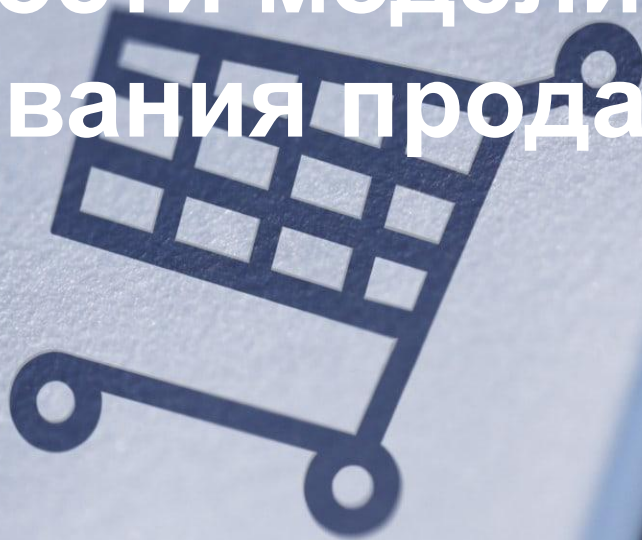


Построение и анализ эффективности модели прогнозирования продаж



Цель проекта

Проведение анализа данных с последующим прогнозированием успешности продаж, с целью разработки эффективных стратегий для увеличения объема продаж и максимизации прибыли.



Исходные данные

Amazon Sale Report.csv

Order ID	Идентификатор заказа
Date, Status	Дата заказа, Статус заказа
Fulfilment	Метод выполнения
Sales Channel	Канал продаж
ship-service-level	Уровень обслуживания доставки
Style	Стиль
SKU	Уникальный идентификатор товара
Category, Size	Категория товара, Размер товара
ASIN	Уникальный идентификатор товара на Amazon
Courier Status	Статус курьера
Qty	Количество товаров в заказе
Amount	Сумма заказа
ship-city, ship-state	Город доставки, Штат доставки
ship-postal-code, ship-country	Почтовый индекс доставки, Страна доставки
promotion-ids	Идентификаторы акций/промо



- 1. Исследование и анализ имеющихся данных***
- 2. Подготовка данных к машинному обучению***
- 3. Подбор оптимальной модели на тренировочных данных***
- 4. Подбор гиперпараметров модели машинного обучения***
- 5. Проверка модели на тестовых данных***
- 5. Итоги и выводы***

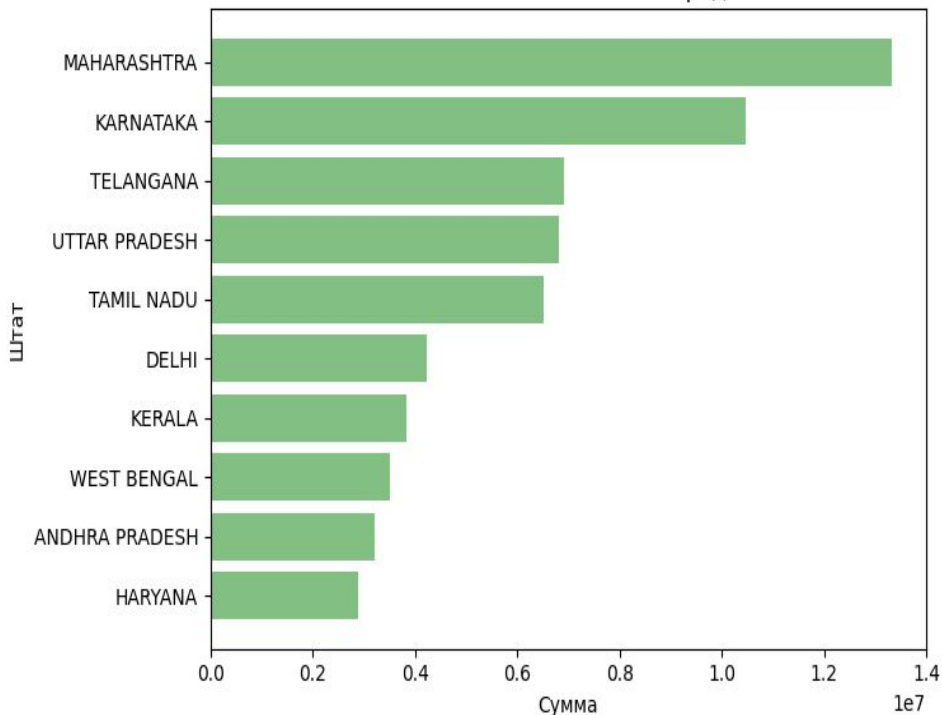
В данном проекте решаем задачу классификации.

Целью построения модели обучения является прогнозирование факта успешной продажи, а также выявление значимых характеристик сделки, влияющих на решение покупателя отказаться от заказа или нет.

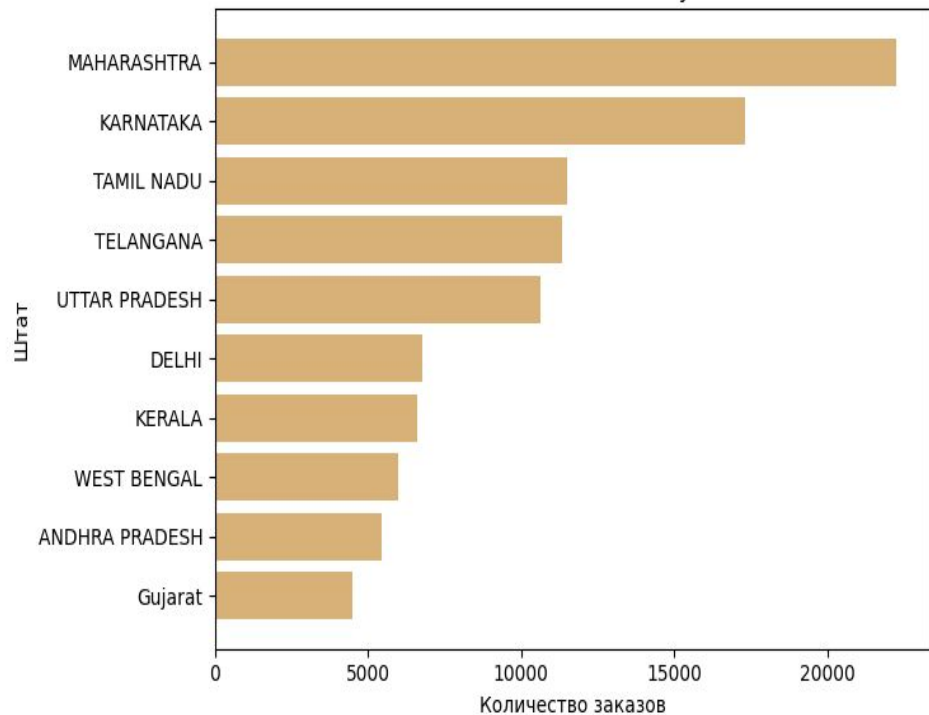


Выделим топ-10 штатов по объему продаж в стоимостном и количественном выражении

Топ-10 штатов стоимости продаж

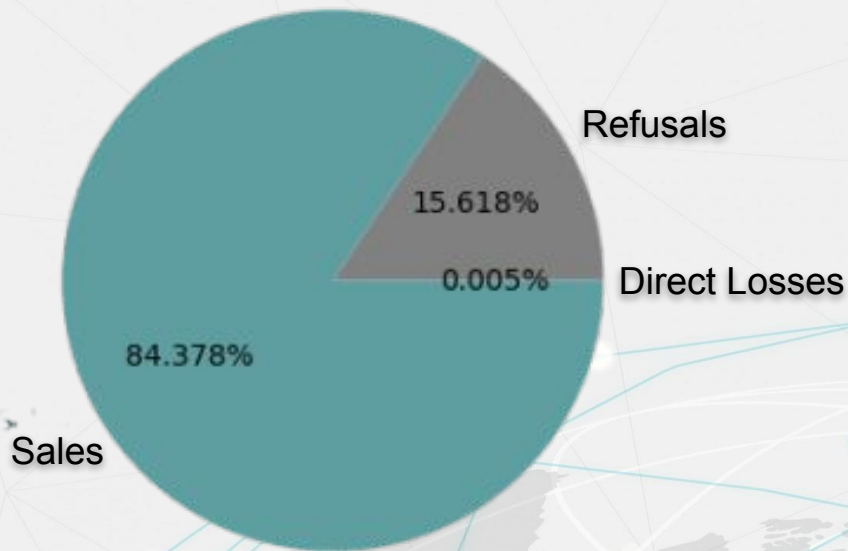


Топ-10 штатов по количеству заказов

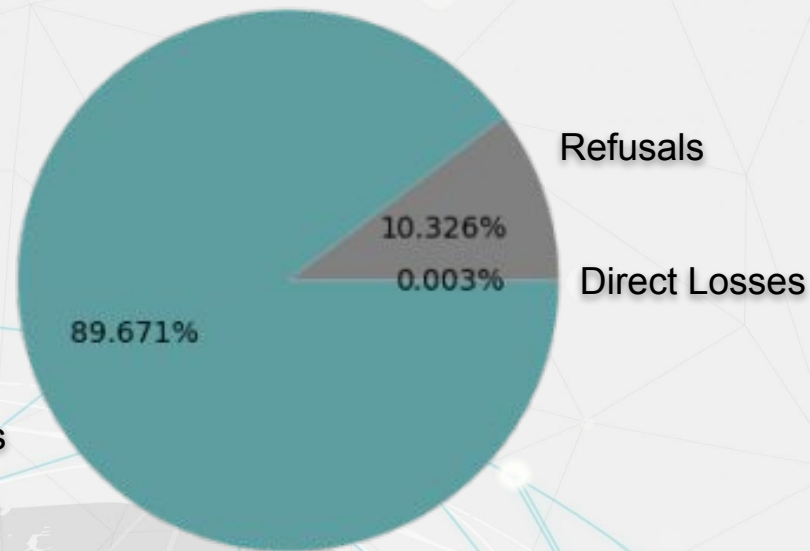


Объединим статусы заказов в 3 основные группы (продажа, отказы и прямые потери) и посчитаем удельный вес каждой группы в общем объеме данных.

Количество заказов по статусу



Стоимость заказов по статусу



Видим, что прямые потери составляют незначительную долю в общей массе заказов как в количественном выражении, так и в стоимостном. Поэтому при обучении модели эти данные будем игнорировать.

Средний процент отказов в общей выручке составляет 10,33%. Посмотрим, как данный процент распределяется между топ-10 штатов.

SHIP-STATE	REFUSALS _PERCENTAGE
ANDHRA PRADESH	14.022680
DELHI	10.362352
HARYANA	10.724739
KARNATAKA	10.052737
KERALA	15.117573
MAHARASHTRA	10.623010
TAMIL NADU	11.156929
TELANGANA	11.837695
UTTAR PRADESH	13.164776
WEST BENGAL	12.246068



Как видно из таблицы, в некоторых штатах данный показатель достигает 15%. Необходимо проведение дополнительных мероприятий по выявлению и более детальному анализу причин отказов.

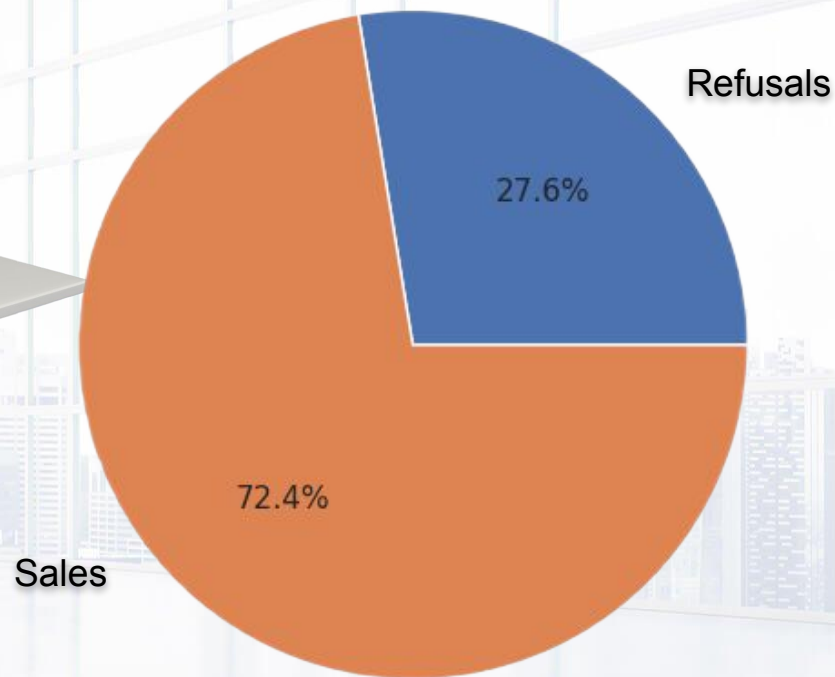
Удельный вес отмененных заказов в период отсутствия акций возрастает до 27,6%, что говорит о значимости данного признака

Поскольку уменьшение удельного веса отказов в периоды акций может быть положительным сигналом, бизнес может рассмотреть проведение более частых акций или разнообразных маркетинговых исследований для определения, какие конкретные акции наиболее успешны.

Это позволит бизнесу реагировать на изменения в потребительском спросе и принимать меры для оптимизации продаж.



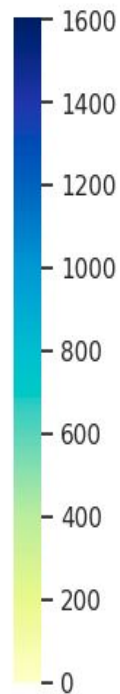
Стоимость заказов в разрезе статусов в период отсутствия акций



Количество отмененных заказов в различных категориях и размерах

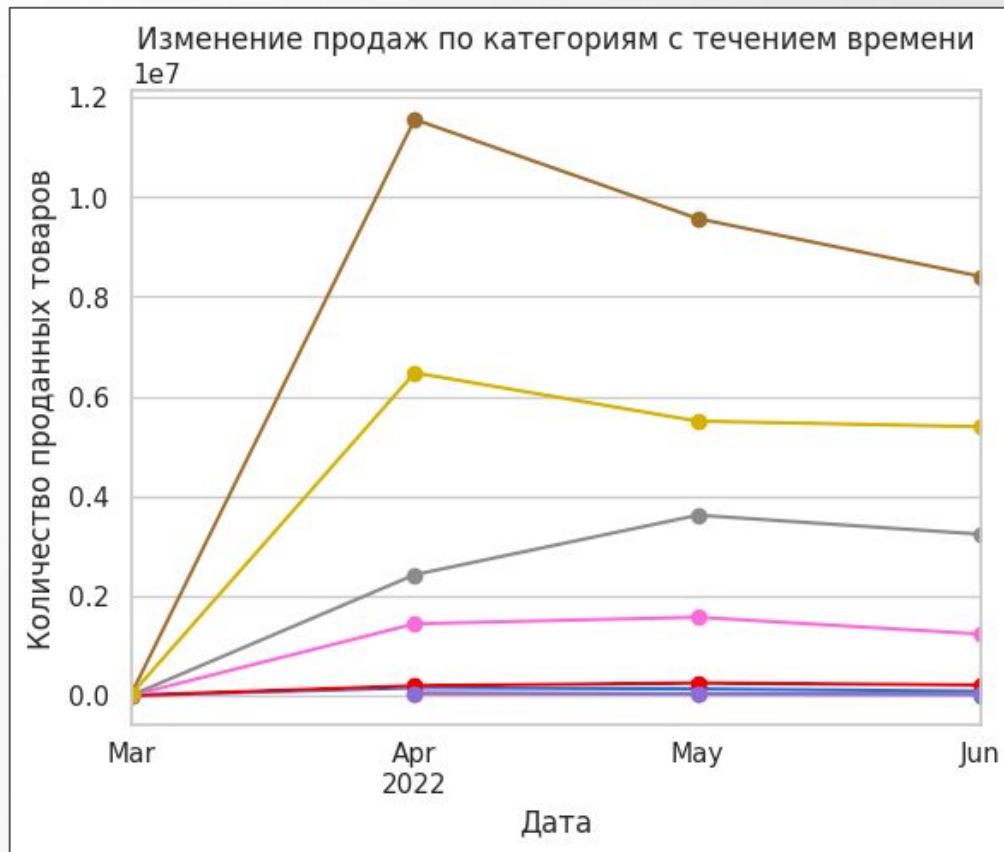
Категория

Blouse	0	0	0	0	31	11	26	20	24	5	11
Bottom	6	0	0	0	0	8	9	12	11	5	15
Ethnic Dress	21	0	0	0	0	36	22	22	23	14	23
Saree	0	0	0	0	22	0	0	0	0	0	0
Set	857	6	6	14	0	1316	1605	1270	1145	1041	922
Top	132	0	0	0	0	238	260	175	243	152	210
Western Dress	256	0	0	0	0	446	410	417	385	250	301
kurta	831	46	60	62	0	1446	1454	967	1359	567	1215
	3XL	4XL	5XL	6XL	Free	L	M	S	XL	XS	XXL
Размер											



Анализ количества отмененных заказов в разрезе категорий товаров и их размерности может помочь идентифицировать категории товаров или размерности товаров, где наблюдаются наибольшие проблемы с отменой заказов. Зная эти факторы, мы сможем принимать более точные решения, оптимизировать запасы, улучшить обслуживание клиентов и повысить общую прибыльность бизнеса.

Рассмотрим общую тенденцию объемов продаж с течением времени в разрезе категорий товаров



Предобработка данных

удаление строк
со статусом
прямых потерь

разделение даты
на день и месяц с
добавлением
столбцов

смена типа
данных "bool" на
"object"

удаление заказов с
нулевой стоимостью

замена пропусков
числовой переменной
"Amount" на 0

замена пропусков
категориальных
переменных на
"unknown"

добавление
признака
проведения акции

применение
encoding к
категориальным
переменным

отбор признаков,
разбиение данных на
тренировочные,
валидационные и
тестовые

В нашем тренировочном наборе данных с целевой переменной наблюдается сильный дисбаланс.
Применим один из методов борьбы с этим - RandomOverSampler. И проверим сбалансированность.



В процессе подбора оптимальной модели были использованы следующие алгоритмы машинного обучения:

LogisticRegression

KNeighborsClassifier

DecisionTreeClassifier

RandomForestClassifier

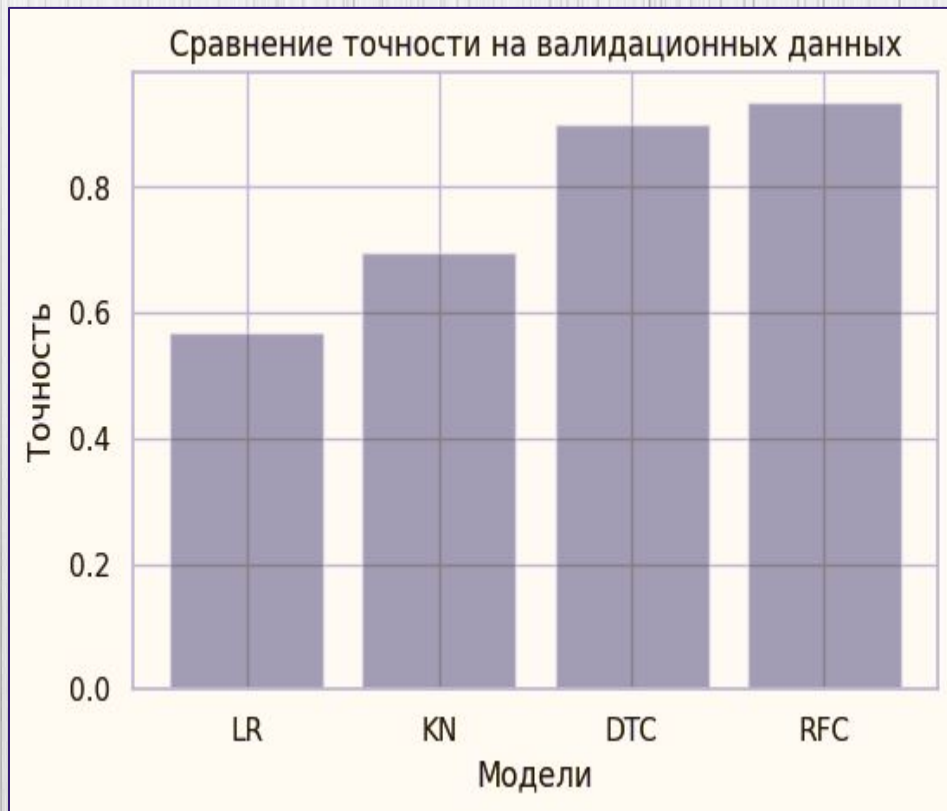
Вот такие результаты средней точности модели на всех фолдах, а также стандартное отклонение точности между фолдами:

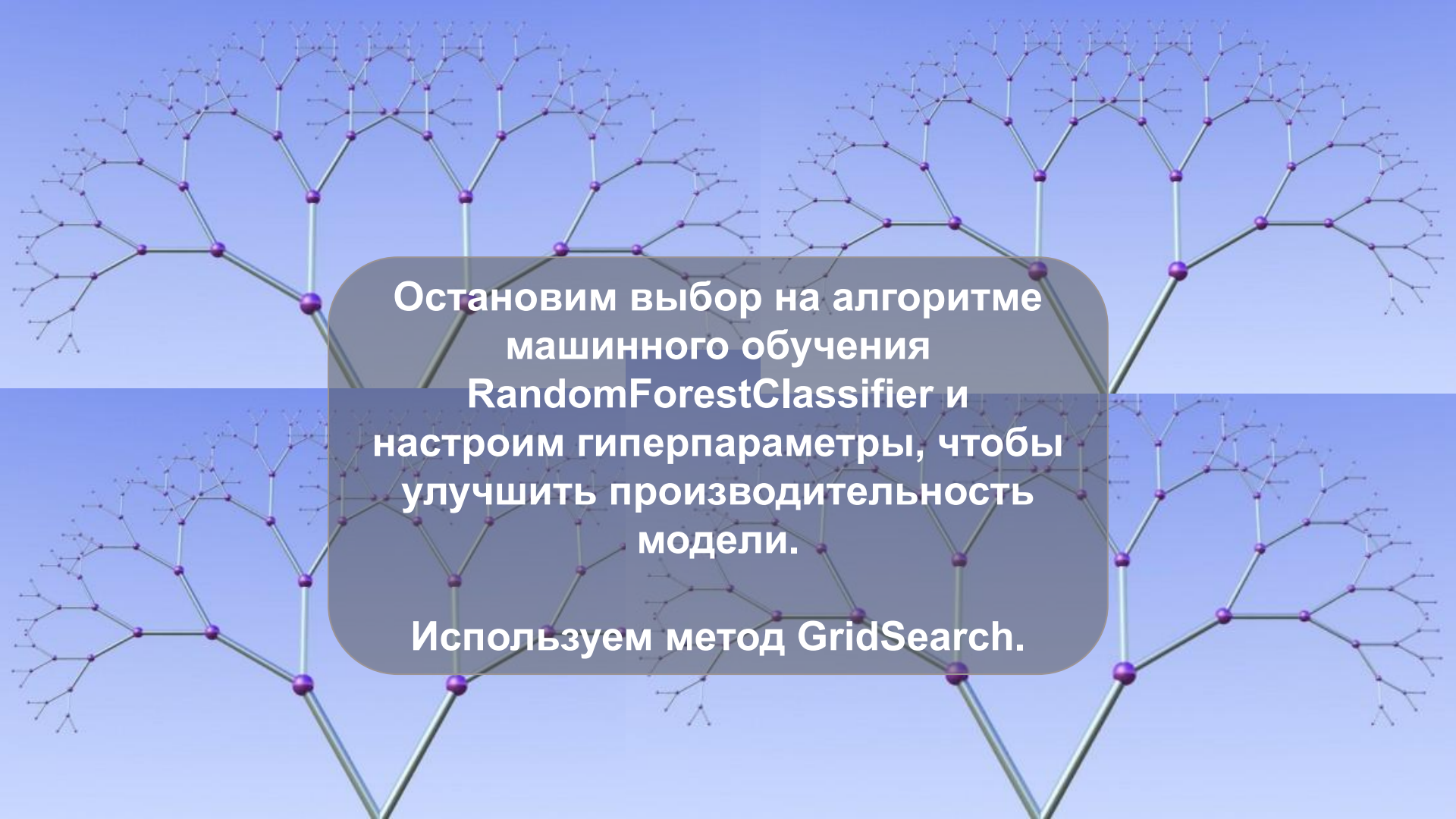
LR: train = 0.513 (0.002) / valid = 0.568

KN: train = 0.835 (0.003) / valid = 0.697

DTC: train = 0.934 (0.001) / valid = 0.901

RFC: train = 0.991 (0.000) / valid = 0.935





Остановим выбор на алгоритме
машинного обучения
RandomForestClassifier и
настроим гиперпараметры, чтобы
улучшить производительность
модели.

Используем метод **GridSearch**.

После обучения модели с
наилучшими гиперпараметрами

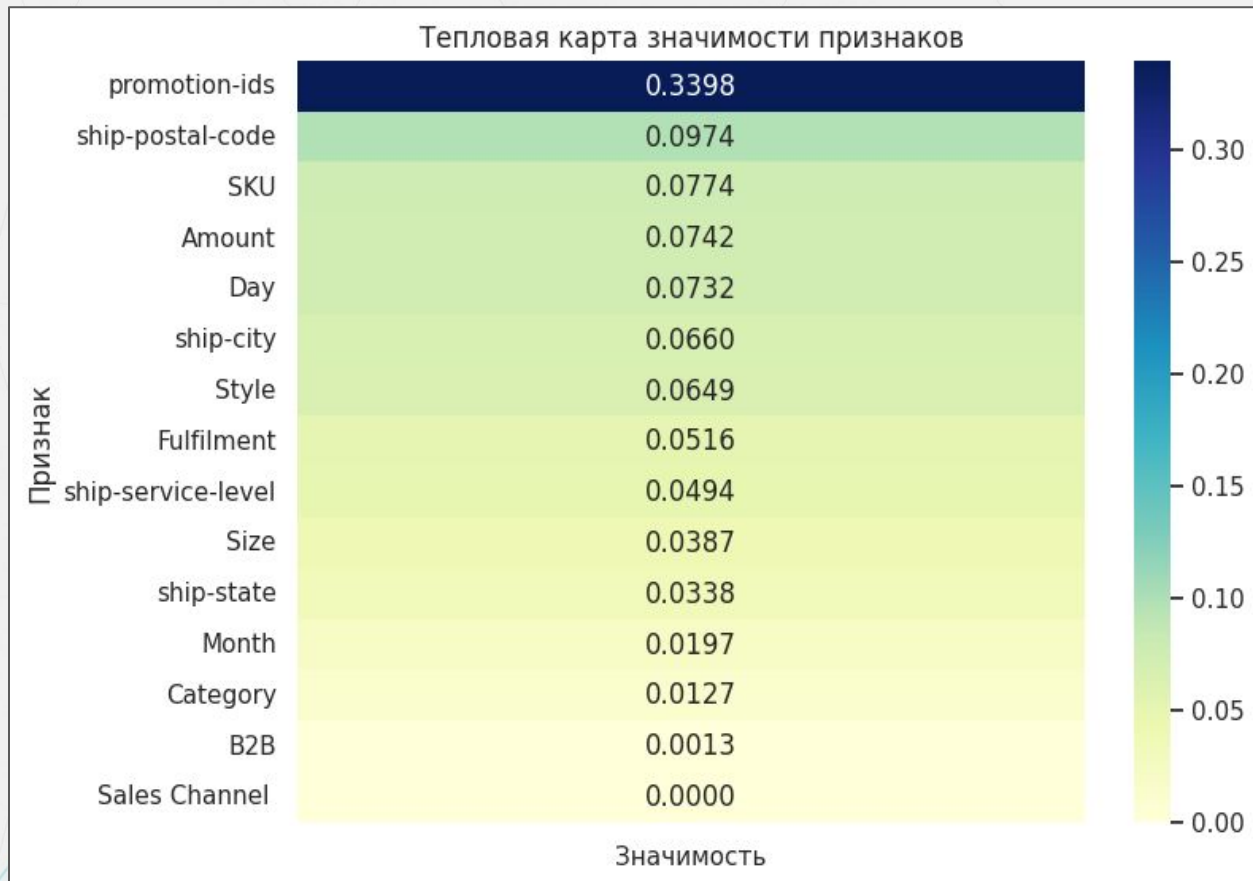
```
{'max_depth': None,  
'min_samples_leaf': 1,  
'min_samples_split': 3,  
'max_features': 'sqrt',  
'n_estimators': 200}
```

получили следующие метрики:



Accuracy	0.935	доля правильно классифицированных данных по отношению ко всем данным
Precision	0.944	доля истинных положительных предсказаний относительно всех положительных предсказаний
Recall	0.986	доля объектов положительного класса, которые модель определила правильно
F1 Score	0.964	гармоническое среднее между точностью и полнотой

Значимость признаков в нашей модели машинного обучения



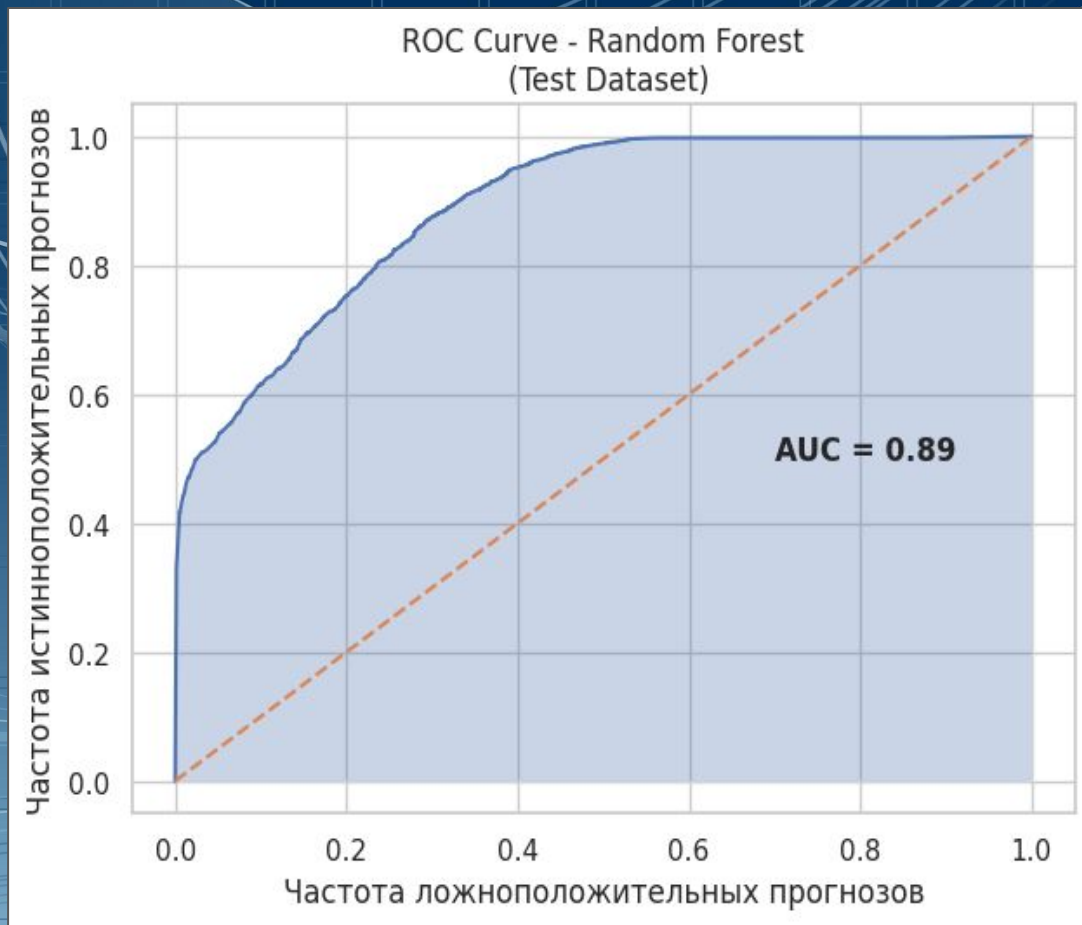
Как и предполагалось ранее, очень важную значимость в прогнозировании успеха в продаже имеет факт проведения либо отсутствия каких-либо акций в период продаж. В связи с чем бизнесу необходимо проводить более углубленный анализ маржинальности, чтобы лучше понимать влияние акций на финансовую эффективность.

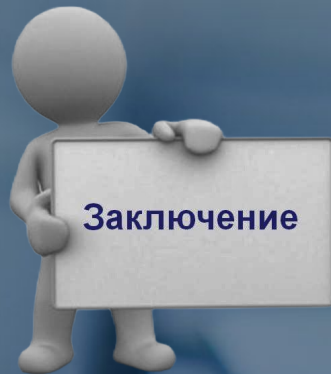
Также региональность как признак (ship-postal-code и ship-city) стоит на втором месте по значимости в прогнозировании успеха. Это обуславливает необходимость более детального изучения проблемных регионов, возможно проведения ненавязчивого анкетирования покупателей и выявления общих причин их негативного поведения.

ROC-AUC показала результат
немного хуже.

Однако...

Выбор метрики должен зависеть от
конкретных бизнес- или задачных целей.
В нашем случае, если ложноотрицатель-
ные прогнозы не дороже ложноположи-
тельных, более важными могут стать
другие метрики, такие как точность
(precision), полнота (recall), F1-мера. Они
обычно более информативны в случаях с
неравномерным распределением классов.





На основе анализа, который мы провели на этом наборе данных о транзакциях электронной торговли, мы смогли выявить несколько тенденций и закономерностей, которые могут быть полезны для оптимизации рекомендаций по продуктам, определения возможностей для дополнительных продаж, улучшения процессов и экономии затрат:

1. Бизнес может сосредоточить усилия на расширении продаж в топ-регионах или на поддержании уже успешных операций.
2. Выделение категорий продуктов, которые наиболее востребованы среди клиентов, может помочь с ассортиментным планированием и инвестициями в более успешные категории.
3. Результаты анализа временных тенденций могут помочь в управлении запасами, маркетинге и предложении акций в нужные периоды.
4. Бизнес может отслеживать количество отмененных заказов и выявлять потенциальные причины или факторы, которые могут способствовать отмене.
5. После создания модели машинного обучения бизнес может проводить A/B-тестирование, чтобы оценить, какие стратегии и действия могут увеличить вероятность успешных продаж.

*СПАСИБО
за внимание*

