2025

Портфолио

Писцова Ксения

Аналитик | Data Scientist

ПОРТФОЛИО

01

ГЕНДЕРНАЯ СТАСТИСТИКА В СФЕРЕ ОБРАЗОВАНИЯ

Проект в Power BI анализирует гендерную статистику преподавателей в Беларуси на уровнях начальной, средней школы и университета. Он включает визуализацию данных и выявление дисбалансов в составе преподавательского состава.

02

ПРОГНОЗ ПАССАЖИРОВ АВИАКОМПАНИИ

Проект посвящён прогнозированию количества пассажиров авиакомпании с использованием метода SARIMA. Он включает обработку данных, создание и обучение модели SARIMAX, оценку качества с помощью метрик MAE и RMSE, а также визуализацию результатов на графике.

03

ДИАГНОСТИКА ЗАБОЛЕВАНИЙ

Проект посвящён диагностике заболеваний с использованием машинного обучения. Он включает обработку данных, создание модели логистической регрессии и оценку её эффективности через метрики качества и визуализацию ROC-кривой.

04

АНАЛИЗ И ВИЗУАЛИЗАЦИЯ РЕЙТИНГОВ ТЕЛЕШОУ

Обработка данных, замена пропусков, сортировка по рейтингу и построение горизонтальной гистограммы для отображения 10 самых высоких рейтингов. Визуализация выполнена с использованием библиотеки Matplotlib.

05

ПАРСИНГ НОВОСТЕЙ

Скрипт для парсинга новостей с сайта Euronews. Используя библиотеки requests и BeautifulSoup, программа извлекает заголовки и ссылки на статьи, сохраняет их в CSV-файл. Скрипт также включает проверку доступности сайта и выводит количество найденных статей.

06

ДРУГИЕ

Проект с Hive
Проект с Docker
Модель классификации изображений

ksushapistsova@gmail.com

02

ГЕНДЕРНАЯ СТАСТИСТИКА В СФЕРЕ ОБРАЗОВАНИЯ

Проект посвящён анализу гендерной статистики преподавателей в Беларуси на различных уровнях образования: начальной, средней школы и университете.

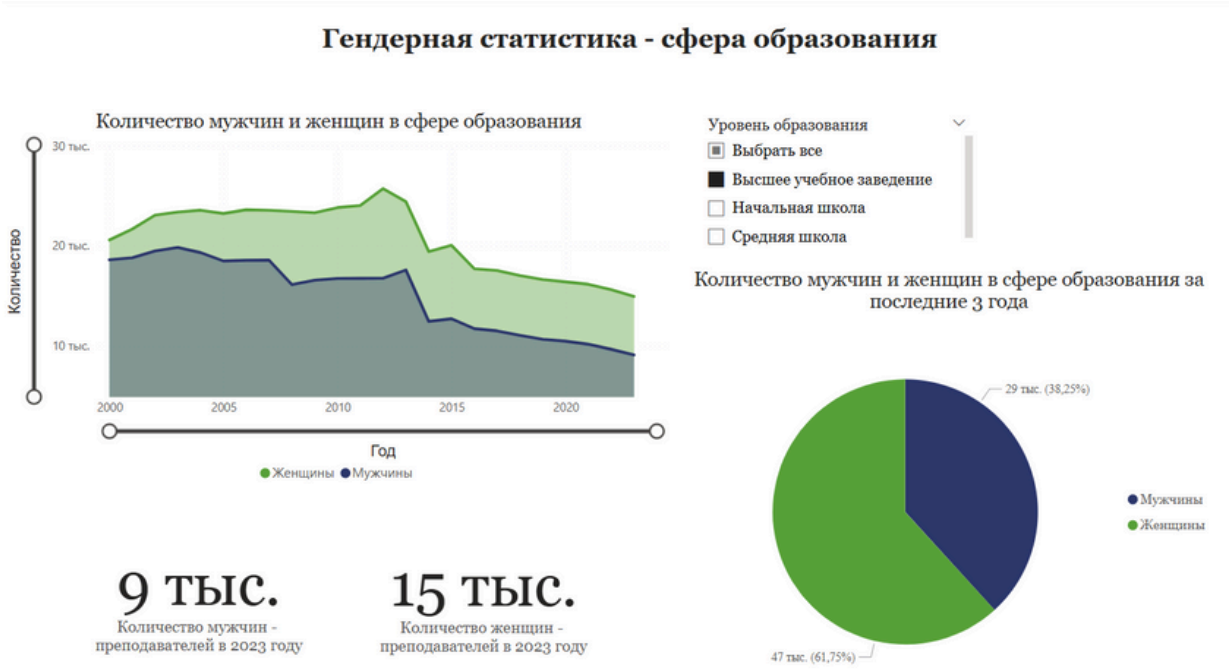
Он включает в себя:

1. Сбор и обработку данных о преподавателях, разбитых по образовательным ступеням.
2. Визуализацию гендерного распределения, что позволяет легко сравнивать количество мужчин и женщин на каждом уровне образования.
3. Анализ тенденций и выявление потенциальных дисбалансов в гендерном составе преподавательского состава.

Проект предоставляет наглядные отчёты и инсайты, способствующие пониманию текущей ситуации и поддержке инициатив по равенству и разнообразию в образовательной сфере.

Используемые технологии

- Power BI: для визуализации и анализа данных.
- Excel (Power Query): для предобработки данных.



02

ПРОГНОЗ ПАССАЖИРОВ АВИАКОМПАНИИ

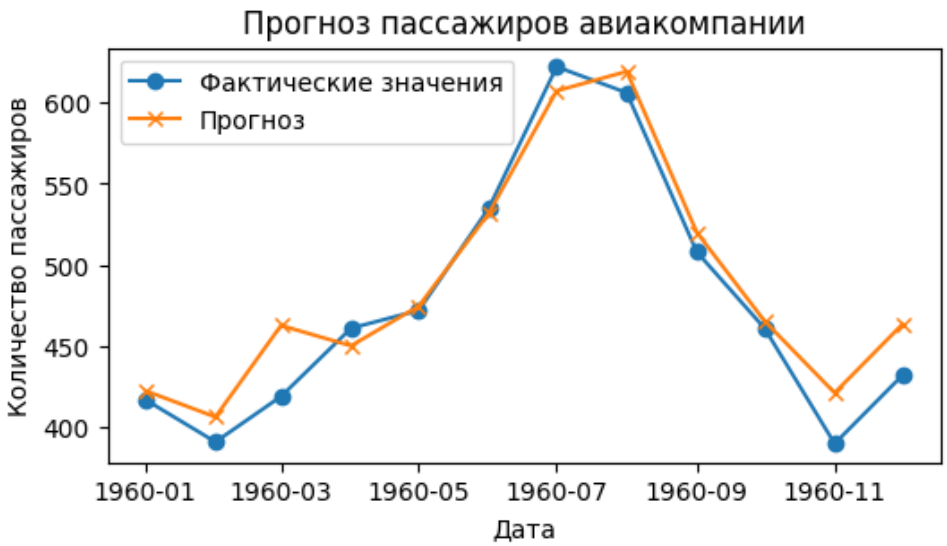
Проект посвящён прогнозированию количества пассажиров авиакомпании с использованием метода SARIMA (Seasonal AutoRegressive Integrated Moving Average).

Основные этапы работы:

1. Обработка данных: Загружаются данные из CSV-файла и преобразуются в формат временного ряда. Устанавливается частота индекса как "месяц".
2. Разделение данных: Данные делятся на обучающую выборку (все, кроме последних 12 месяцев) и тестовую (последние 12 месяцев).
3. Моделирование: Создаётся модель SARIMAX, которая обучается на обучающей выборке, и генерируются прогнозы на тестовой.
4. Оценка качества: Рассчитываются метрики MAE (средняя абсолютная ошибка) и RMSE (корень из средней квадратичной ошибки) для оценки точности прогноза.
5. Визуализация: Строится график с фактическими и прогнозируемыми значениями.

Технологии

- Python: Язык программирования, используемый для реализации проекта.
- Библиотеки:
  - pandas: Для обработки и анализа данных.
  - statsmodels: Для создания и оценки модели SARIMAX.
  - sklearn: Для вычисления метрик ошибок.
  - matplotlib: Для визуализации результатов.



03

ДИАГНОСТИКА ЗАБОЛЕВАНИЙ

Проект посвящён диагностике заболеваний с использованием машинного обучения. Цель проекта — разработать модель для предсказания вероятности заболевания на основе различных признаков.

Результаты:

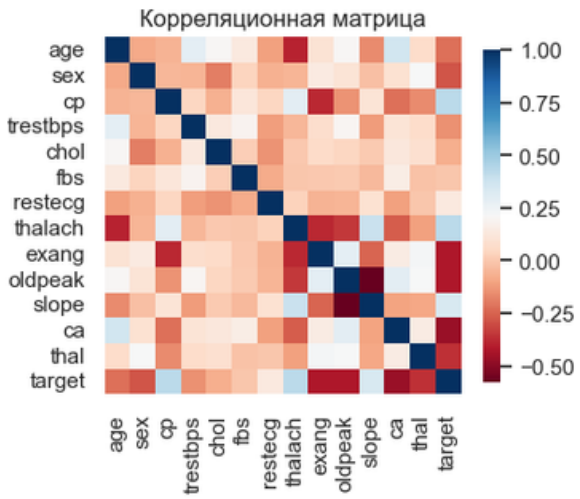
- Качество модели: Достигнуты высокие значения точности, полноты и F-меры, что свидетельствует о хорошей производительности модели.
- ROC-кривая: Построена ROC-кривая с вычисленным значением AUC, что позволяет визуально оценить качество классификации.

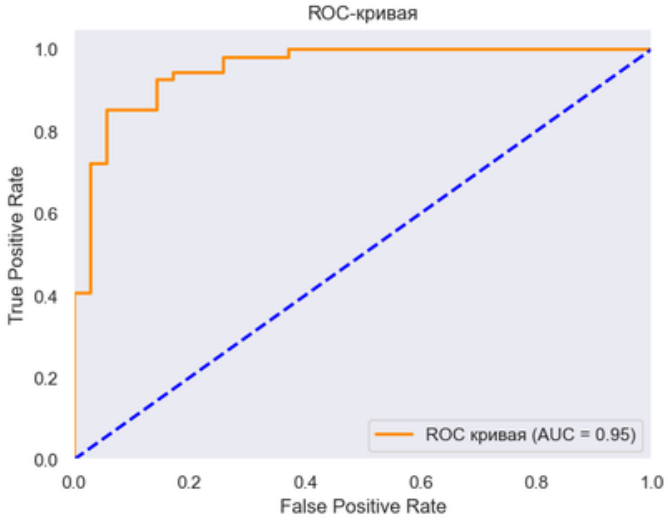
Технологии:

- Python: Основной язык программирования для реализации проекта.

- Библиотеки:
  - scikit-learn: Для построения и оценки модели.
  - matplotlib: Для визуализации ROC-кривой.

- Методы:
  - Логистическая регрессия для классификации.
  - Стандартизация данных с помощью StandardScaler.
  - Разделение данных на обучающую и тестовую выборки.





04

АНАЛИЗ И ВИЗУАЛИЗАЦИЯ РЕЙТИНГОВ ТЕЛЕШОУ

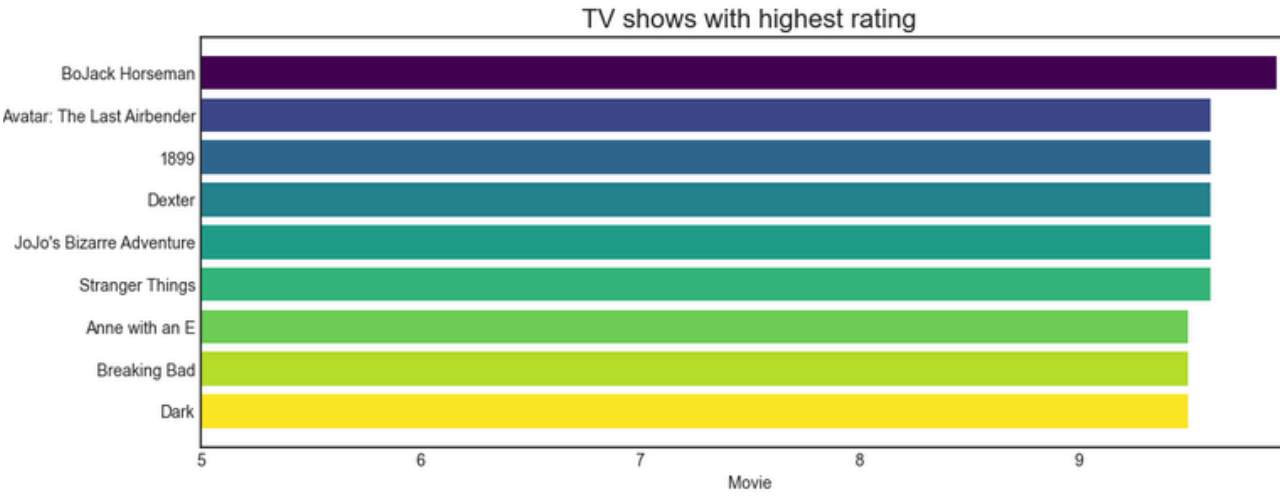
В этом проекте я провёла анализ данных о телешоу, сосредоточившись на их рейтингах.
Основные этапы работы включали в себя:

1. Обработка данных:
   - Заменила недостающие значения в столбце rating на 0, чтобы избежать искажений в анализе.
   - Преобразовала значения рейтинга в тип float для дальнейших вычислений.
2. Сортировка и выбор данных:
   - Отсортировала телешоу по рейтингу в порядке убывания и выбрала 10 шоу с наивысшими рейтингами.
3. Визуализация:
   - Создала горизонтальную столбчатую диаграмму с использованием библиотеки Matplotlib для наглядного представления результатов.
   - Настроила стиль графика и цветовую палитру, чтобы улучшить восприятие данных.

Результаты:
График демонстрирует 10 телешоу с наивысшими рейтингами, что позволяет быстро оценить их популярность.

Использованные технологии:

- Python
- Pandas
- Matplotlib
- NumPy



05

ПАРСИНГ НОВОСТЕЙ

Проект представляет собой скрипт на Python для парсинга новостей с сайта Euronews.

Основные функции:

- Запрос данных: Использует библиотеку requests для получения HTML-контента страницы новостей.
- Парсинг: С помощью BeautifulSoup извлекает заголовки и ссылки на статьи, находя их в тегах <h3>.
- Сохранение данных: Сохраняет извлечённые данные в CSV-файл с заголовками и ссылками на статьи.
- Обработка ошибок: Проверяет статус ответа, чтобы убедиться в доступности сайта.
- Вывод информации: Информирует пользователя о количестве найденных статей и успешном сохранении данных.

06

ДРУГИЕ

РАБОТА С HIVE

Проект демонстрирует основные операции с таблицами в Hive, включая создание, вставку и загрузку данных.

ПРАКТИКА С DOCKER

Проект включает два микросервиса, развернутых с помощью Docker: один для управления книгами и другой для рецензий. Оба сервиса созданы с использованием Flask.

КЛАССИФИКАЦИЯ ИЗОБРАЖЕНИЙ

Проект классификации изображений использует сверточную нейронную сеть (CNN) для распознавания объектов на изображениях с применением аугментации данных, такой как вращение и сдвиги. Модель обучается на наборе данных с 10 классами и оценивается по точности на тестовом наборе.