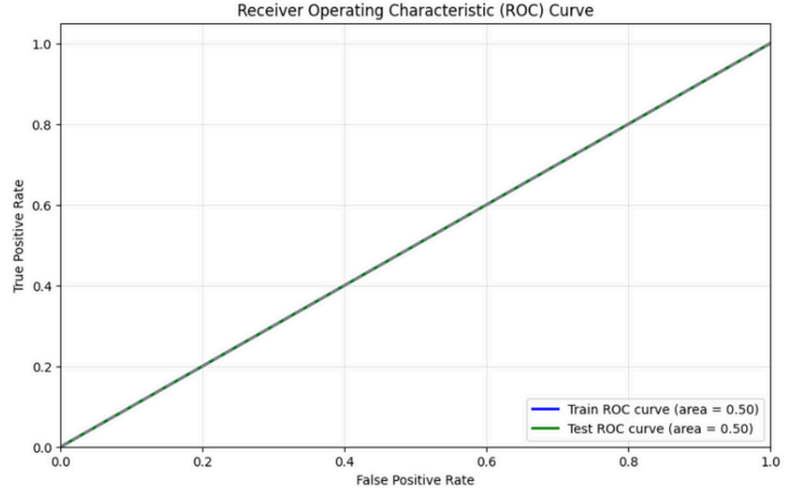
XGBoostClassifier



Logistic regression

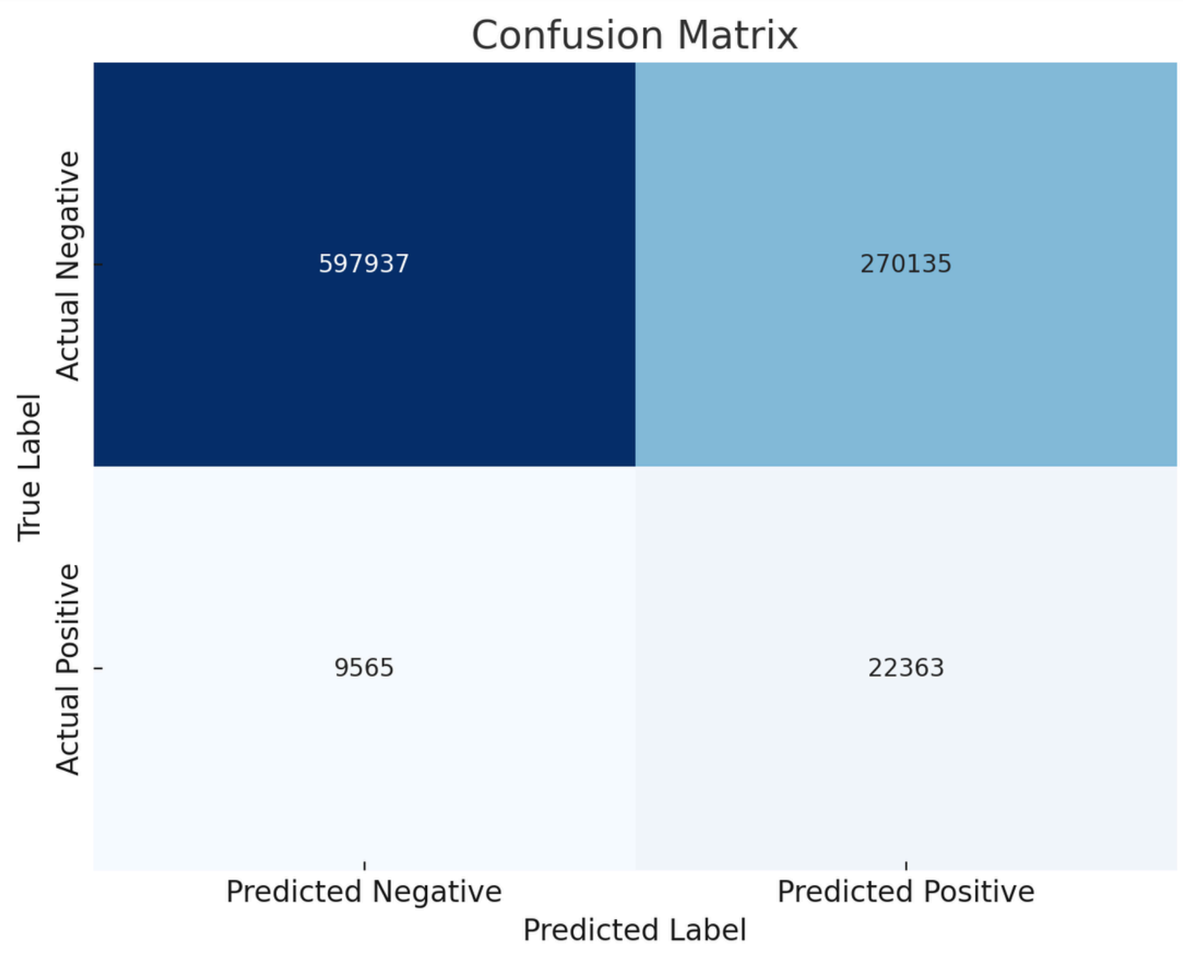


AdaBoostClassifier



MLPClassifier

CONCLUSION



В этом проекте по управлению кредитным риском я работал с различными моделями классификации и разными методами, чтобы преодолеть проблему классового дисбаланса. Проблема была сложной из-за размера набора данных (больших данных) и требуемой очень большой вычислительной мощности. Даже я потратил некоторую сумму денег на покупку дополнительной оперативной памяти, чтобы решить проблему нехватки памяти. Однако мне удалось создать модель машинного обучения с показателем roc_auc более 0,76, чтобы решить проблему управления кредитным риском для банков. Эта модель машинного обучения работает с более высокой точностью и достойно прогнозирует оба класса пользователей банка. Он не перетренирован и работает на очень высокой скорости. Это может легко решить проблему управления кредитным риском для банков.

CONTACT ME

E-mail sattorov7474@gmail.com

Phone +998909250890

Linkedin Umidjon Sattorov