**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №1**

**по дисциплине «Машинное обучение»**

Тема: Исследование набора данных

|  |  |  |
| --- | --- | --- |
| Студент гр. |  | Полуянов В. Н. |
| Преподаватель |  | Татчина Я.А. |

Санкт-Петербург

2024

**Цель работы.**

Выбор набора данных для дальнейшего исследования алгоритмов кластеризации и классификации. Получение практических навыков работы с инструментами анализа и визуализации данных.

**Постановка задачи.**

1. Создать Jupyter Notebook
2. Выбрать исследуемый набор данных из предложенных источников
3. Описать выбранный набор данных, а именно:
   1. Предметную область, источник данных, характер данных;
   2. Атрибуты, их тип, что они обозначают;
   3. Задачу анализу.
4. Для каждого атрибута:
   1. Определить среднее значение, СКО;
   2. Построить гистограмму распределения значений, определить наличие выбросов;
   3. Определить наличие пропущенных значений и их количество;
   4. Предложить вариант обработки пропущенных значений.
5. Определить:
   1. Высоко коррелированные атрибуты, характер корреляции;
   2. Не коррелированные атрибуты;
   3. Графики рассеивания.
6. Проанализировать полученные результаты.

**Выполнение работы.**

Для исследования был выбран синтетический набор данных оценки кредитных рисков, применимый для задач прогнозирования и классификации. Проанализированы признаки, которые содержит набор (см. приложение А).

Далее прочитаны данные, обработаны категориальные типы признаков EmploymentStatus, EducationLevel, MaritalStatus, HomeOwnershipStatus, LoanPurpose.

С помощью средств pandas.DataFrame.describe() и seaborn.histplot() вычислены основные параметры признаков и построены гистограммы распределения соответственно. Из гистограмм видно, что набор данных не содержит выбросов и пропусков.

Применив seaborn.heatmap() были выделены коррелированные атрибуты и выдвинуты основные предположения для предобработки, а именно:

1. Для возраста 20-30 лет разброс кредитного рейтинга от возраста будет больше, чем для возраста 60-80;
2. Корреляция возраста и опыта составляет более 98%, поэтому в будущем один из параметров можно будет удалить как избыточный;
3. Корреляция годового и месячного доходов так же более 98% и один из параметров можно удалить;
4. Фактическая кредитная ставка сильно зависит от базовой и эта зависимость требует дополнительного изучения.

Далее были проверены выдвинутые гипотезы. Так, зависимость опыта от возраста подтвердилась (рис. 1).

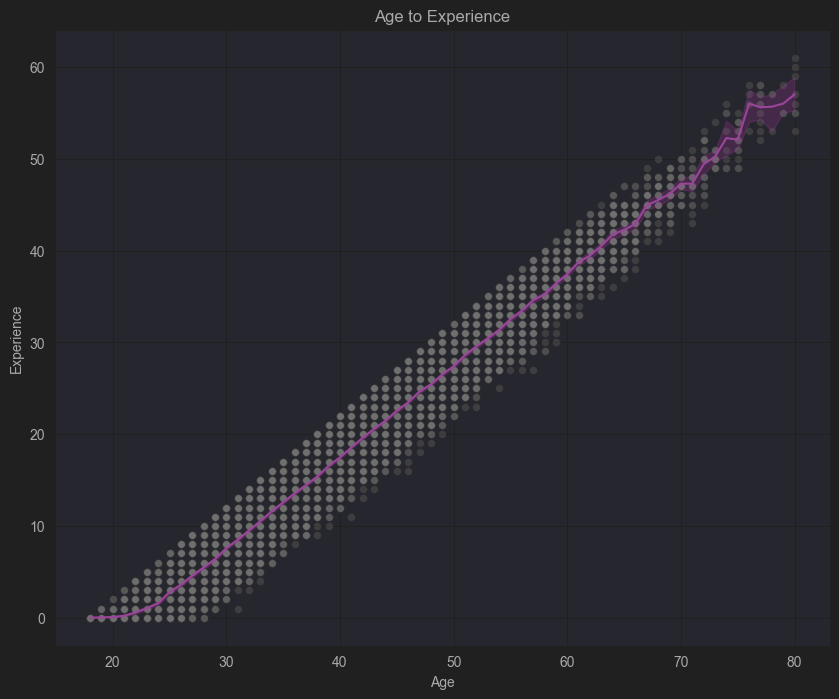


Рисунок 1. Зависимость Опыта от Возраста

Кредитный рейтинг действительно зависит от возраста (опыта), причём с ростом возраста (опыта) разброс уменьшается. Так же была снова подтверждена почти идеальная линейная взаимосвязь опыта и возраста (рис. 2).



Рисунок 2. Зависимости Рейтинга от возраста(опыта)

Корреляция месячного и годового доходов содержит выбросы, связанные с возможностью внеплановых заработков. Однако, для оценки кредитных рисков важнее стабильный месячный доход, нежели разовые крупные доходы, влияющие на годовые показатели.

Корреляция фактической и базовой ставки подтвердилась, в дальнейшем можно будет избавиться от одного из параметров (рис. 3).

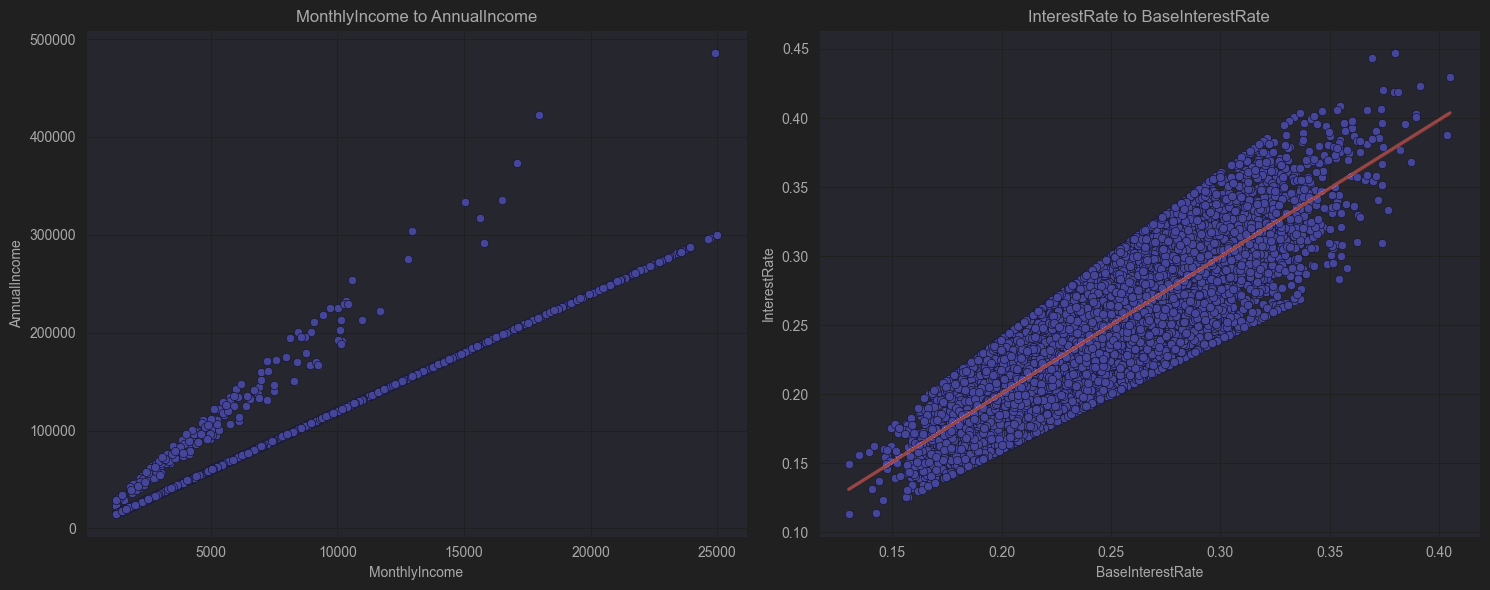


Рисунок 3. Зависимости месячного и годового доходов, фактической и ключевой ставки

**Выводы.**

В ходе работы были выбран набор данных для исследования. Получены практические навыки работы с анализом и визуализацией данных с помощью инструментов numpy, pandas, matplotlib, seaborn.

Приложение А

**Атрибуты набора данных**



