Путь поиска решения

1. Изучил датасеты, количество числовых и категориальных признаков.

   Пытался отобрать числовые признаки по матрице корреляции. Пытался отобрать категориальные признаки по количеству уникальных значений. Удалил колонки, у которых больше половины значений это `Nan`.

2. Для категориальных признаков пробовал различные методы кодирования: `One Hot, Label, Target`; Для числовых признаков пробовал нормировку.

3. Начал с моделей `catboost` и `xgboost`: предсказывали хорошо, но были не уверены в своих предсказаниях.

4. Попробовал в качестве модели `Random forest` с разными методами кодирования, `SVM` с различными ядрами, `нейросеть` с разным количеством слоёв, `логистическую регрессию`, различные комбинации `стекинга`. Признаки выбирал с помощью `catboost`, `xgboost`, `Random Forest`.

5. В итоге лучше всего себя показала модель случайного леса с признаками, подобранными с помощью `xgboost`, методом кодирования `Target Encoding`. Пропущенные значения в столбцах заменял средним. Гиперпараметры подбирал руками.

6. На тестовой выборке `mean log loss` $= 0.3$.