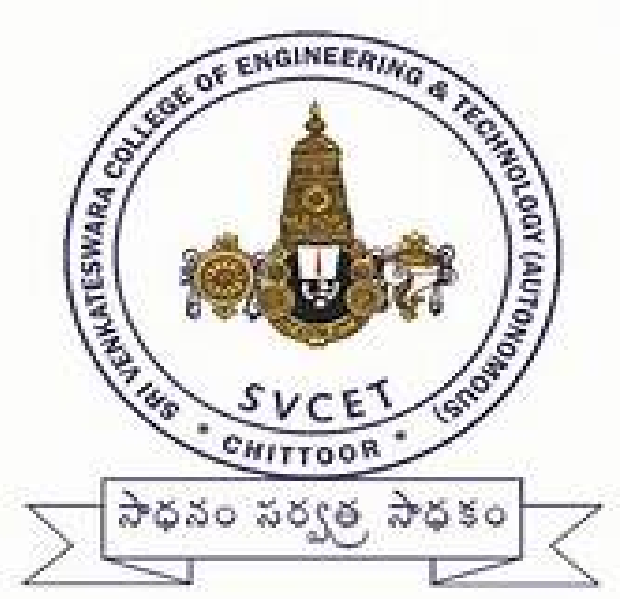


SRI VENKATESWARA COLLEGE OF ENGINEERING AND TECHNOLOGY (AUTONOMOUS)

DATA SCIENCE WITH PYTHON

Presented by:

G.ANUPA

21781A3240

Contents:

- What is Data Science?
- Python in Data Science
- Data Science Workflow
- Python libraries
- Data Visualization with Matplotlib and Seaborn
- Data Manipulation with Pandas
- Hands-On Project
- Challenges and Solutions
- Conclusion

What is Data Science?

- Data Science is the study of data to extract meaningful insights for business.
- Data Science is used in asking problems, modelling algorithms, building statistical models.
- Data science is an interconnected field that involves the use of statistical and computational methods to extract insightful information and knowledge from data.
- Python is popular and versatile programming language, now has become a popular choice among data scientists for its ease of use, extensive libraries,and flexibility.

Data science workflow

- Problem Definition: clearly define the problem you're trying to solve and establish goals.
- Data collection: Gather relevant data from various sources, ensuring it aligns with various sources, ensuring it aligns with your problem and goals.
- Data Cleaning: Preprocess and clean the data handling missing values, outliners, and ensuring consistency.
- Exploratory Data Analysis(EDA): Explore the data to gain insights, visualize patterns, and understand its characterstics.
- Model Selection: Choose appropriate algorithims/models based on the nature of your problem and data.
- Model Training: Train the selected models using your prepared data.

Python libraries

- Python has libraries with large collections of mathematical functions and analytical tools.
- In this course, we will use the following
- libraries:
- Pandas- This library is used for structured data operations, like import CSV files, create data frames, and data preparation.
- Numpy- This is a mathematical library has a powerful N-dimensional array object, linear algebra, fourier transform, etc. Visualization of data.
- Matplotlib- This library is used for visualization of data.
- SciPy- This library has linear algebra modules.

Data visualization with Matplotlib and seaborn

- Matplotlib and Seaborn are powerful Python libraries that offer a wide range of tools for creating appealing and informative visualizations.
- Matplotlib:
- Matplotlib is a versatile library for creating static, animated, and interactive visualizations. let's start with a simple line plot:
- Import matplotlib.pyplot as plt
- import matplotlib.pyplot as plt
- # Sample data
- x = [1, 2, 3, 4, 5]
- y = [2, 4, 6, 8, 10]
- # Plotting the data
- plt.plot(x, y, label='Linear Function')
- plt.xlabel('X-axis')
- plt.ylabel('Y-axis')
- plt.title('Simple Line Plot')
- plt.legend()
- plt.show()

- Seaborn:
- Seaborn is built on top of Matplotlib and provides a high-level interface for statistics data visualization. let’s create a histogram using Seaborn:
- import seaborn as sns
- # Sample data
- data = [1, 2, 2, 3, 3, 3, 4, 4, 5]
- # Creating a histogramsns.histplot(data, bins=5, kde=True, color='skyblue’)
- plt.xlabel('Values’)
- plt.ylabel('Frequency’)
- plt.title('Histogram with Seaborn’)
- plt.show()

Data Manupulation with Pandas

- DataFrame in Pandas
- A DataFrame is two-dimensional table in pandas.Each column can have different data types like int, float,or string.Each column is of class series in pandas.
- Creating a DataFrame in Pandas
- # import the library as pd
- import pandas as pd
- df = pd.DataFrame(
- {
- 'Name': [“Vandana”,”Hyma”]
- Age': [20, 20],
- 'Country': ['India’,’India’]
- }
-)
- print(df)
- # output
- Name Age Country

0 Vandana 20 India

1 Hyma 20 India

Project:

- **Problem statement:** Create a classification model to predict whether CREDIT RISK is good Or bad.

- **Context:**

- Financial institution, is interest is to know the potential financial whereabouts of the customers in order to determine whether the credit risk associated with them is good or bad.

- The data set could be used to predict if the customer could be given credit. Many features require data cleaning.
- After that, we will use two data sets that emulate real credit applications on business values.

What exactly is credit risk?

- Credit Risk is when lender lends money to a borrower but may not be paid back.

Loans are extended to borrowers based on the business or the individual's ability to service future payment obligations (of principal and interest).

Calculated risk is the difference between lending someone money and a Government bond.

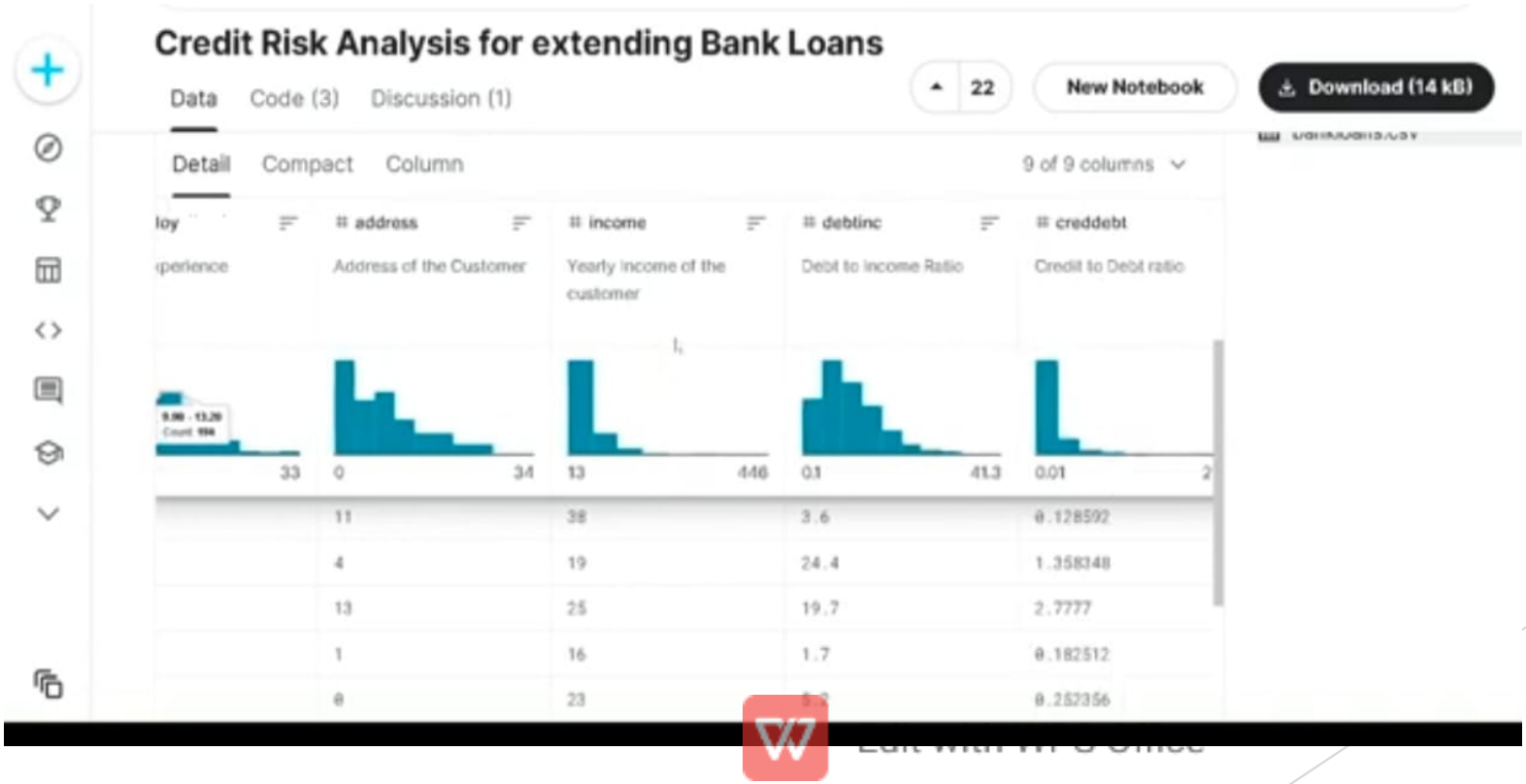
Dataset:

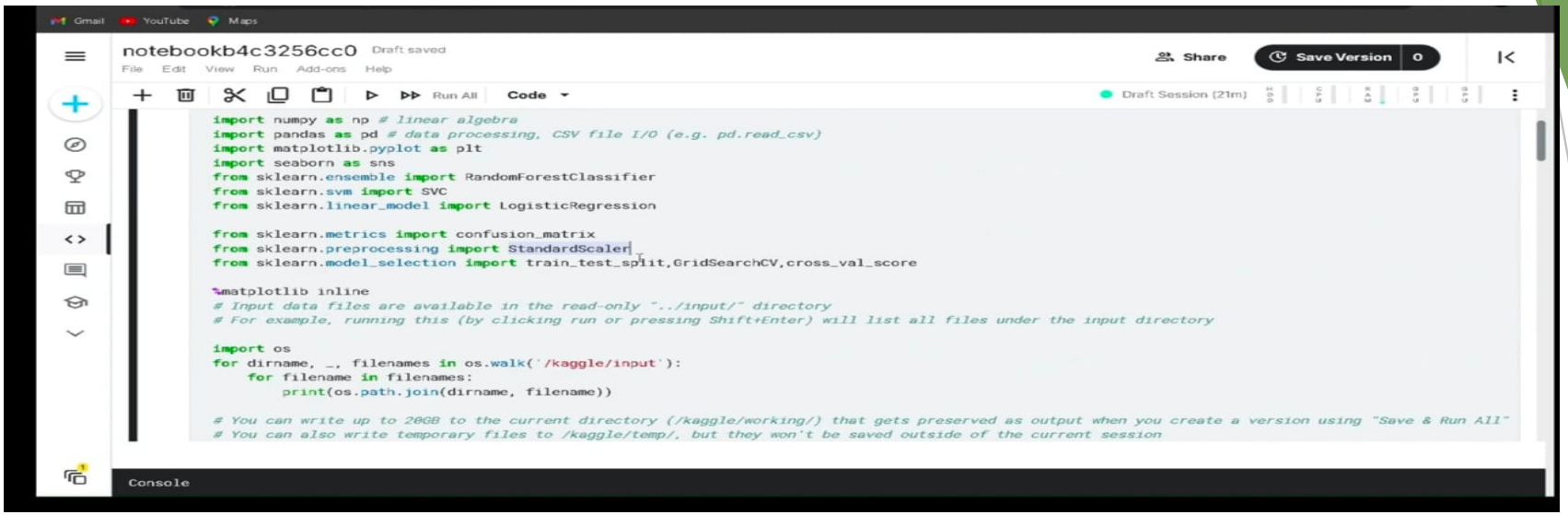
- For modelling probability of default we generally have two primary types of data available:
- Application data: Which is data that is directly tied to the loan application like loan grade.
- Behavioral data: Which describes the recipient of the loan,such as employment length.

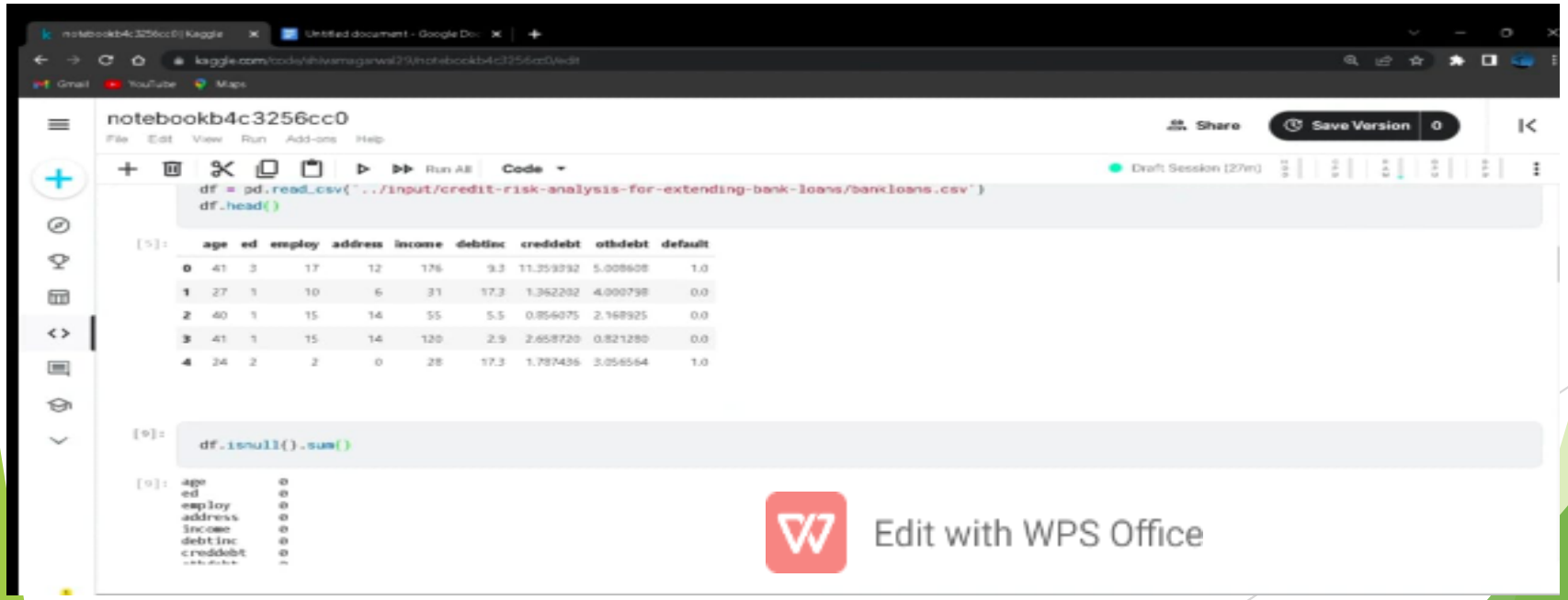
- The data will use for our predictions of probability of default includes a mix.
- This important because application data alone is not as good as application and behavioral data together.

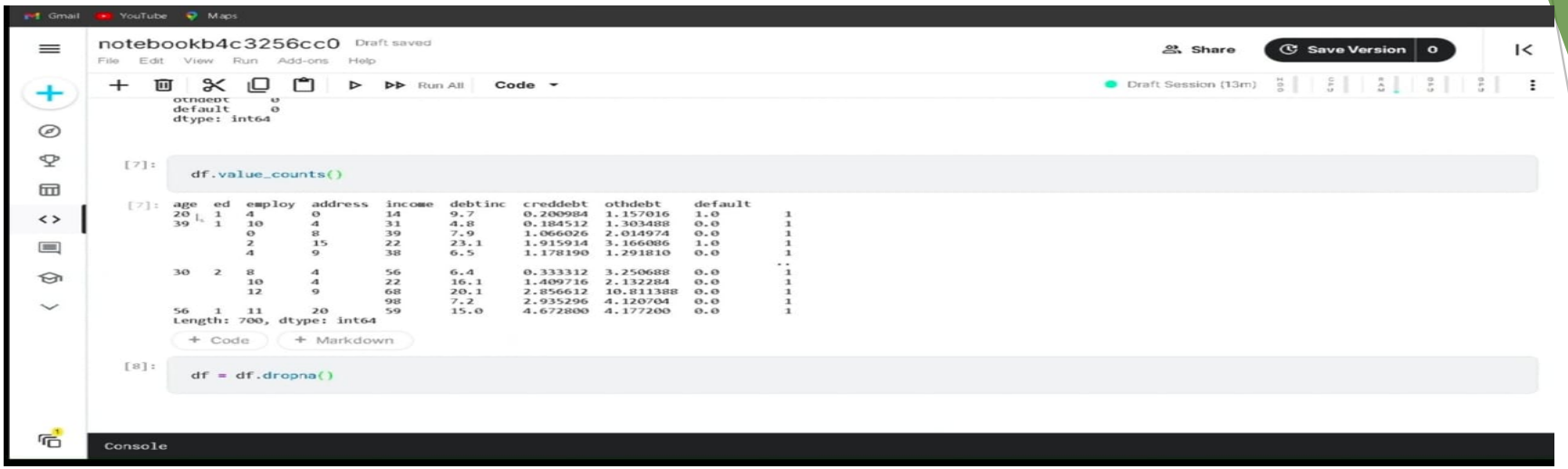
Credit Risk Analysis for extending bank loans:

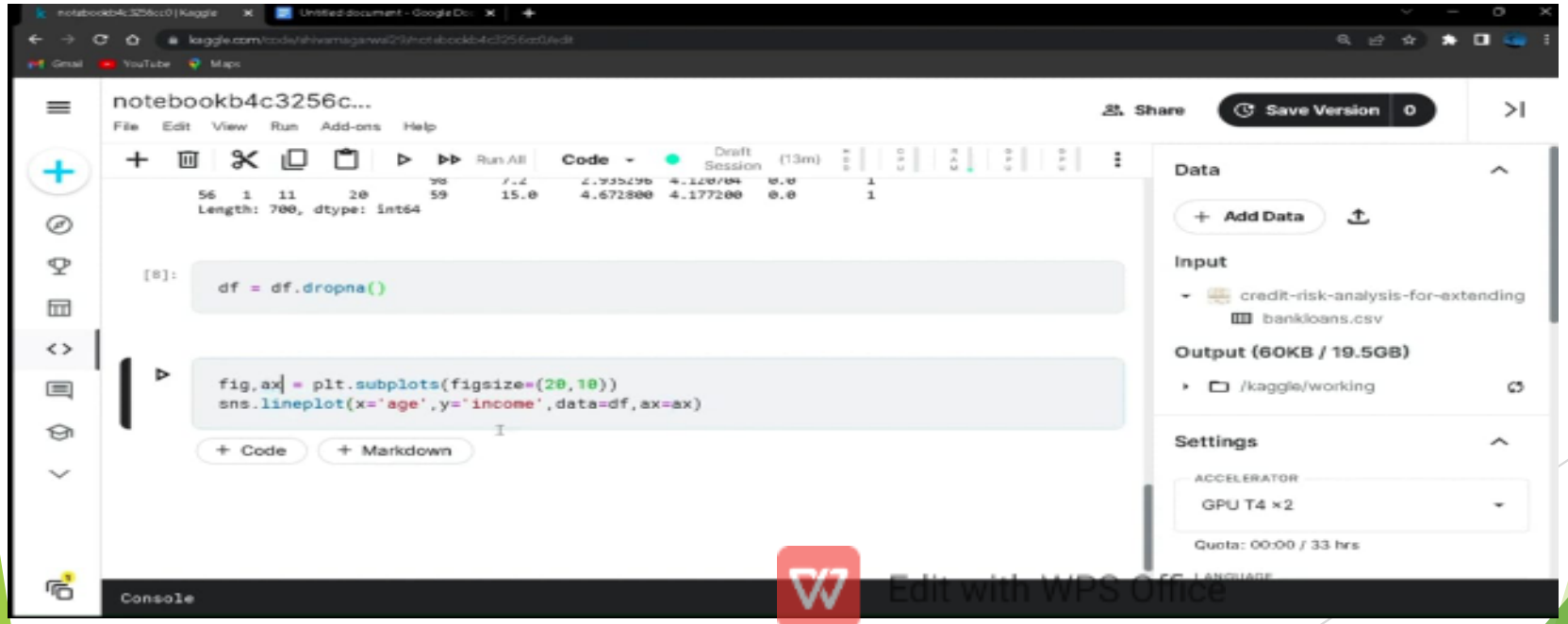
- Credit risk is perhaps one of the most ‘classic’ applications for predictive modelling, to predict whether or not credit extended to an applicant will likely result in profit or losses for lending institution.

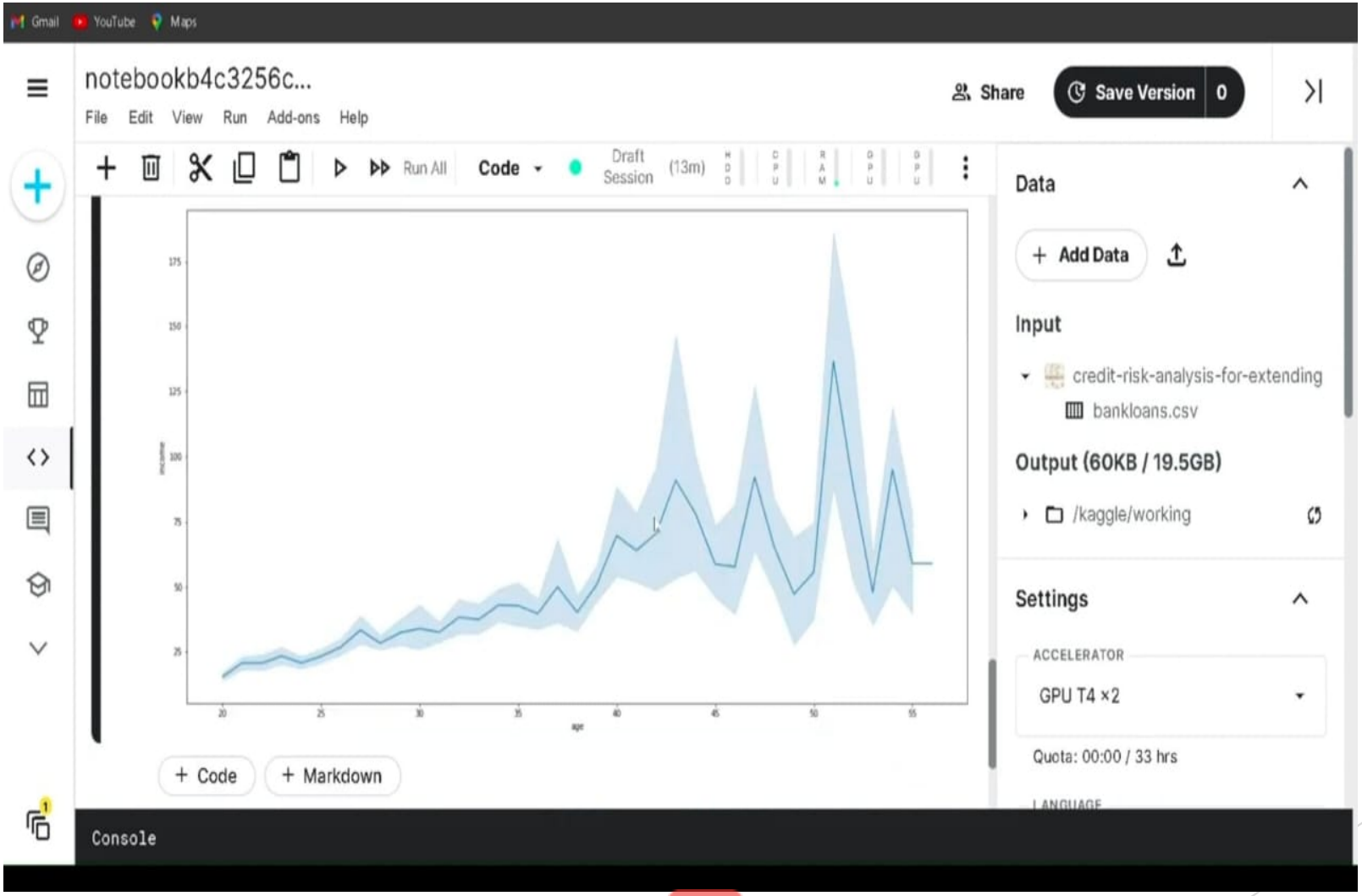



notebookb4c3256cc0
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (21m)
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split,GridSearchCV,cross_val_score
%matplotlib inline
# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
for filename in filenames:
print(os.path.join(dirname, filename))
# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
Console

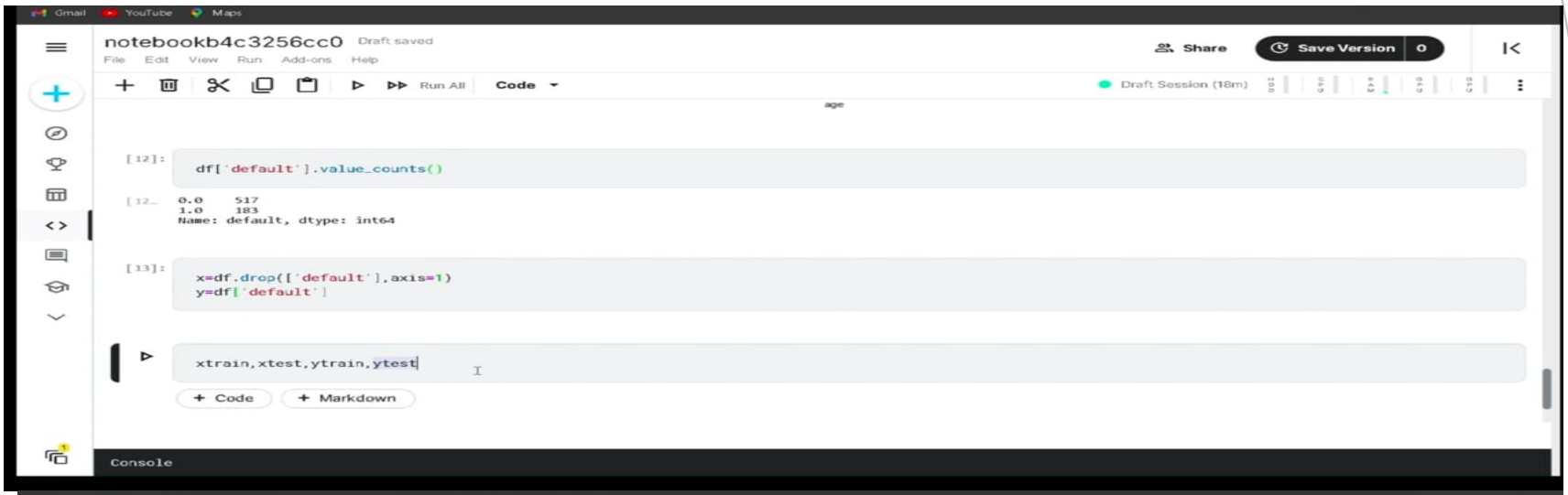

notebookb4c3256cc0
Share
Save Version 0
Run All
Code
Draft Session (27m)
df = pd.read_csv('../input/credit-risk-analysis-for-extending-bank-loans/bankloans.csv')
df.head()
[5]:
age ed employ address income debtinc creddebt othdebt default
0 41 3 17 12 176 9.3 11.359392 5.008608 1.0
1 27 1 10 6 31 17.3 1.362202 4.000798 0.0
2 40 1 15 14 55 5.5 0.856075 2.168925 0.0
3 41 1 15 14 120 2.9 2.658720 0.821280 0.0
4 24 2 2 0 28 17.3 1.787436 3.056564 1.0
[9]:
df.isnull().sum()
[9]:
age 0
ed 0
employ 0
address 0
income 0
debtinc 0
creddebt 0
Edit with WPS Office

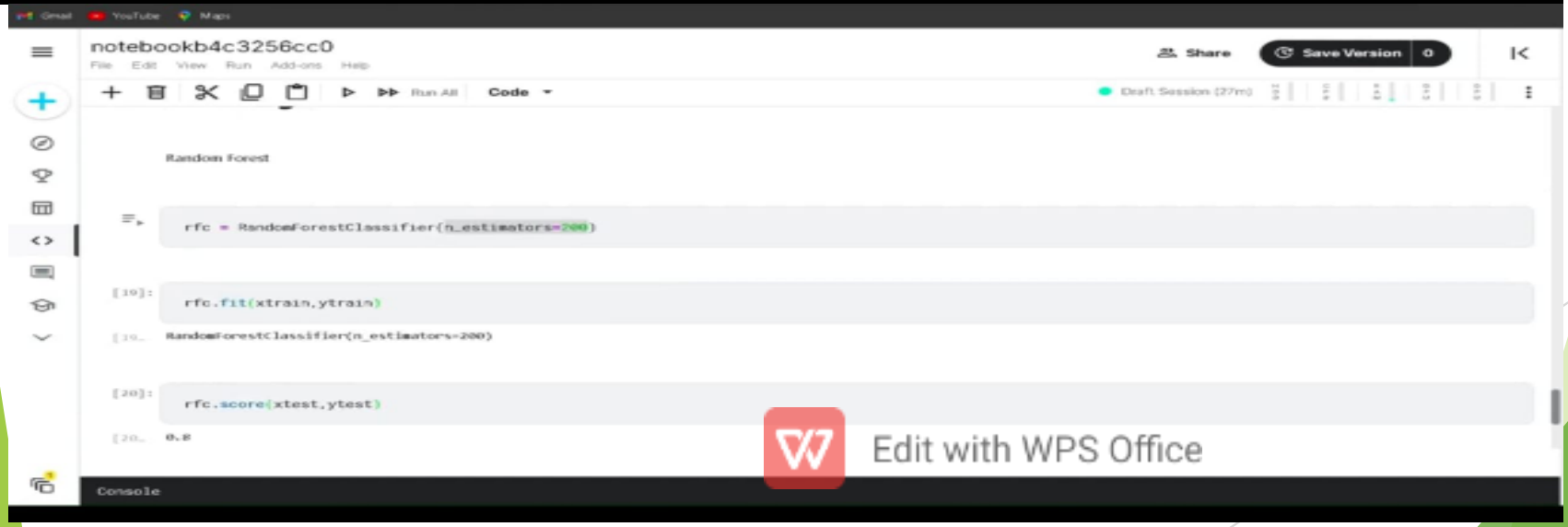


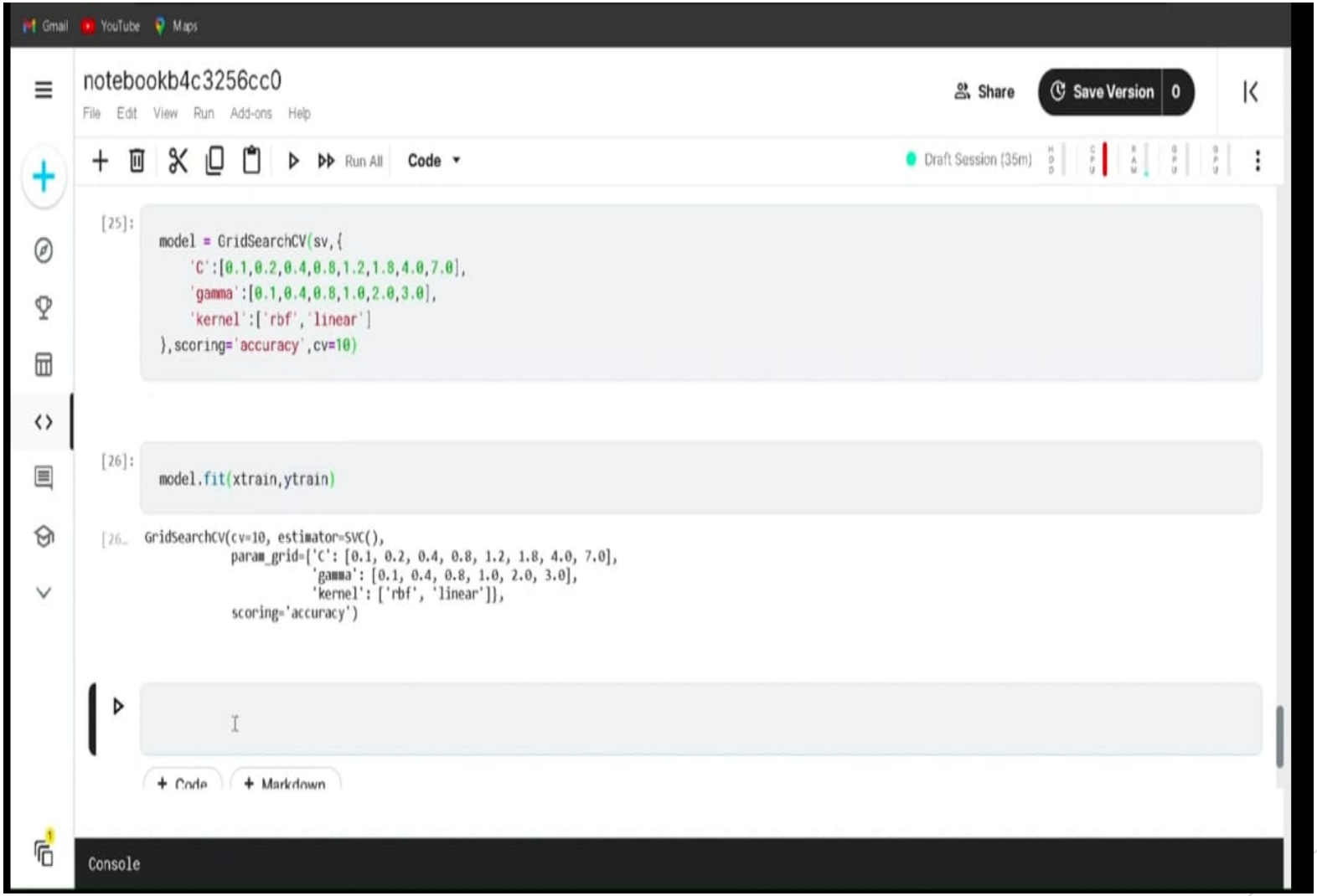


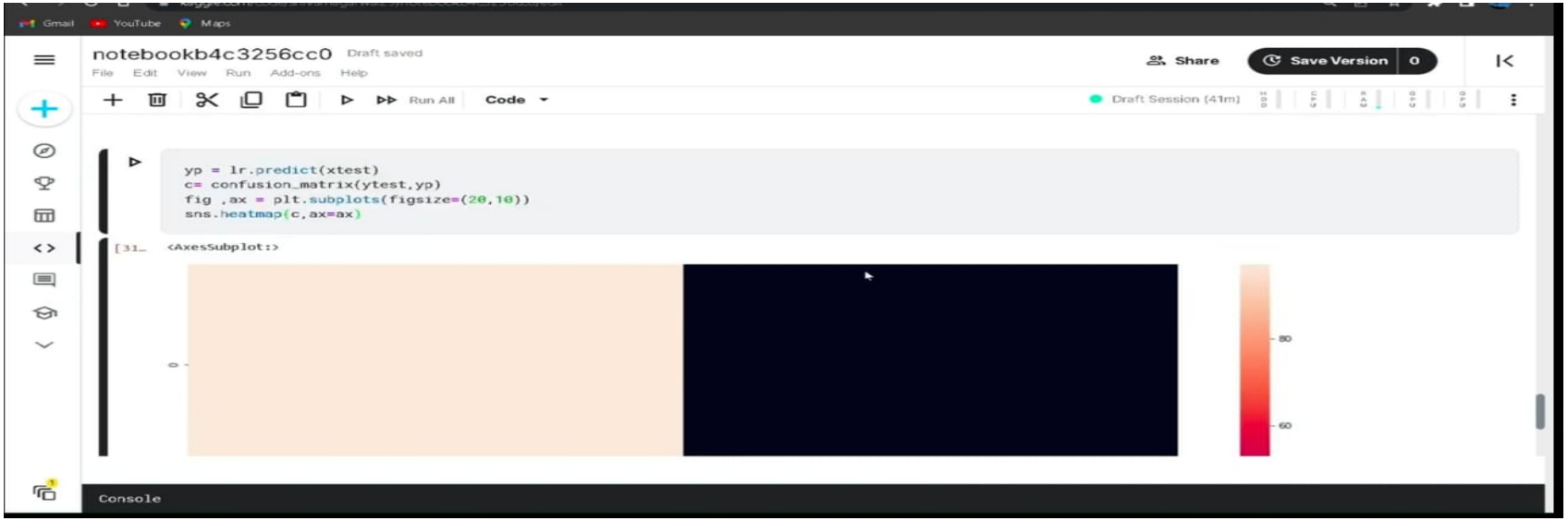


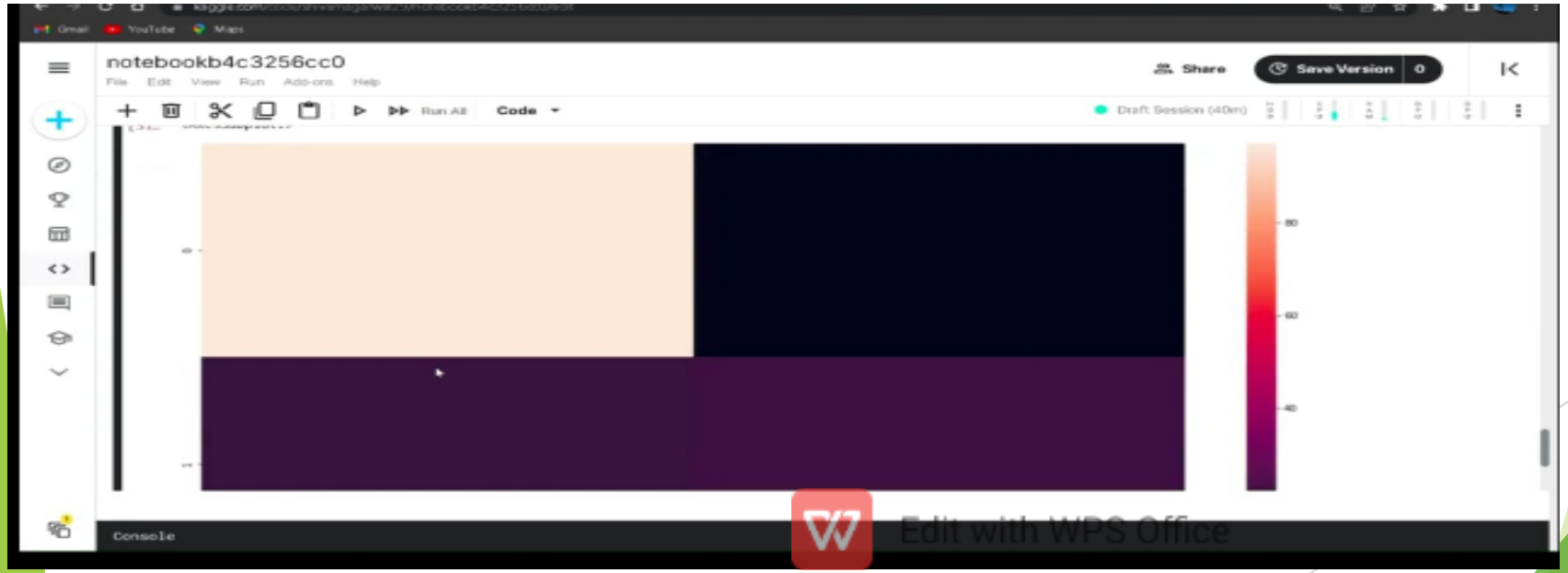
Creating Model:








notebookb4c3256cc0
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (41m)
yp = lr.predict(xtest)
c= confusion_matrix(ytest,yp)
fig ,ax = plt.subplots(figsize=(20,10))
sns.heatmap(c,ax=ax)
[31]
<AxesSubplot:>
80
60
Console


notebookb4c3256cc0
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (40m)
80
60
40
Console
Edit with WPS Office

Conclusion:

- Predicting credit risk involves a comprehensive analysis of various factors, including financial history ,payment behaviour and economic indicators.

- Payment patterns and consistency serve as crucial indicators for credit worthiness.

Economic conditions plays a vital role in influencing credit risk.

However, it’s crucial to acknowledge the dynamic nature of financial markets and inherent uncertainities in predicting credit outcomes.