Анализ оттока клиентов

Кузнецов Г.И.

Введение

Любой бизнес хочет максимизировать количество клиентов.

Для достижения этой цели важно не только попытаться привлечь новых, но и **сохранить уже существующее**. Удержать клиента дешевле, чем привлечь нового.

Кроме того, новый клиент может оказаться слабо заинтересованным в услугах бизнеса, и тогда с ним будет сложно работать, поскольку у старых клиентов уже есть необходимые данные для взаимодействия с сервисом.

Цель: провести разведочный анализ данных оттока клиентов.

Задачи:

- Исследование зависимости и формулирование гипотез
- Проверка гипотез
- Построение моделей для прогнозирования оттока
- Сравнение качества моделей

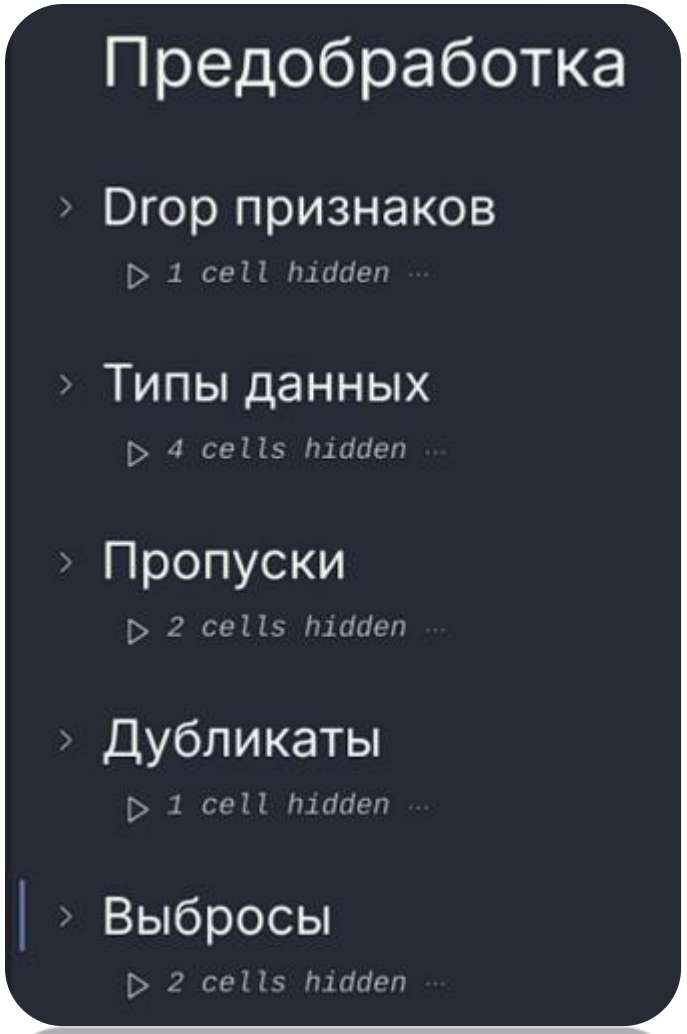
Описание данных

В данных содержится информация примерно о **шести тысячах пользователей**, их демографических характеристиках, услугах, используемых ими странах, долговечности услуг оператора, методе оплаты, стандартах оплаты.

```
data = pd.read_csv('telecom_users.csv')
data.head()
```

	Unnamed: 0	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneS
0	1869	7010-BRBUU	Male	0	Yes	Yes	72	
1	4528	9688-YGXVR	Female	0	No	No	44	
2	6344	9286-DOJGF	Female	1	Yes	No	38	
3	6739	6994-KERXL	Male	0	No	No	4	
4	432	2181-UAESM	Male	0	No	No	2	

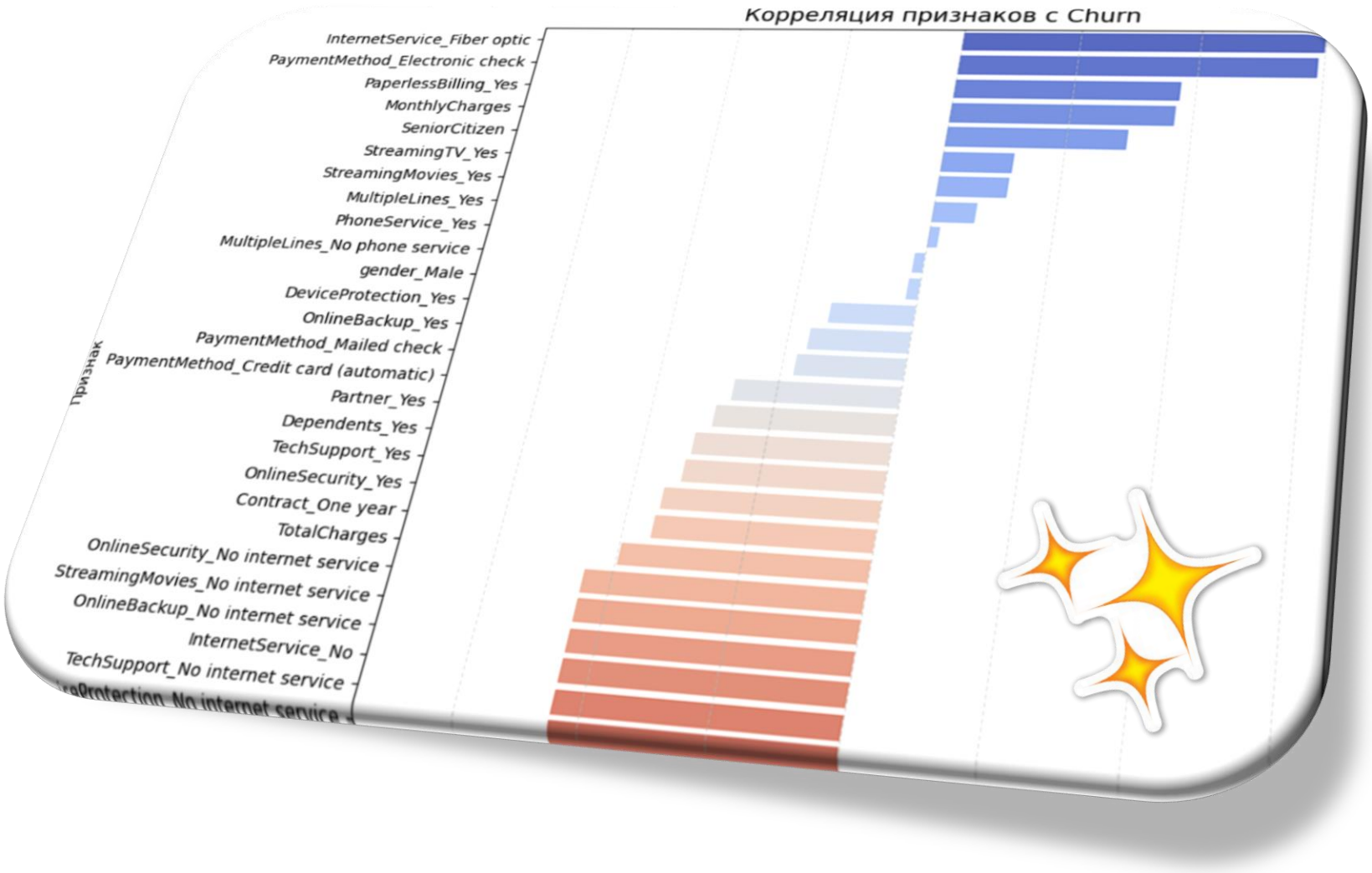
Предобработка



Данные оказались достаточно чистыми, однако некоторые корректировки пришлось сделать:

- убрать колонки с ID и CustomerID (как минимум на время обучения)
- обработать колонку Churn как целевую
- поменять тип данных для признака TotalCharges

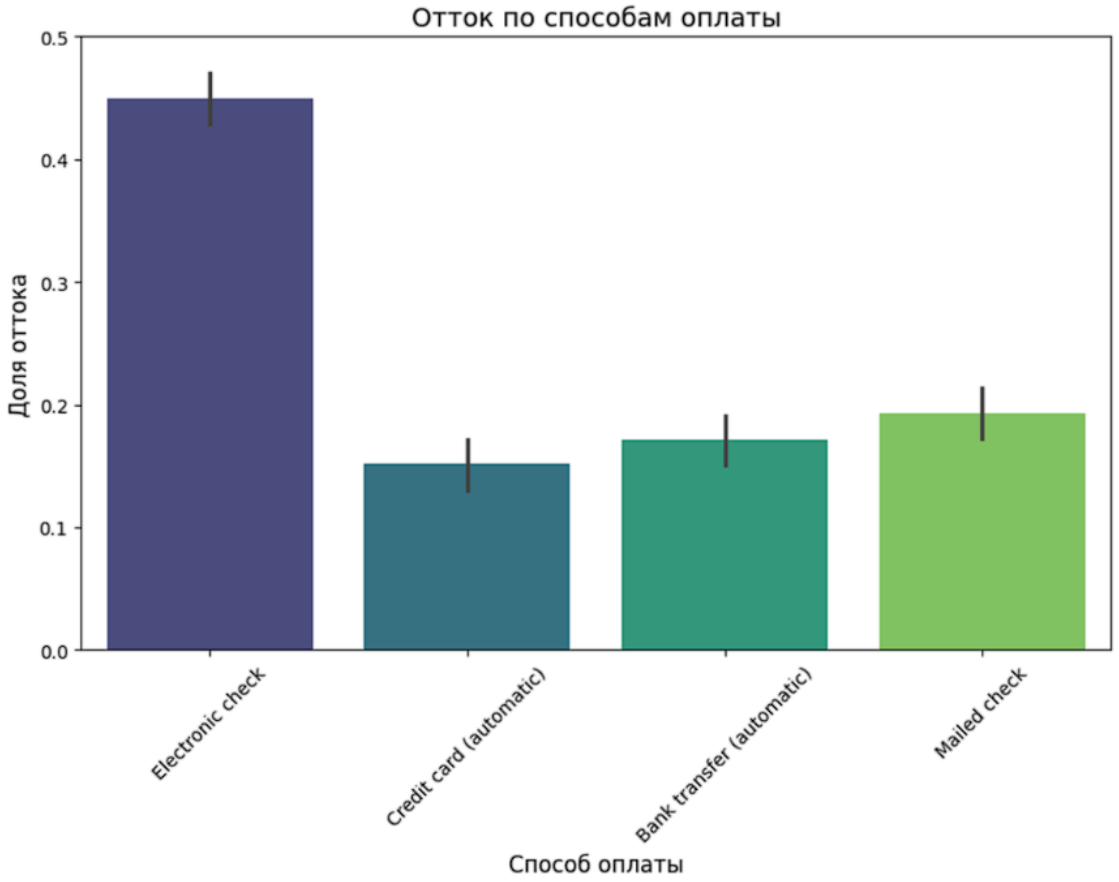
Гипотезы: построение

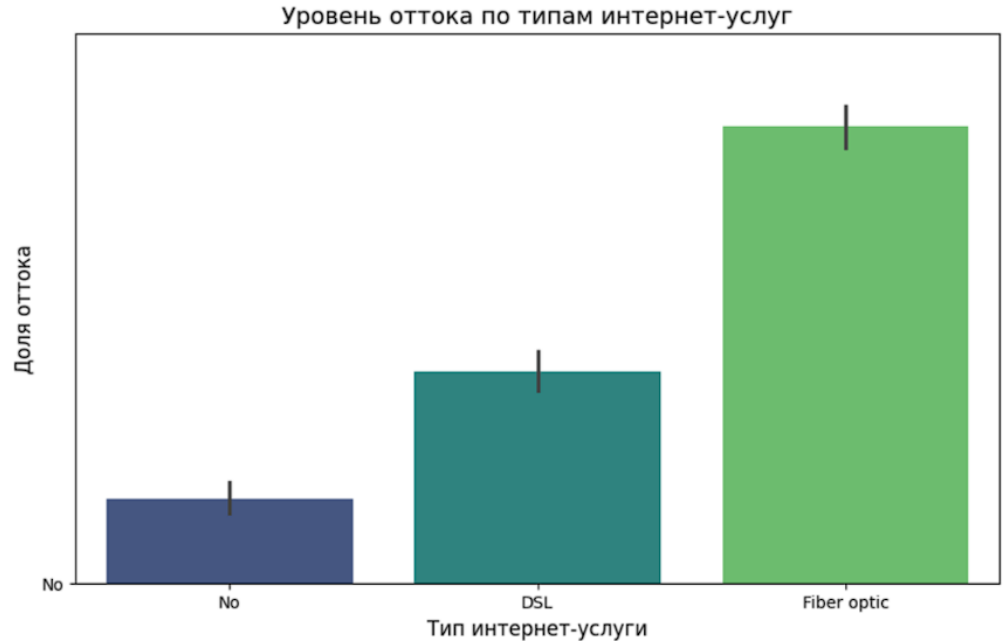


Гипотезы: текст

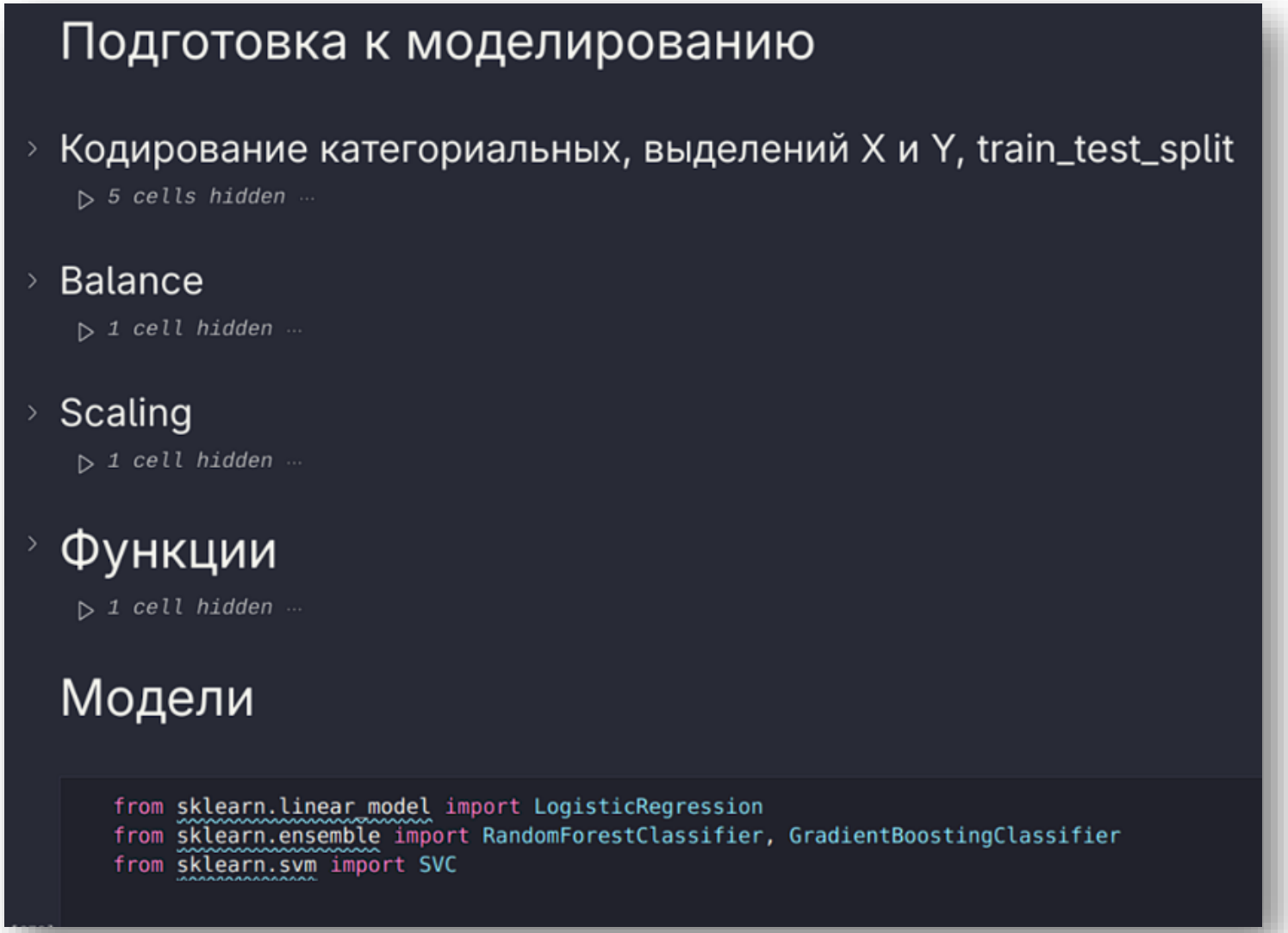
1. Гипотеза: Клиенты с оптоволоконным интернетом (Fiber optic) чаще уходят, несмотря на высокую скорость, из-за высокой стоимости или нестабильности сервиса.
   Как проверить:
   - Сравнить средние MonthlyCharges для Fiber optic vs. другие значения
   - Построить график: sns.barplot(x='InternetService', y='Churn', data=df).
2. Гипотеза: Клиенты, использующие электронные чеки (Electronic check), чаще уходят, потому что этот метод менее удобен или не поддерживает автоматические платежи.
   Как проверить:
   - Сравнить % оттока между Electronic check, Credit card и Bank transfer.
3. Гипотеза: Клиенты с безбумажными счетами чаще уходят, так как они более технологичны и чувствительны к альтернативам.
   Как проверить:
   - Сравнить отток в группах PaperlessBilling_Yes и No.
   - Проверить, связано ли это с возрастом (SeniorCitizen), у нас есть параметр является ли клиент пенсионером.
4. Гипотеза: Клиенты с высокими MonthlyCharges чаще уходят из-за несоответствия цены и качества.
   Как проверить:
   - Разделить клиентов на 3 группы по расходам (низкие/средние/высокие) и сравнить отток.
   - Построить sns.histplot(x='MonthlyCharges', hue='Churn', data=df).
5. Гипотеза: Клиенты, которые недавно подключились (tenure < 12 месяцев), чаще уходят, так как еще не успели привыкнуть к сервису или столкнулись с "разочарованием после покупки".
   Как проверить:
   - Разделить клиентов на группы по tenure (например, 0–6, 6–12, 12+ месяцев).
   - Построить график
6. Гипотеза: Клиенты с многоканальной связью (MultipleLines) чаще уходят, потому что это дорогая услуга с низкой добавленной ценой.
   Как проверить:
   - Сравнить средние MonthlyCharges для клиентов с MultipleLines_Yes и No.
   - Построить график: sns.countplot(x='MultipleLines', hue='Churn', data=df).

Проверка гипотез





Модели и метрики



Метрики:

cohen_kappa_score,

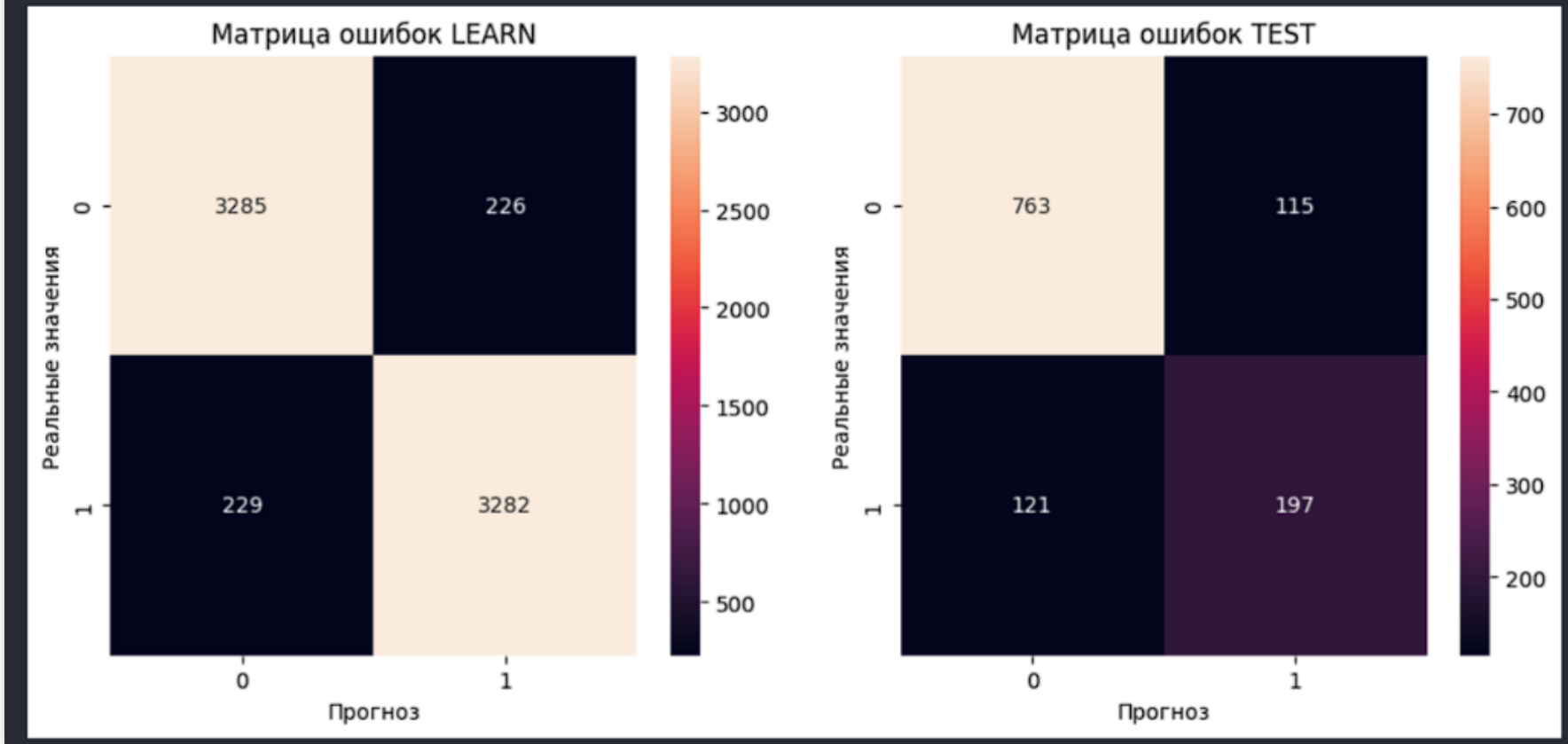
accuracy,

precision,

recall

Выбор лучшей модели

```
Начинаем подбор параметров для GradientBoosting...
Fitting 5 folds for each of 108 candidates, totalling 540 fits
Лучшие параметры: {'learning_rate': 0.2, 'max_depth': 4, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200}
Learn: kappa=0.8704, acc=0.9352, pre=0.9352, rec=0.9352, f1=0.9352, support=None
Test:  kappa=0.4915, acc=0.8027, pre=0.8015, rec=0.8027, f1=0.8021, support=None
```



Выводы и интерпретация результатов

В результате сравнения моделей мы видим, что самой точной оказалась модель **GradientBoosting**, оптимизированная **GridSearch**'ем, в целом я использовал несколько метрик и подошел к анализу довольно комплексно.
Также в результате проведенного анализа было установлено, что наиболее важные показатели при прогнозировании оттока — это **PaymentMethod, tenure и TotalCharges**.
Остальные параметры либо не оказывают прямого влияния на результат, либо сильно связаны друг с другом, поэтому может быть достаточно этих трех параметров для оценки, если появится какое-то желание оптимизировать модель.

Спасибо за внимание