

COMPETITION REPORT

1. Анализ данных

1.1 Описание набора данных

Набор данных состоит из следующих характеристик:

- **Размер:** 58 645 строк (50295 отрицательных классов, 8350 положительных классов)
- **Признаки:** 11 числовых и 4 категориальных
- **Целевая переменная:** loan_status (0 - не одобрен, 1 - одобрен)

1.2 Визуализация распределений

- Построены диаграммы распределения категориальных признаков.

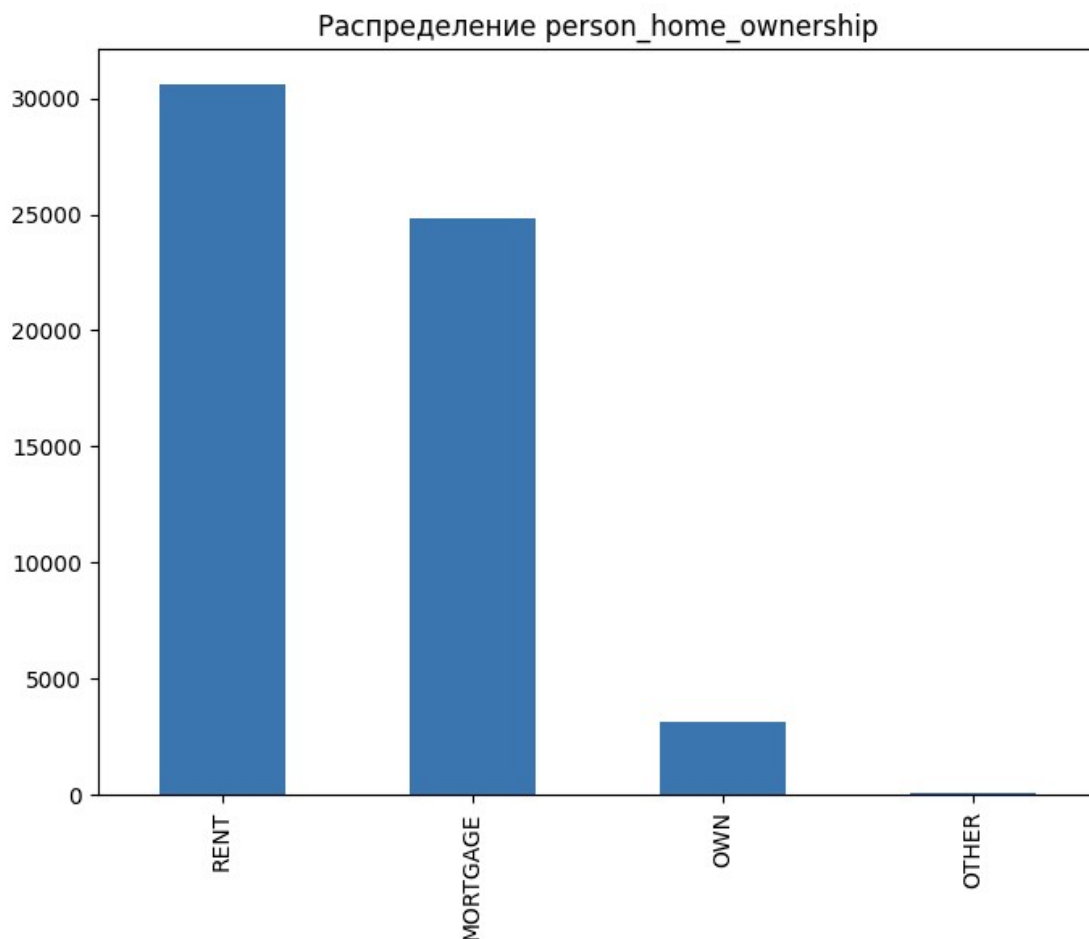


Рисунок 1

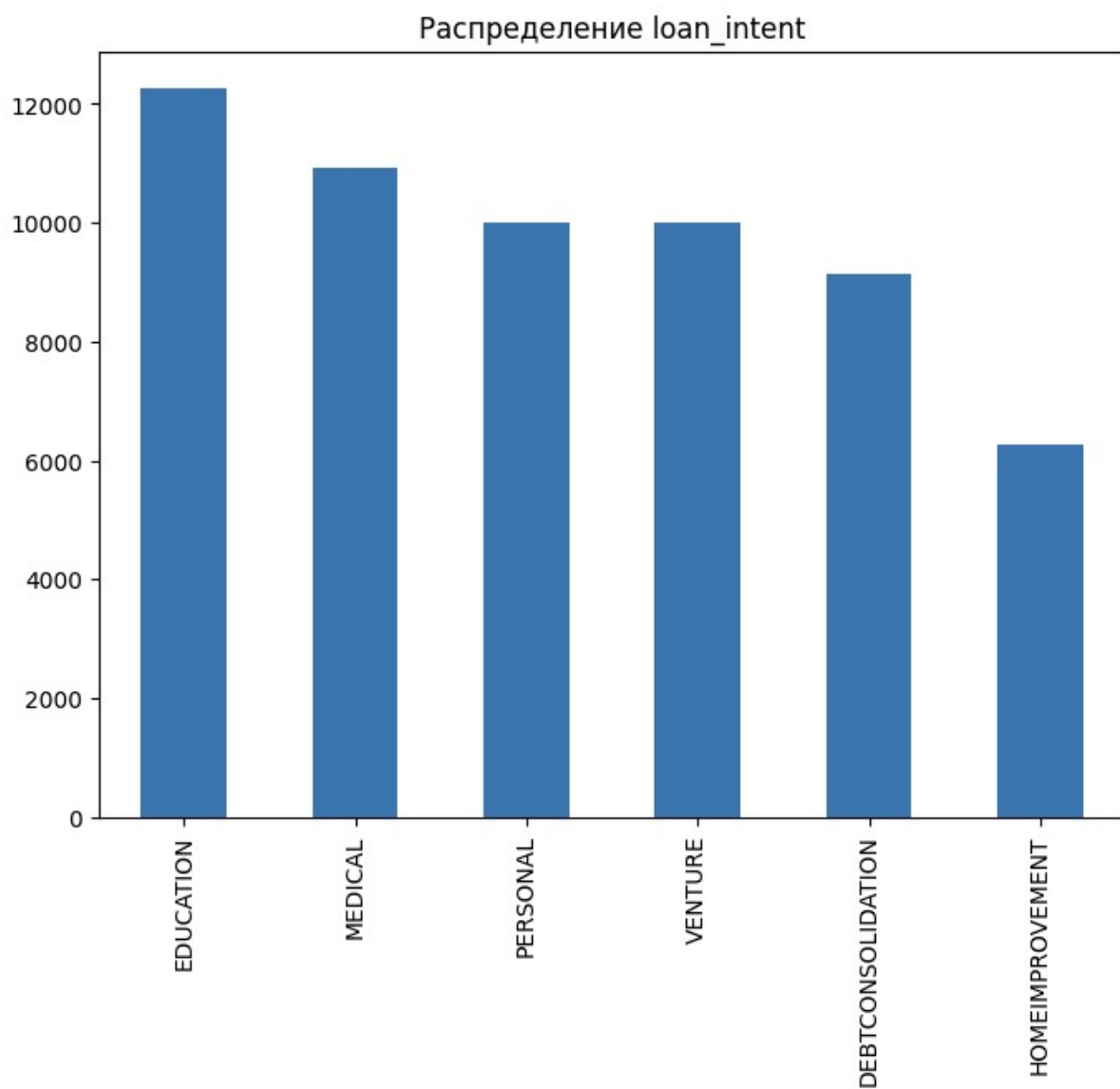


Рисунок 2

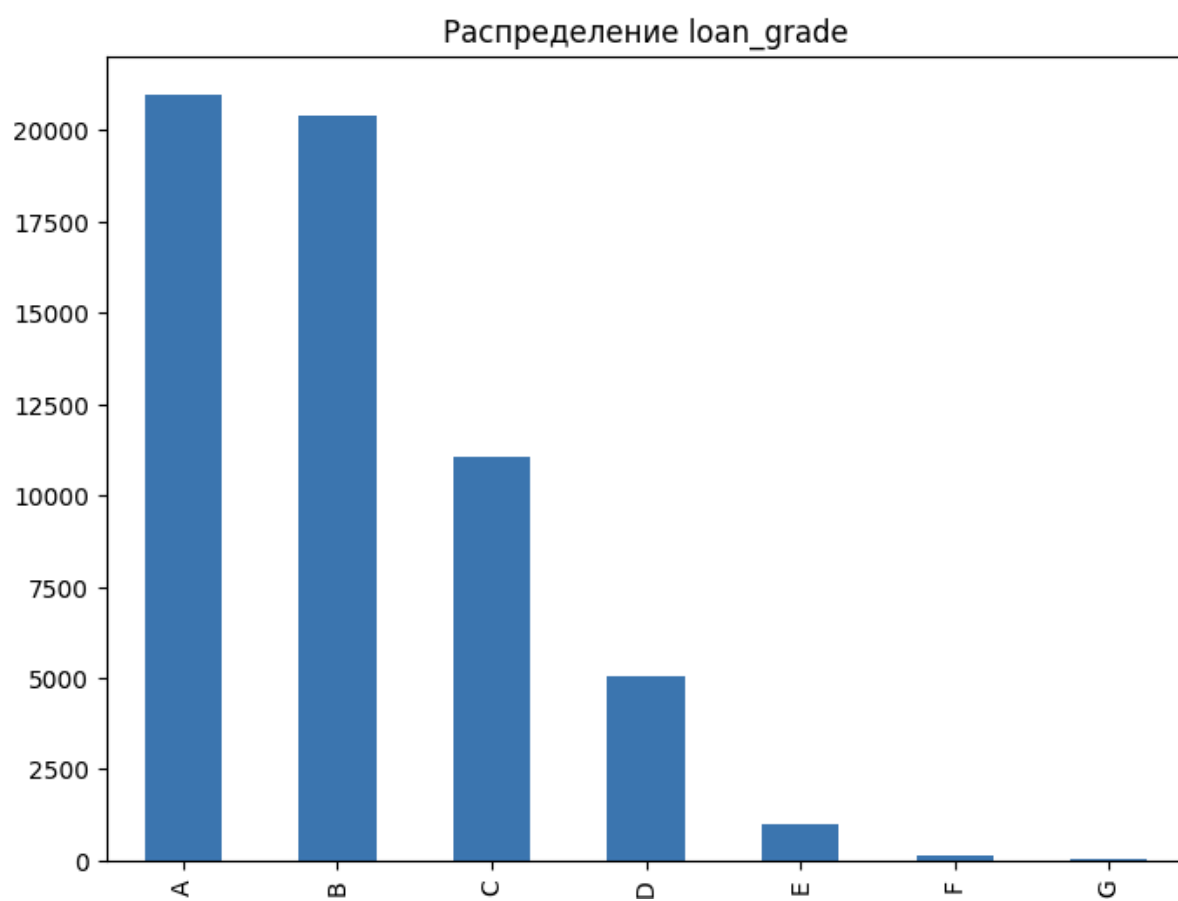


Рисунок 3

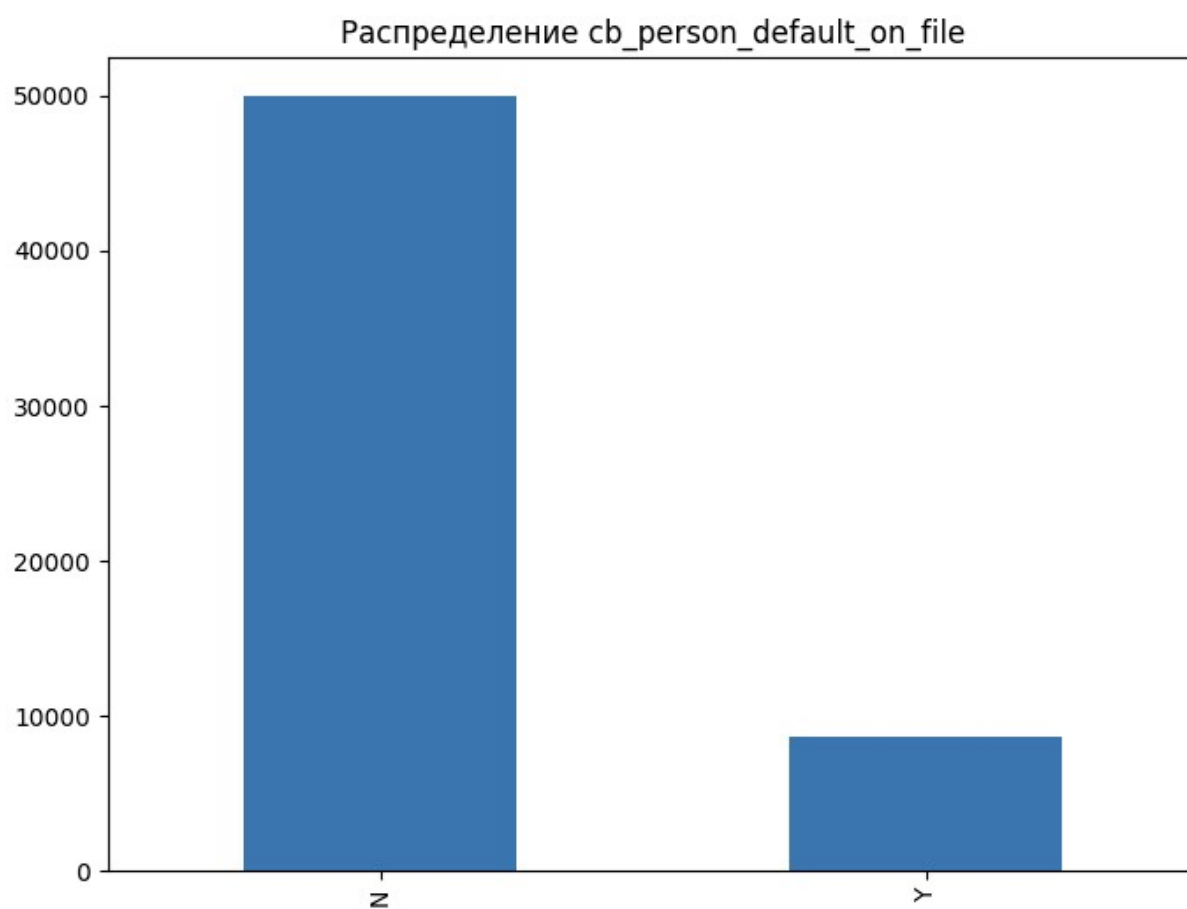


Рисунок 4

- Корреляционная матрица

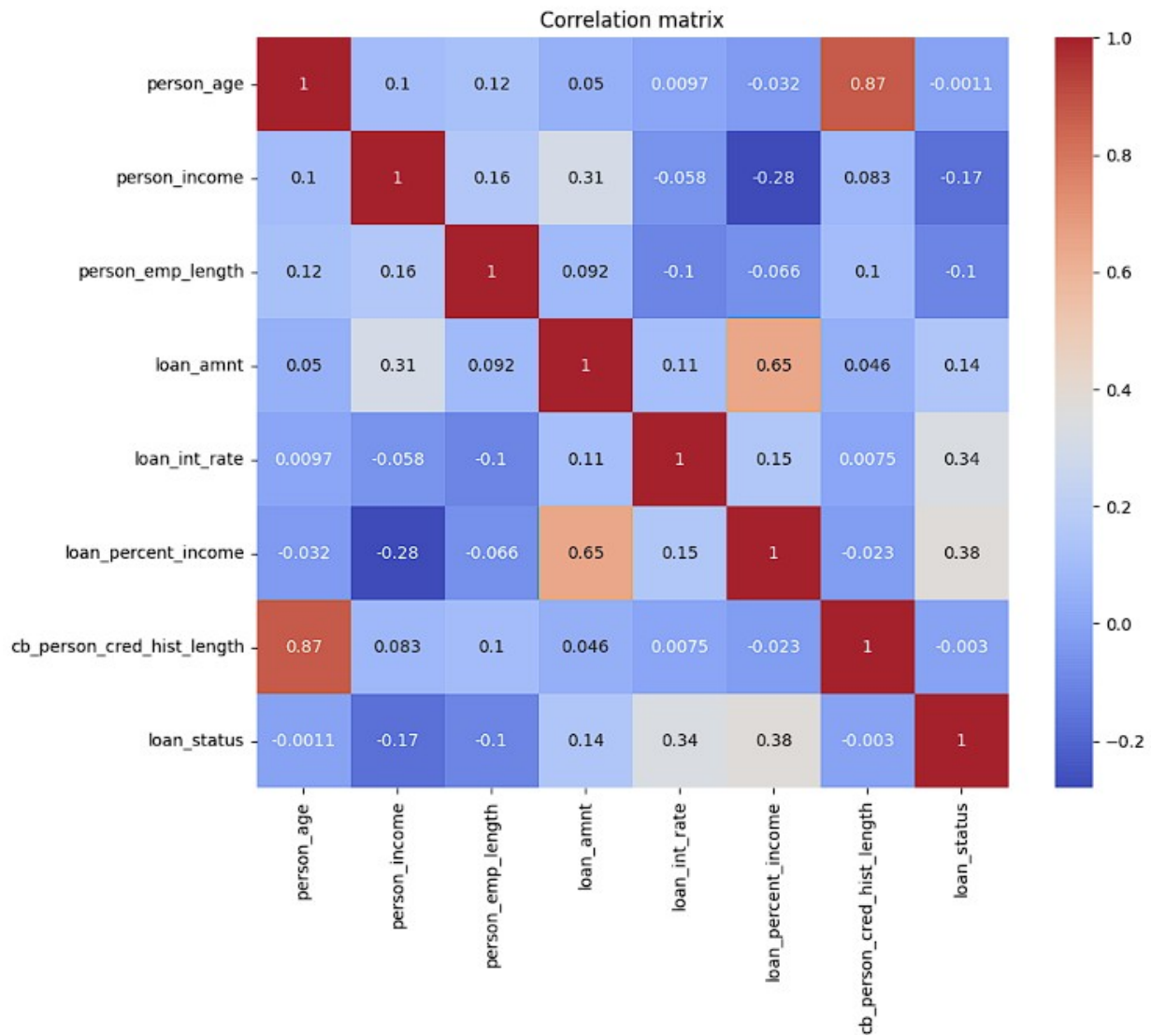


Рисунок 5

1.3 Проверка пропусков и выбросов

- Пропуски в данных отсутствуют (из за искусственной природы)

```
In [4]: data.isnull().sum()

Out[4]:
person_age                0
person_income              0
person_home_ownership      0
person_emp_length          0
loan_intent                0
loan_grade                 0
loan_amnt                  0
loan_int_rate              0
loan_percent_income        0
cb_person_default_on_file  0
cb_person_cred_hist_length 0
loan_status                0
dtype: int64
```

Рисунок 6

- Признак `cb_person_cred_hist_length` удалён из-за низкой корреляции.
- Категориальные признаки закодированы с помощью `OneHotEncoding`.
- Числовые признаки стандартизированы `StandartScaler`.

2. Выбор модели

2.1 Архитектура нейронной сети

Используется полносвязная нейронная сеть (MLP) со следующими характеристиками:

- Входной слой: 21 нейрон (количество признаков после кодирования)
- Два скрытых слоя:
 - 64 нейрона (BatchNorm, Dropout 0.2, ReLU)
 - 32 нейрона (BatchNorm, Dropout 0.2, ReLU)
- Выходной слой: 1 нейрон (без активации, используется `BCEWithLogitsLoss`)

2.2 Выбор функций и гиперпараметров

- **Функция активации:** ReLU. Выбрана из-за её способности решать проблему исчезающего градиента, а также простоты вычислений и эффективности при работе с глубокими нейронными сетями.
- **Функция потерь:** BCEWithLogitsLoss. Подходит для бинарной классификации, так как объединяет в себе сигмоидную активацию и бинарную кросс-энтропию, обеспечивая стабильное обучение.
- **Оптимизатор:** AdamW ($lr=0.0005$, $weight_decay=1e-5$). Выбран благодаря его способности адаптивно изменять шаг градиента и встроенному L2-регуляризатору, который помогает избежать переобучения.
- **Число эпох:** 20
- **Размер батча:** 128

3. Результаты

3.1 Оценка качества модели

- **F1-score:** 0.7744
- **Accuracy:** 0.9397
- **Precision:** 0.8290
- **Recall:** 0.7265
- **ROC-AUC:** 0.9307

3.2 График обучения

- Построен график функции потерь (loss) по эпохам. Наблюдается хорошая сходимость модели.

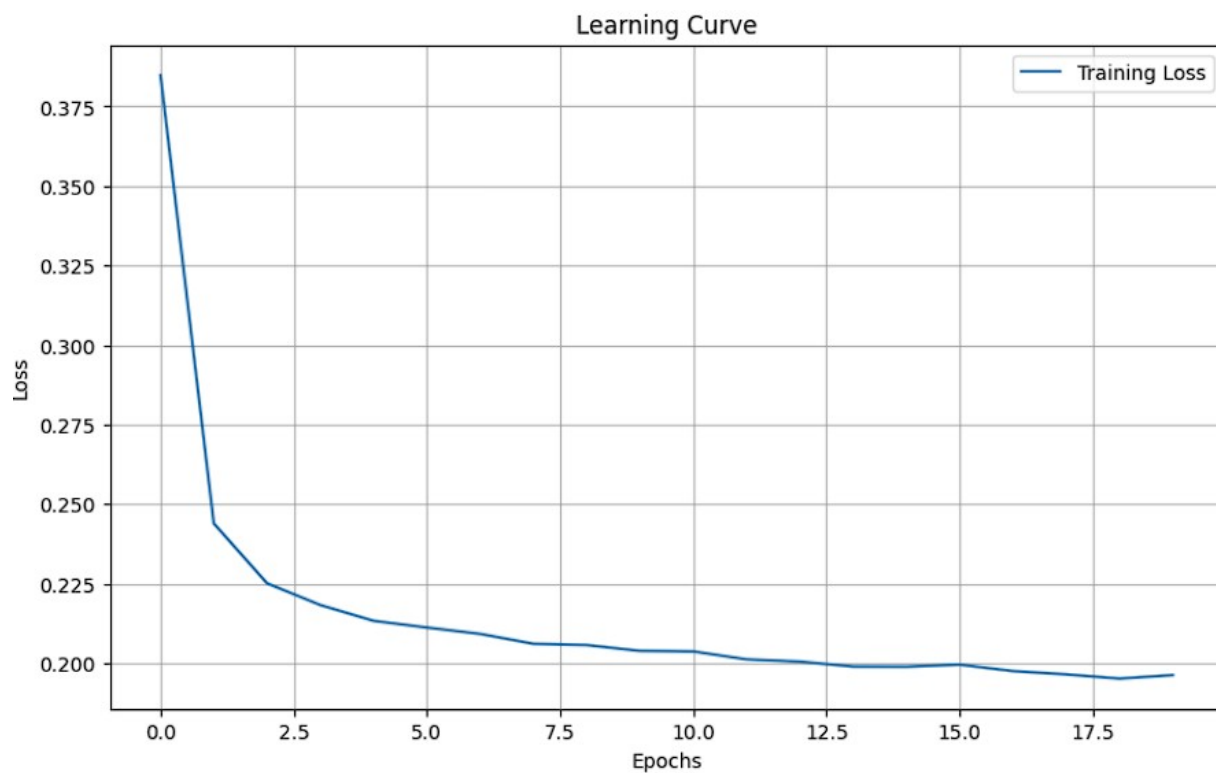


Рисунок 7

- Построен график ROC

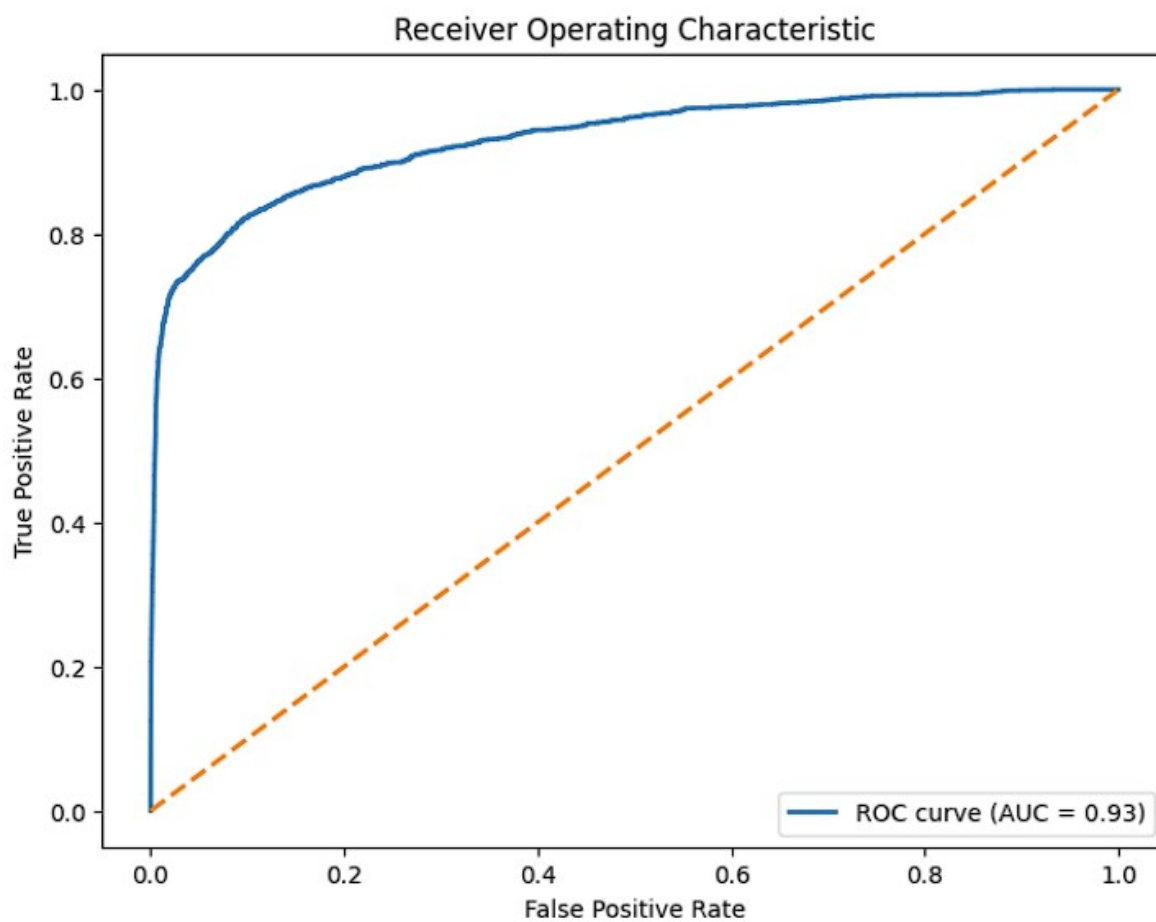


Рисунок 8

3.3 Ссылка на submission в leaderboard

[Loan_Approval_Predcition](#)