ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ

АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

АЭРОФЛОТ
Российские авиалинии

Политех им Баумана

Задача 13. Рекомендательный сервис динамического прогнозирования спроса на авиарейсы

ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Содержание

ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Контакты

Максим Кузнецов

- Data Science
- @max_ii_m
- +7 999 552 59 17

Дмитрий Коротков

- Frontend
- @Nupellot
- +7 902 896 2417

Данил Кочура

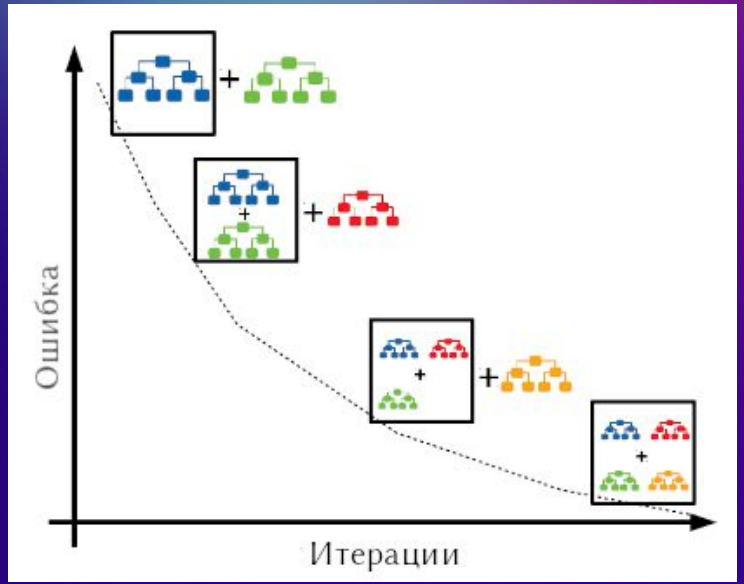
- Fullstack
- @kochuradanil
- +7 988 281 1407

Прогнозирующая модель

Модель - это логическое дерево, для построения которого перебираются разные параметры таким образом, чтобы оно давало наиболее точное предсказание.

Использовался способ обучения под названием градиентный бустинг.

Для отбора наилучшей модели использовалась функция потерь, считающая среднюю квадратичную ошибку.



ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Предсказание спроса

Предсказание спроса

Эта модель учитывает: аэропорт прилета и вылета, часть дня, когда самолет вылетит и прилетит, день недели, месяц, глубину бронирования, количество людей, которые полетят этим рейсом (предсказывается вспомогательной моделью)

	Feature Id	Importances
0	sscl1	43.999299
1	dtd	17.041089
2	equip	11.015320
3	month	9.445063
4	sorg	4.384703
5	tt_dep	4.089823
6	weekday	3.906906
7	tt_arr	3.746849
8	sdst	2.370948

Важность атрибутов для прогноза

Предсказание количества улетевших людей

Эта модель учитывает: аэропорт прилета и вылета, часть дня, когда самолет вылетит и прилетит, день недели, месяц, класс бронирования, салон.

Она нужна для улучшения результатов предсказания модели, предсказывающей спрос.

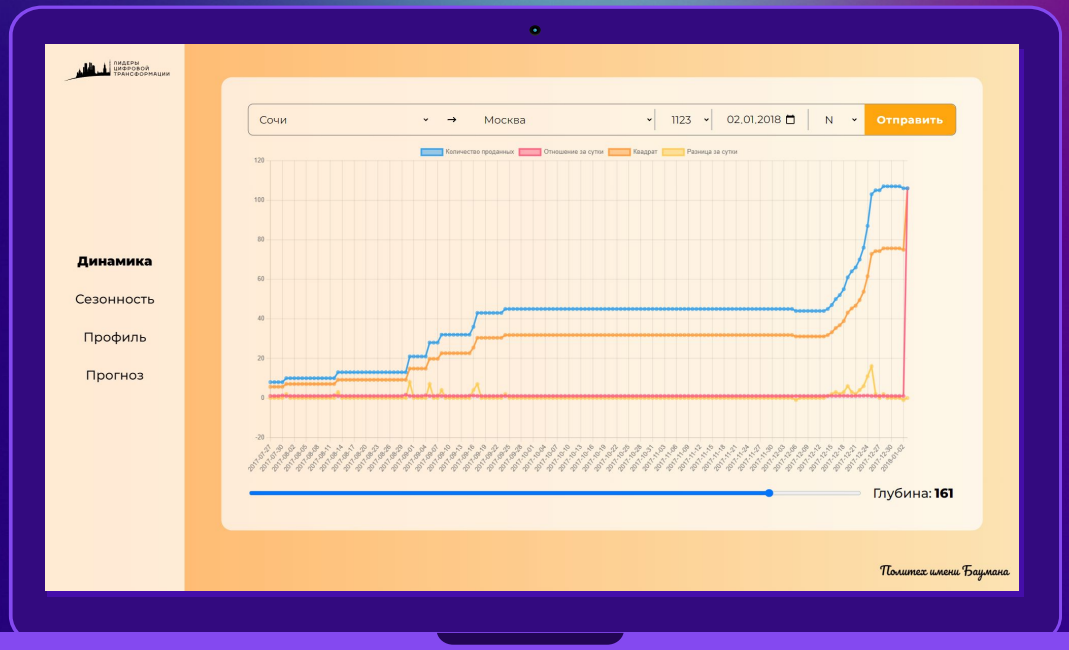
	Feature Id	Importances
0	seg_class_code	56.134148
1	month	12.340629
2	equip	9.114389
3	weekday	5.313152
4	tt_arr	3.661820
5	tt_dep	3.646850
6	sscl1	3.537700
7	sorg	3.379240
8	sdst	2.872073

Функциональная архитектура системы

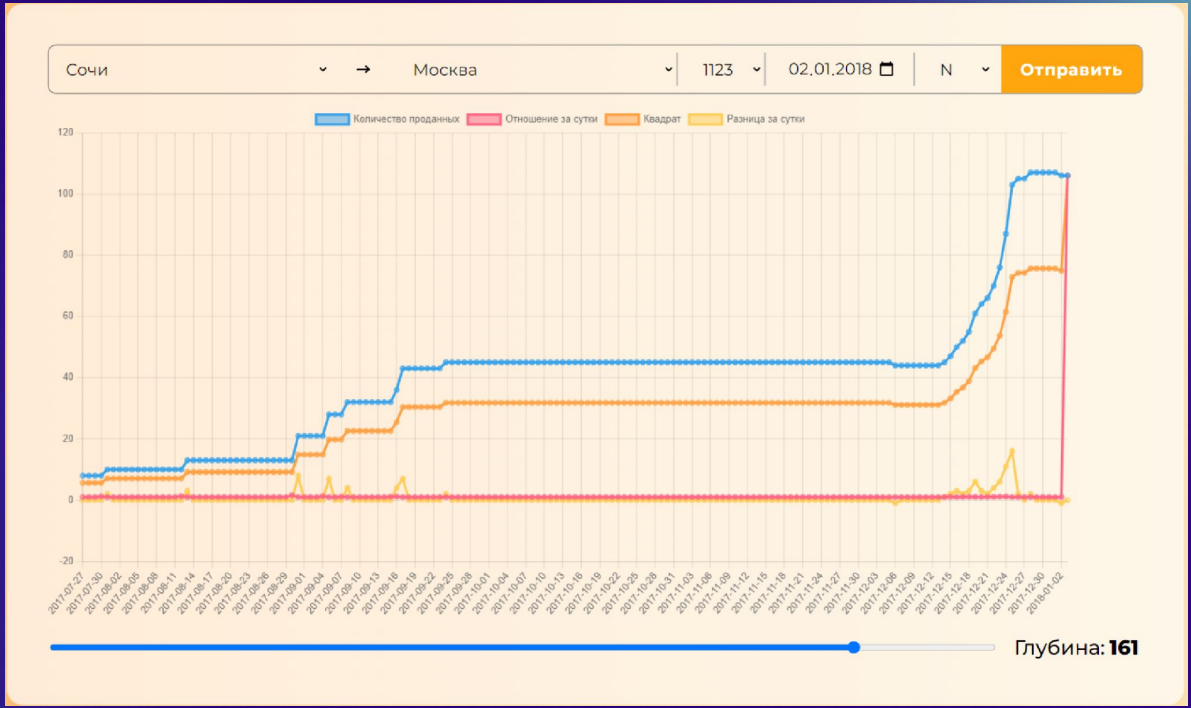
На сайте в левом меню представлена возможность выбора одного из четырех разделов:

«Динамика», «Сезонность», «Профиль», «Прогноз».

После выбора одного из данных разделов перед пользователем открывается форма для ввода информации об интересующем его рейсе. После выбора пунктов отправления и прибытия, даты вылета и класса обслуживания система динамически предложит пользователю номера рейсов для анализа. После нажатия кнопки «Отправить» пользователю будет показан график с информацией.



Динамика



Интерпретация получившегося графика

Синяя линия:

Суммарное количество проданных на данный момент билетов

Красная линия:

Отношение общего количества проданных билетов между соседними днями

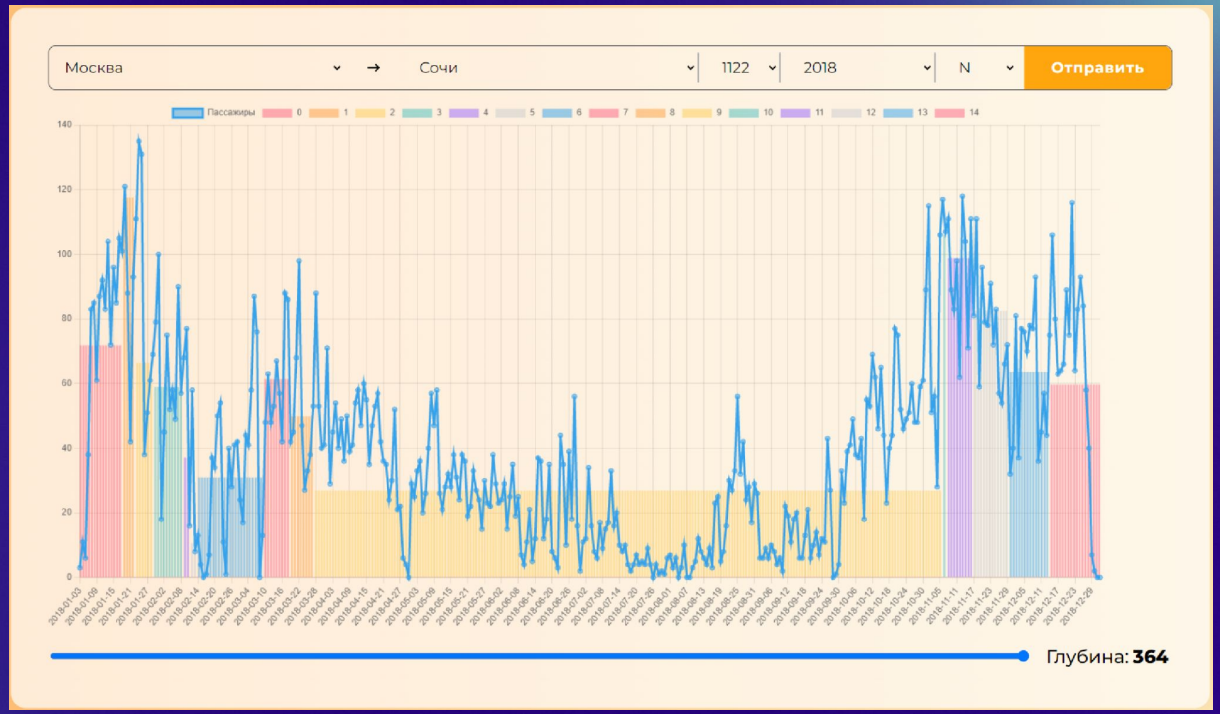
Оранжевая линия:

Среднее квадратичное соседних дней

Желтая линия:

Прирост билетов за день (Отрицательный в случае преобладания сдачи билетов над их покупкой)

Сезонность



Интерпретация получившегося графика

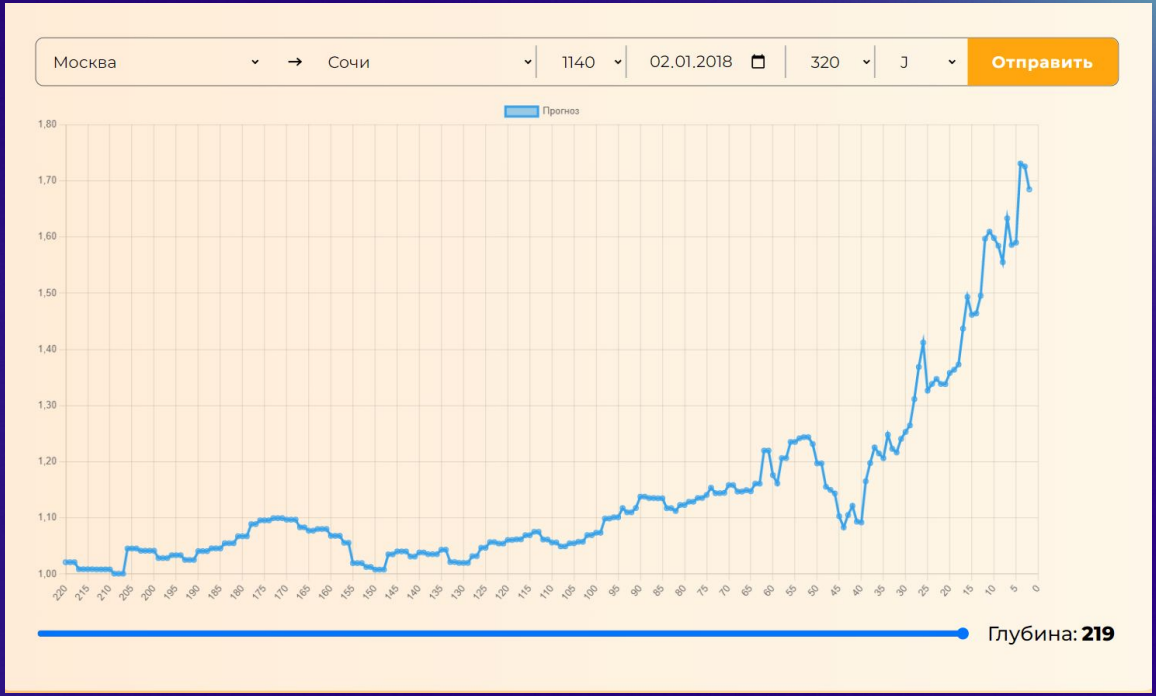
Синяя линия - количество пассажиров, вылетевших этим рейсом в конкретный день.

Прямоугольники разных цветов - определенные с помощью специального алгоритма сезоны.

Высота прямоугольника отражает среднее количество вылетевших за один день пассажиров в рамках текущего сезона.

Ширина прямоугольника отвечает за длительность сезона.

Прогноз



Интерпретация получившегося графика

Синяя линия:

Предполагаемый спрос на авиабилеты

Программно-техническая архитектура системы

Проект состоит из двух основных систем - заранее обученной ML-модели и веб-сервиса

- Серверная часть реализована на MySQL и PHP
- Клиентская часть, помимо самописных компонентов CSS и JavaScript, использует ChartJS и JQuery
- Модель реализована на языке Python и обучена посредством компонентов NumPy, Pandas и CatBoost
- Скрипт, который отвечает за прогнозирование, написан на Python с использованием упомянутых выше библиотек. Он взаимодействует с серверной частью приложения через консоль ОС, где запускается из-под интерпретатора PHP

Подробная файловая структура описана в Документации продукта

Технологии, задействованные в разработке

Pandas – библиотека для анализа данных на python

NumPy – продвинутая математика на python

CatBoost - библиотека для градиентного бустинга

Scikit learn – инструмент оценки ML-модели

Chart.js – инструмент для построения адаптивных графиков

Ресурсы системы

Эксплуатация

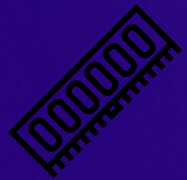
Для использования системы подойдёт любой современный web-браузер

Развертывание

Для развертывания системы на каком-либо из облачных сервисов понадобится в пределах 75ГБ дискового пространства для хранения исторических данных

Доработка

В случае необходимости дообучения модели понадобится мощное железо с обязательным наличием большого количества оперативной памяти



ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Сложности в разработке

Проблемы

Бэкенд

С точки зрения разработки серверной части веб-приложения основную сложность представляло развертывание базы исходных данных. Большое количество записей требовало индексации базы под распространенные запросы, что увеличивало ее потенциальный размер. Отсутствие и на локальном, и на виртуальном сервере достаточного количества дискового пространства, сильно замедляло разработку в первые дни.

ML-модели

Неудобные датасеты.
Не хватало ОЗУ для соединения классов и кабин
Из-за нехватки мощностей не удалось добавить все 26 класса бронирования в модель по предсказанию спроса.
Из-за нехватки времени и данных о полетах не удалось сделать более точные прогнозы и подобрать параметры для более эффективного обучения моделей

Решения

Фильтрация исходных CSV-файлов и частичная загрузка исходных данных.

Предобработал данные (убрал лишние пробелы, переписал колонки в нижнем регистре)
Сжал данные и удалил ненужные колонки, далее запустил цикл и, перебирая каждый класс, собрал датасет
Сделали предсказание количества полетевших людей.

Возможности развития системы

Как можно улучшить модель?

Добавить:

- Флаги с массовыми мероприятиями
- Флаг плохой сезонности в регионе
- Флаг о факте рецессии в экономике страны (для предсказания небольшой глубины)
- Сильный рост или падение индексов, около 10% за месяц
- Улучшенные предсказания по количеству полетевших людей
- Дообучить две модели
- Подобрать более эффективные параметры для обучения моделей
- Классы бронирования

Как можно улучшить сайт?

- Добавить авторизацию пользователей
- Добавить адаптивность
- Улучшить структуру вёрстки
- Улучшить динамичность формы
- По возможности интегрировать данные формы с данными внутренних систем аэрофлота

Оценка разработанных решений относительно мировых трендов

CatBoost

ML-модель обучалась с помощью библиотеки CatBoost, которая использует технологию градиентного бустинга. В представленных бенчмарках показано, что CatBoost не уступает аналогичным современным решениям.

	CatBoost	XGBoost	LightGBM
CPU (Xeon E5-2660v4)	527 sec	4339 sec	1146 sec
GTX 1080Ti (11GB)	18 sec	890 sec	110 sec

Dataset Epsilon (400K samples, 2000 features). Parameters: 128 bins, 64 leafs, 400 iterations.

scikit-learn

scikit-learn - библиотека для машинного обучения на python, широко используемая множеством современных компаний и активно снабжаемая обновлениями

Выводы

Статистика

В результате плодотворной работы нашей командой был разработан продукт, способный существенно облегчить жизнь сотрудникам Аэрофлота за счёт предоставления информативной статистики динамики бронирования и сезонности.

Прогноз

Имея данные за прошедшие периоды времени модель достаточно точно предсказывает спрос на авиабилеты в будущем.