**Федеральное государственное образовательное бюджетное учреждение высшего профессионального образования**

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ**

**ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»**

**(Финансовый университет)**

**Факультет «Факультет информационных технологий и анализа больших данных»**

**Курсовая работа**

по дисциплине: «Технологии анализа данных и машинного обучения»

на тему:

**«Предварительный анализ данных и построение признаков в задачах оценки активов»**

Вид исследуемых данных:

Набор данных ….

Выполнила:

студентка группы ПМ19-4

Качуляк Маргарита Григорьевна

Научный руководитель:

к.т.н., доцент

Болтачев Эльдар Филаридович

Москва 2022г.

Оглавление

[1. ВВЕДЕНИЕ 3](#_Toc120986187)

[2. ТЕОРЕТИЧЕСКАЯ СПРАВКА 4](#_Toc120986188)

[2.1 Активы компании. Факторы, влияющие на увеличение экономической выгоды для организации 4](#_Toc120986189)

[2.2 Технические методы предобработки и обогащения данных для построения модели машинного обучения 4](#_Toc120986190)

[3. ПРАКТИЧЕСКАЯ ЧАСТЬ 5](#_Toc120986191)

[3.1 Анализ и предобработка выбранных данных 5](#_Toc120986192)

[3.2 Построение признаков 5](#_Toc120986193)

[3.3 Оценка качества обработки 5](#_Toc120986194)

[4. ЗАКЛЮЧЕНИЕ 6](#_Toc120986195)

[5. СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ 7](#_Toc120986196)

[6. ПРИЛОЖЕНИЕ 8](#_Toc120986197)

[6.1 Характеристики компьютера 8](#_Toc120986198)

[6.2 Коды программ 8](#_Toc120986199)

[6.3 Список файлов 8](#_Toc120986200)

[6.4 Время работы программ 8](#_Toc120986201)

## 1. ВВЕДЕНИЕ

Цель:

Задачи:

Актуальность:

## 2. ТЕОРЕТИЧЕСКАЯ СПРАВКА

### 2.1 Активы компании. Факторы, влияющие на увеличение экономической выгоды для организации

* Можно прогнозировать для одной компании на следующий период
* Или на основании показателей ряда компаний (из одной сферы/области) спрогнозировать показатель для другой компании

На что обращать внимание… финансовые показатели прибыли

### 2.2 Технические методы предобработки и обогащения данных для построения модели машинного обучения

* Изучение размера, типов переменных
* Заполнение пропусков
* Понижение размерности данных – агрегация (до 2 с помощью метода главных компонент)
* Поиск выбросов (ящиковая диаграмма) и удаление аномальных значений
* Приведение всех признаков к бинарной либо числовой шкале: кодирование нечисловых характеристик
* Уникальные имена. Обработка дубликатов
* Генерация данных: категоризация данных
* Сокращение нерелевантный признаков: удаление несущественных либо избыточных признаков

## 3. ПРАКТИЧЕСКАЯ ЧАСТЬ

### 3.1 Анализ и предобработка выбранных данных

### 3.2 Построение признаков

### 3.3 Оценка качества обработки

* Поиск смещений

## 4. ЗАКЛЮЧЕНИЕ

В результате данной курсовой работы

## 5. СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

* + - 1. Online community of data scientists and machine learning practitioners Kaggle, Spotify Playlists [Электронный ресурс]. URL: https://www.kaggle.com/datasets/andrewmvd/spotify-playlists (дата обращения: 3.05.2022)

## 6. ПРИЛОЖЕНИЕ

### 6.1 Характеристики компьютера

Тип процессора – Apple M1 (8 ядер, 4 производительных и 4 энергоэффективных)

Тактовая частота – 3.20 ГГц

Частота системной шины – 1066 МГц

Объем кэш-памяти второго уровня (L2) – 16 MB

### 6.2 Коды программ

Установка пакетов и библиотек:

Обработка датасета:

### 6.3 Список файлов

*Kachulyak\_ПМ19-4\_2022.ipynb* – файл JupiterNotebook, техническая реализация практической части работы.

*Курсовая\_2022\_Качуляк\_МГ\_ПМ19-4.pdf* – Печатная работа.

*.csv* – база данных платформы Kaggle.

### 6.4 Время работы программ

|  |  |
| --- | --- |
| Программа | Время работы, с. |
| Установка пакетов и библиотек |  |
| Обработка датасета |  |
|  |  |
|  |  |
| Таблица 1. |  |