ИНСТИТУТ ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
УНИВЕРСИТЕТА ИННОПОЛИС

УНИВЕРСИТЕТ ИННОПОЛИС

Прогнозирование оттока клиентов (Customer Churn Prediction)

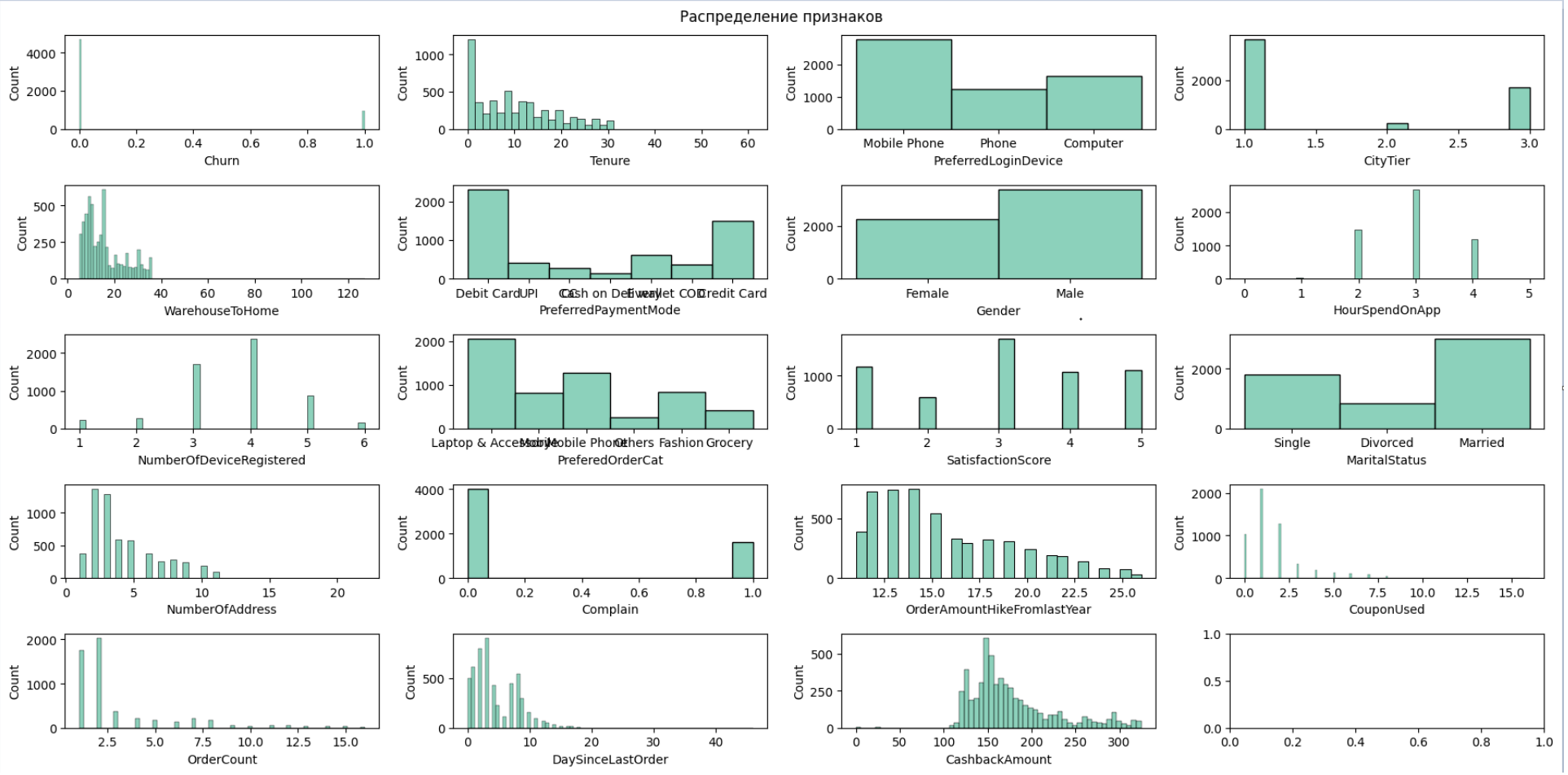
Описание задачи

Построить систему, которая предсказывает вероятность того, что клиент перестанет совершать покупки в течение следующих 3 месяцев

Описание набора данных

Название признака	Тип	Описание
CustomerID	строковый	Уникальный идентификатор клиента
Tenure	числовой (месяцы)	Срок, в течение которого клиент пользуется услугами компании (в месяцах)
PreferredLoginDevice	категориальный	Предпочтительное устройство для входа в аккаунт: «Mobile Phone» или «Computer»
CityTier	категориальный	Классификация города проживания клиента: 1 — крупный мегаполис, 2 — город среднего размера, 3 — небольшой город
WarehouseToHome	числовой (км)	Расстояние от ближайшего склада компании до адреса клиента (в километрах)
HourSpendOnApp	числовой (часы)	Среднее количество часов, проводимых клиентом в мобильном приложении в день
NumberOfDeviceRegistered	числовой (шт.)	Количество устройств, привязанных к аккаунту клиента
SatisfactionScore	категориальный (1–5)	Уровень удовлетворённости клиента, оценённый по пятибалльной шкале (1 — крайне недоволен, 5 — полностью доволен)
NumberOfAddress	числовой (шт.)	Количество адресов доставки, сохранённых в профиле клиента
Complain	бинарный (0/1)	Наличие обращения в службу поддержки с жалобой за последний период: 0 — нет, 1 — есть
OrderAmountHikeFromlastYear	числовой (%)	Процентный рост среднего чека по сравнению с предыдущим годом
Churn	целевая переменная, бинарная (0/1)	Факт ухода клиента: 0 — клиент активен, 1 — клиент прекратил использование сервиса

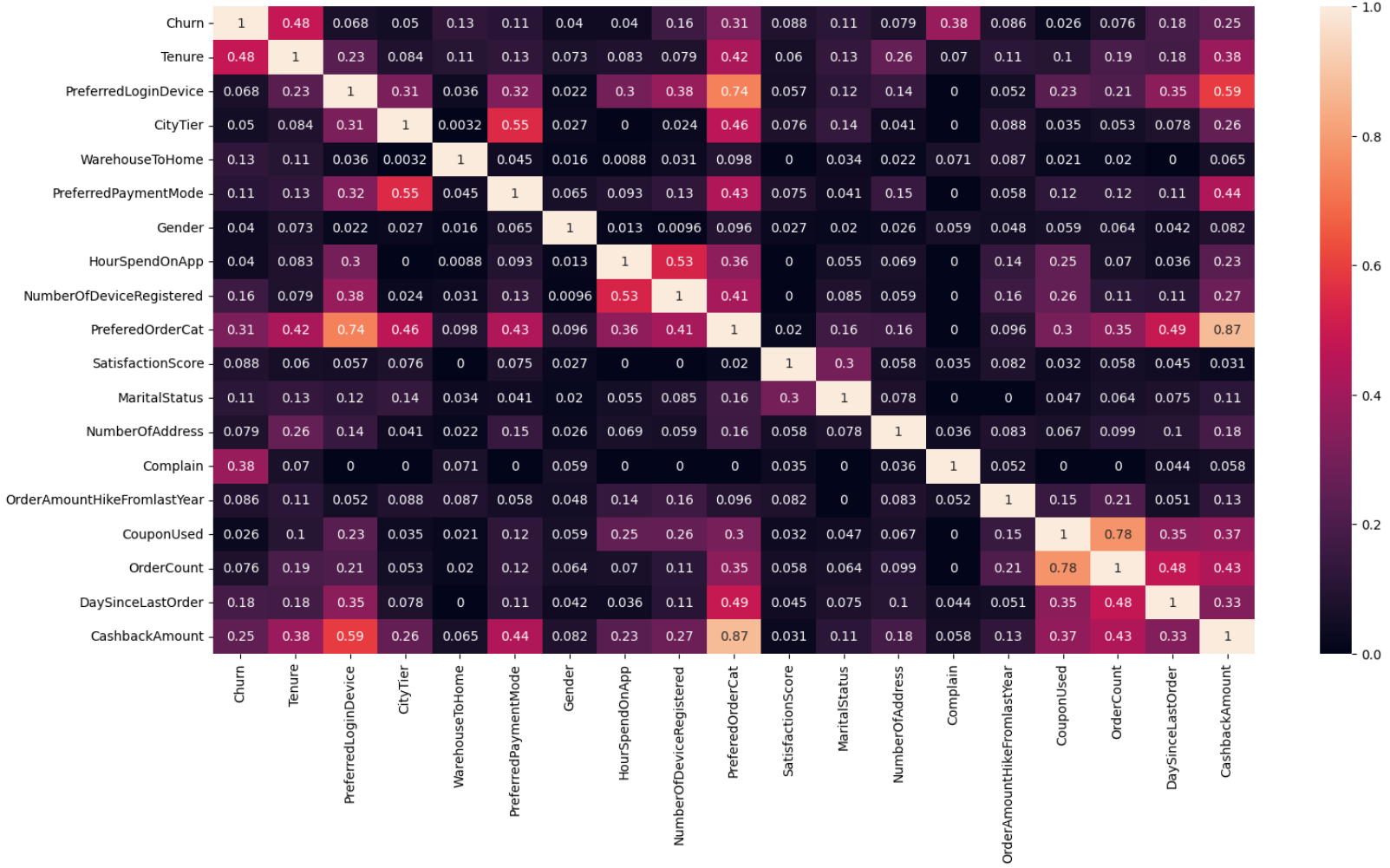
Распределение признаков



Анализ распределения признаков

1. Churn. Есть дисбаланс таргета. Клиентов, которые отстались, больше. 0 - это те, кто не ушли. 1 - кто, ушли.
2. Tenure. Большинство клиентов новые. Но есть, которые уже ранее пользовались услугами
3. PreferredLoginDevice. Большинство заказывают через телефон.
4. CityTier. 1 и 3 преобладают
5. WarehouseToHome. Время доставки, предположу, что в минутах. 99 % времени доставки меньше 35 минут. Есть огромный выброс, который больше 120 минут.
6. PrefferdPaymentMode. Способ оплаты. Большинство платят через карты, дебетувую или кредитную.
7. Male. Больше мужчин, чем женщин.
8. HourSpendOnApp. Большинство проводят около 3 часов в приложении.
9. NumberOfDeviceRegistred. Количество устройств у пользователя. У большинства 4.

Корреляция признаков



Анализ корреляции признаков

1. Видим, что некоторые признаки коррелируют между собой. Например количество заказов и количество использованных купонов. Это логично, значит, что клиенты пользуются купонами. Количество кешбека и колиество времени в организации. Логично, что у лояльных клиентов больше кешбека.
2. Признаки кореллируют с таргетом, Чем больше времени в организации клиент, тем меньше он подвержден "Оттоку". Если есть жалобы, то клиент больше подвержен оттоку.

Методология

1. Сначала используем логистическую регрессию, как базовую модель. Для корректной работы стандартизируем числовые данные и переведем категориальные признаки в вещественные при помощи one-hot-encoder. Посмотрим на метрики precision, recall,f1,roc-auc время инференса.
2. Попробуем 3 модели – catboost, RandomForest, MLP. Выберу лучшую из них по методике выбора лучшей.

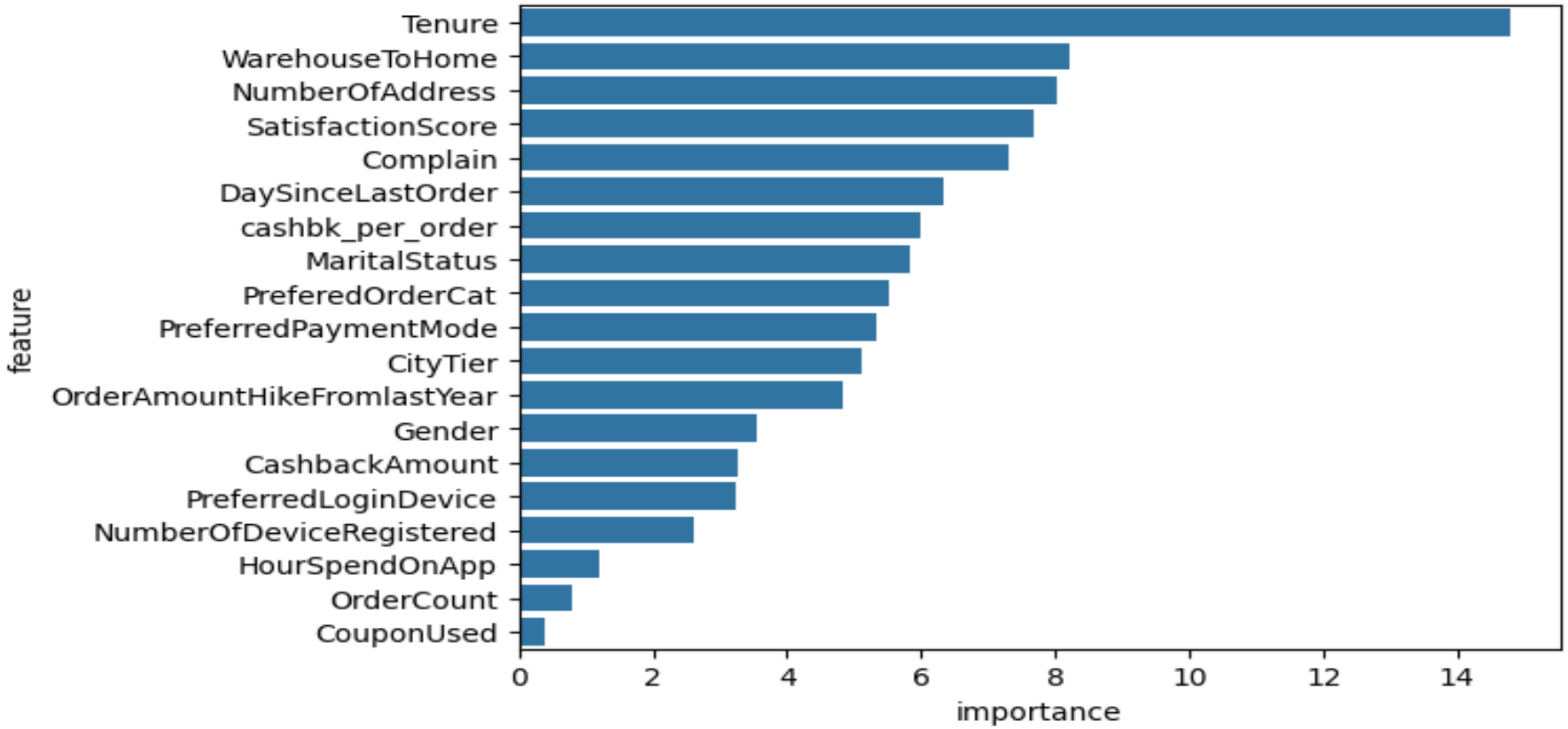
Результаты моделей при параметрах по умолчанию

№	Model	Precision	Recall	F1	ROC_AUC	time_inferense(ms)
0	Catboost	0.9420	0.8904	0.9155	0.9933	0.0107
1	RandomForest	0.9667	0.7945	0.8722	0.9923	1.2502
2	MLP	0.8030	0.7260	0.7626	0.9747	0.0096
3	LogReg	0.7037	0.5205	0.5984	0.8851	0.0105

Результаты моделей после гиперпараметрической оптимизации

№	Model	Precision	Recall	F1	ROC_AUC	time_inferense(ms)
0	Catboost	0.974	0.9934	0.9836	0.999	0.005
1	RandomForest	0.9823	0.8632	0.9189	0.997	1.203
2	MLP	0.8627	0.7631	0.8098	0.9823	0.001
3	LogReg	0.7037	0.5205	0.5984	0.8851	0.0105

Важность признаков



Анализ важности признаков

1. Видно, что самый важный признак это Tenure - сколько клиент находится в организации. Это логично, старые клиенты более лояльны и они не уходят, а новые клиенты могут прийти за конкретным товаром или купить по акции товар и уйти.

2. WarehouseToHome - то же важно, вероятно из - за того, что удаленность от склада влияет на время доставки товаров.

3. SatisfactionScore – важно, чтобы был высоким, чтобы клиент не уходил

4. Complain – обращение в поддержку важно. Если клиент не обращался, то значит все хорошо, меньше риск, что клиент уйдет.

5. HourSpendOnApp – чем меньше времени, тем выше риск ухода

На основе анализа датасета поведения клиентов e-commerce-платформы была разработана и обучена модель машинного обучения, способная прогнозировать вероятность оттока клиента. Модель продемонстрировала высокое качество предсказаний, что подтверждает её пригодность для практического применения.

ИНСТИТУТ ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
УНИВЕРСИТЕТА ИННОПОЛИС

Спасибо за внимание!

Контакты

Сайт