Paper ###-2021 (if SAS author, use SAS###-2021 or SAS####-2021)

**Prediction of customers churners – Machine Learning**

Guide: Prof. Ramanath Krishnamoorthy, Welingkar Institute of Management;

Author-1: Muskan Agarwal, Welingkar Institute of Management;

Author-2: Prasanna Kulkarni, Welingkar Institute of Management

# Abstract

# This study and research show that many of the banks struggle to retain their profitable customers (credit card, general commercial purpose, etc.), the whole business model depends upon their key stakeholders i.e., customers. Banks need to analyze and start making necessary changes to retain customers, make them loyal, make them committed to the organization and they should feel that services are trustworthy, which leads to increased revenue and goodwill to banks. Data is the main thing every organization is recording in this new age of managing firms, every decision higher management personnel taken is through data and what stories they tell about the organization, same goes for banks, collection of data on customers and target customers will help firms to make better decisions regarding their current offerings and future offerings they want to make which will create win-win situations for both parties. Data collected can be subjected to various data exploration and analysis techniques to predict whether certain customers will leave the organization or enjoy the service (or what else extra is expected from that customer), hence firms (here bank) can never make their clients dissatisfied, making them a committed customer. The main purpose of this study is to understand the variables which affect the decisions of customers when they leave or stay. Using these variables and data of a considerable number of customers (existing and non-existing), a machine learning model can be built to predict if any customer who is about to get bank services or already availing bank services will leave the bank’s offerings (credits, credit cards, bank accounts, etc.) or not. The project went into the research of available data for this problem statement and applied analysis methodology to arrive at the required solution (model), learnings are: exploration of data, data cleaning techniques, feature engineering, building models, and visualization.

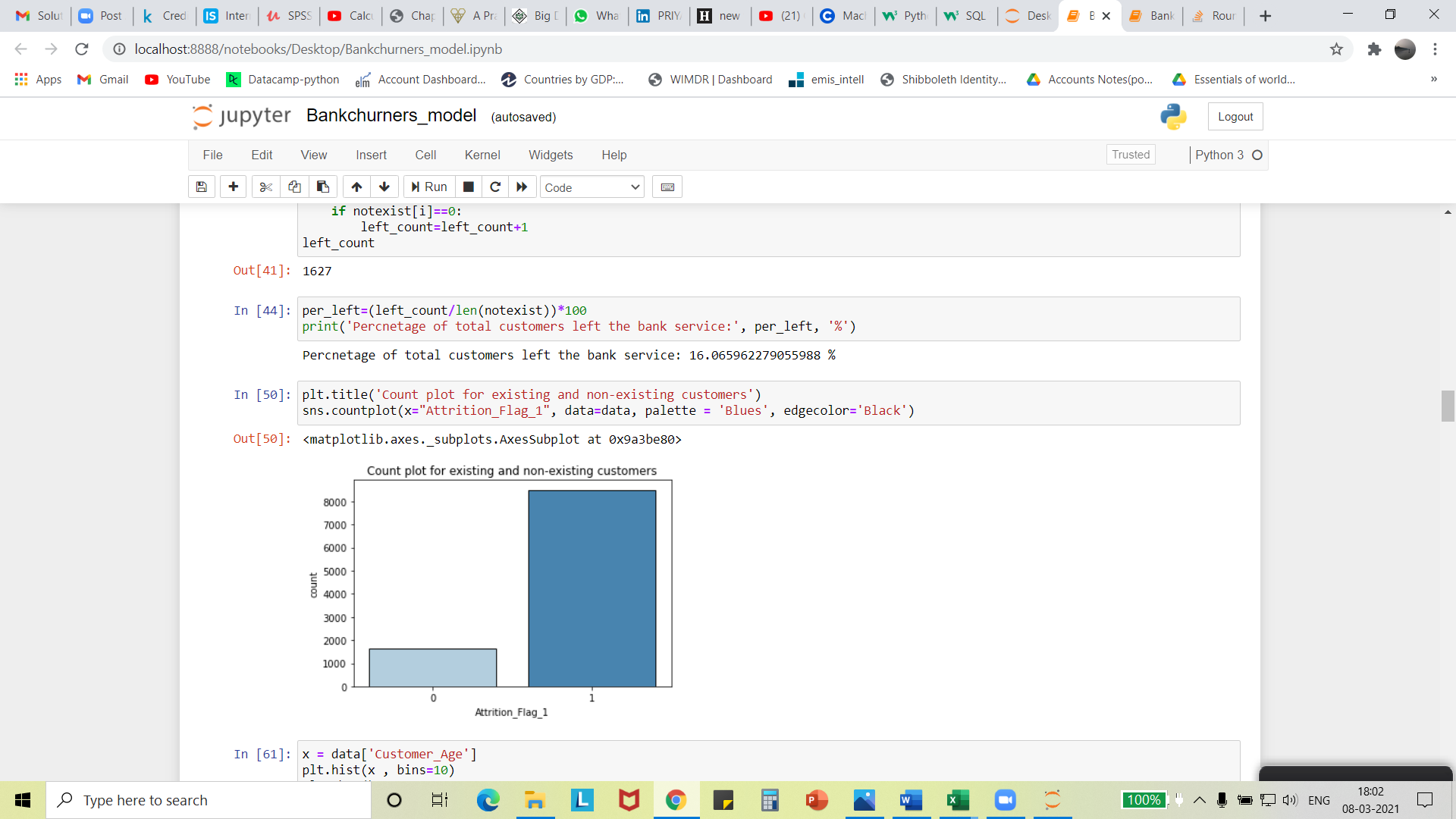
# Introduction

Every organization needs revenue to survive in markets and every business’s objective is to create the offerings and exchange them with desired customers, clients and improve the society as a whole, getting these customers and making them commit to the organization is not an easy task in this ever-changing market trends and customers expect. Problem consideration in this project relates to the identification of target customers to particular offerings, retaining the existing customers, and exploring the reasons for which customers have left. The objective is to provide a robust machine learning model using a considerable amount of data to predict (if the customer stays or leaves) and suggest the target group (TG) (through proper study of necessary attributes about TG). Here we have taken a bank’s credit card story (retrieved customers data which includes information of both existing and churned customers) to build the model and conclude, but this approach can be applied to every aspect of every industry (healthcare, educational, corporate firms, banking, insurance, manufacturing, real estate, and financial services), the condition is that right type and amount of variables has to be considered for each problem type and data retrieval has to be done for analysis.

# Data

This is a main. Data was researched and taken up from the Kaggle data source site (Kaggle is the world's largest data science community with powerful tools and resources to help you achieve your data science goals).

Below is the link pasted which directly leads to the data set description and data.

<https://www.kaggle.com/sakshigoyal7/credit-card-customers>

**Data set info:**

Number of columns/ attributes/ features/ variables: 23

Number of rows/records/data: 10,127

**Attributes and their description:**

1. Client number: Unique number assigned to every credit card customer (existing and non-existing, 2)
2. Attrition flag: Variable that records whether the customer is currently availing credit card services or quit the bank. (categorical variable, 2)
3. Customer age: Age of the customers. (continuous variable)
4. Gender: Gender of the customers (male or female) (categorical variable, 2)
5. Dependent count: numbers of dependent that he/she carries in that household or family. (continuous variable)
6. Education level: maximum education a customer has done till that day of being a customer. (categorical variable, 7)
7. Marital status: Customers marital status of existing and non-existing customers (categorical variable, 4)
8. Income category: To which income category slab the customers belong. (categorical variable, 6).
9. Card Category: Bank has four categories/types of cards (blue, gold, platinum, and silver), (ordinal categorical variable), to which card category the customer has availed himself/herself with.
10. Months on months: Relationship the customer holds with bank availing services. (continuous variable)
11. Total relationship counts Numbers of products a customer avails from the bank. (continuous variable)
12. Months inactive in a year (12 months): Numbers of months a customer was inactive in 12 months period. (continuous variable)
13. Contacts count in a year (12 months): Frequency with which a customer has contacted bank personnel in a year. (continuous variable)
14. Credit limit: Limit of credit that is eligible for a particular card. (continuous variable)
15. Total revolving balance: revolving balance held by a customer during his tenure as a customer. (continuous variable)
16. Average open to buy: Open to Buy Credit Line (Average of last 12 months)
17. Total amount changes from q4 to q1: Difference or change in amount transactions done by customers. (continuous variable)
18. Total transfer amount: Whole or total amount transacted by the customer in a year. (continuous variable)
19. Total transfer count: Number of transactions done by a customer. (continuous variable)
20. Total count changes from Q4 over Q1: Change in transaction count. (continuous variable)
21. Average utilization ratio: Average card utilization ratio (blue, gold, platinum, and silver).
22. Naive\_bayes\_classifier\_attrition\_flag\_card\_category\_contacts\_count\_12\_mon\_dependent\_count\_education\_level\_months\_inactive\_12\_mon\_1.
23. Naive\_bayes\_classifier\_attrition\_flag\_card\_category\_contacts\_count\_12\_mon\_dependent\_count\_education\_level\_months\_inactive\_12\_mon\_2.

Data set is a combination of existing and non-existing customers in the ratio of 1627:8500 majority being existing customers.

# Problem statement

# “Customers” is an answer to many questions that are asked within the organization, main source revenue, word of mouth about offerings, product development through empathy and feedback, and so on. Acquiring these customers and retaining them by providing value-added products and services will be the main goal. Here, bank management wants to know why their credit card customers are leaving the services. They would like to know the attributes which relate to this cause, predict future attritions based on these attributes and evaluate future customers if they are going to commit to bank services or not. The objective is to develop a machine learning model to predict pre-purchase behavior and post-purchase behavior of customers about attrition and generalize this model type to predict many such problems’ result in different sectors.

# Data cleaning

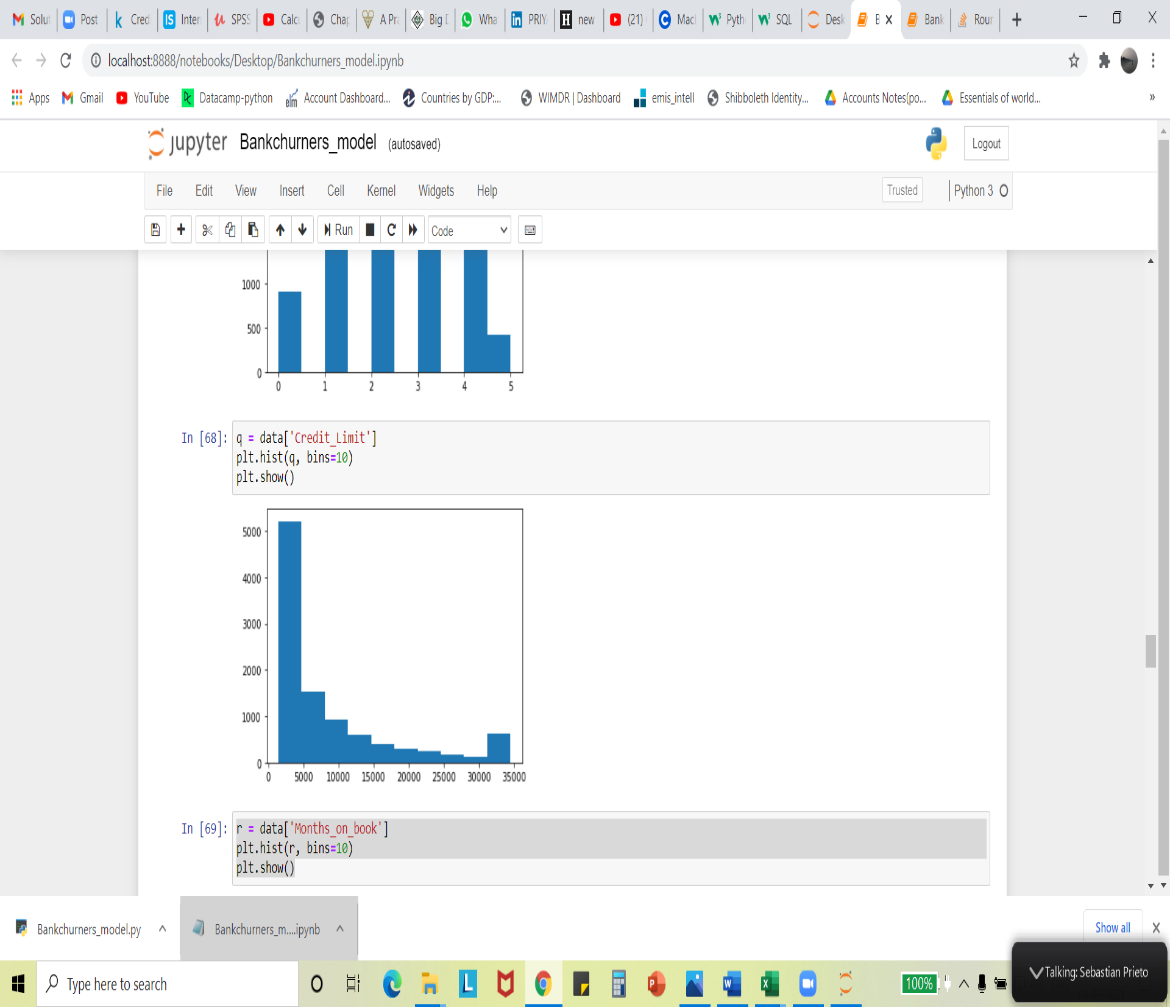
The process helps to provide quality data for analysis.

Methods or steps undertaken for data cleaning are:

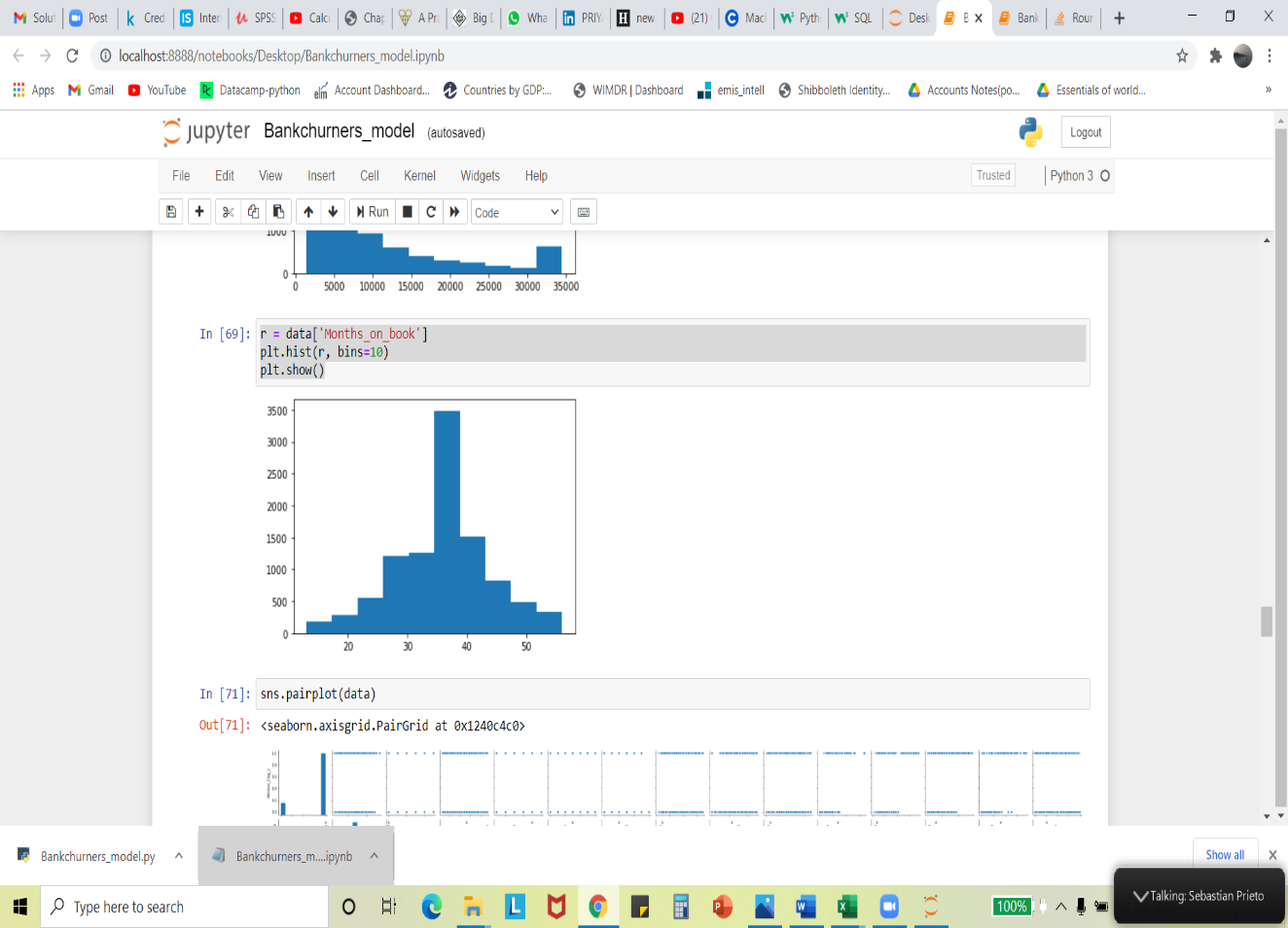
1. Eliminate the variables which are reconsidered as redundant or non-value adding attribute and manual encoding of attrition flag in excel.
2. Check for the duplication of records among 10127 entries.
3. Dealing with Missing values along with the whole dataset.
4. Inappropriate values recognition in an attribute. (example. Age and months on the book cannot be value zero)
5. Attribute consistency has to be checked for any outlier elimination. (box plots or histogram as univariate analysis for continuous variables).
6. Range and central tendency checking and evaluation by descriptive statistics.

**ReSults from the exploratory analysis (data exploration with visualisation)**

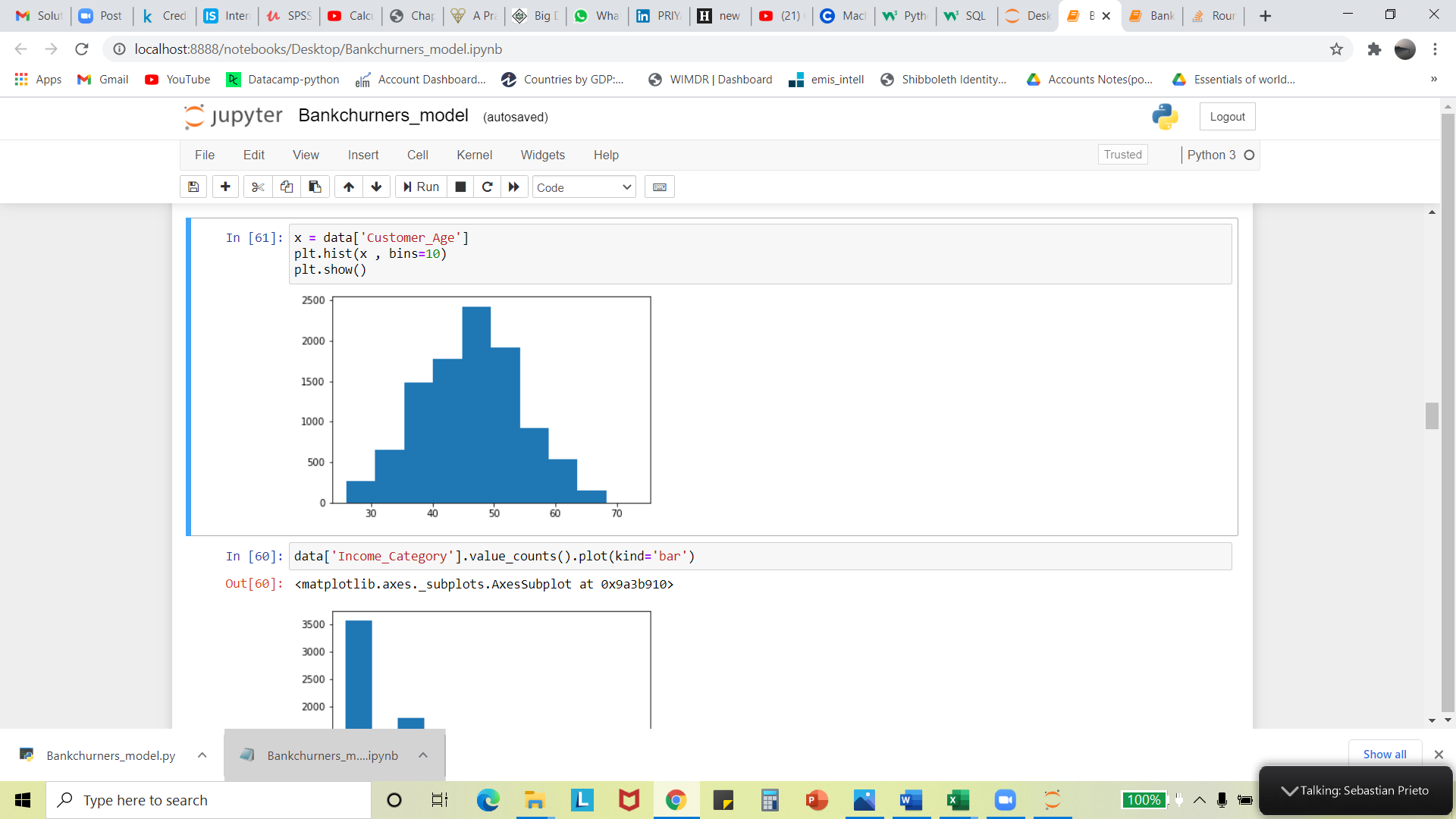
1. Columns: client number and naïve Bayes were deleted, Existing-1, Non-Existing-0.
2. Using Excel’s remove duplicate function, all the records of the dataset were checked and found zero duplications.
3. Using python Jupyter notebook editor with pandas library missing values were checked, no missing values found, using following code.
   1. Code: null\_column = pd.DataFrame({'column': data\_df.isna().sum().index, 'no\_null\_values': data\_df.isna().sum().values,'percentage': data\_df.isna().sum().values.shape[0]})
4. No inappropriate values are found in the attributes like age and months on the book.
5. Consistency of attributes like age, months on books, and credit limit are checked through histogram representation using python’s matplotlib library.
6. Descriptive analysis is being done for attributes to check the consistency in ranges (0%, 25%, 50%, 75%, 100%).



Credit limit



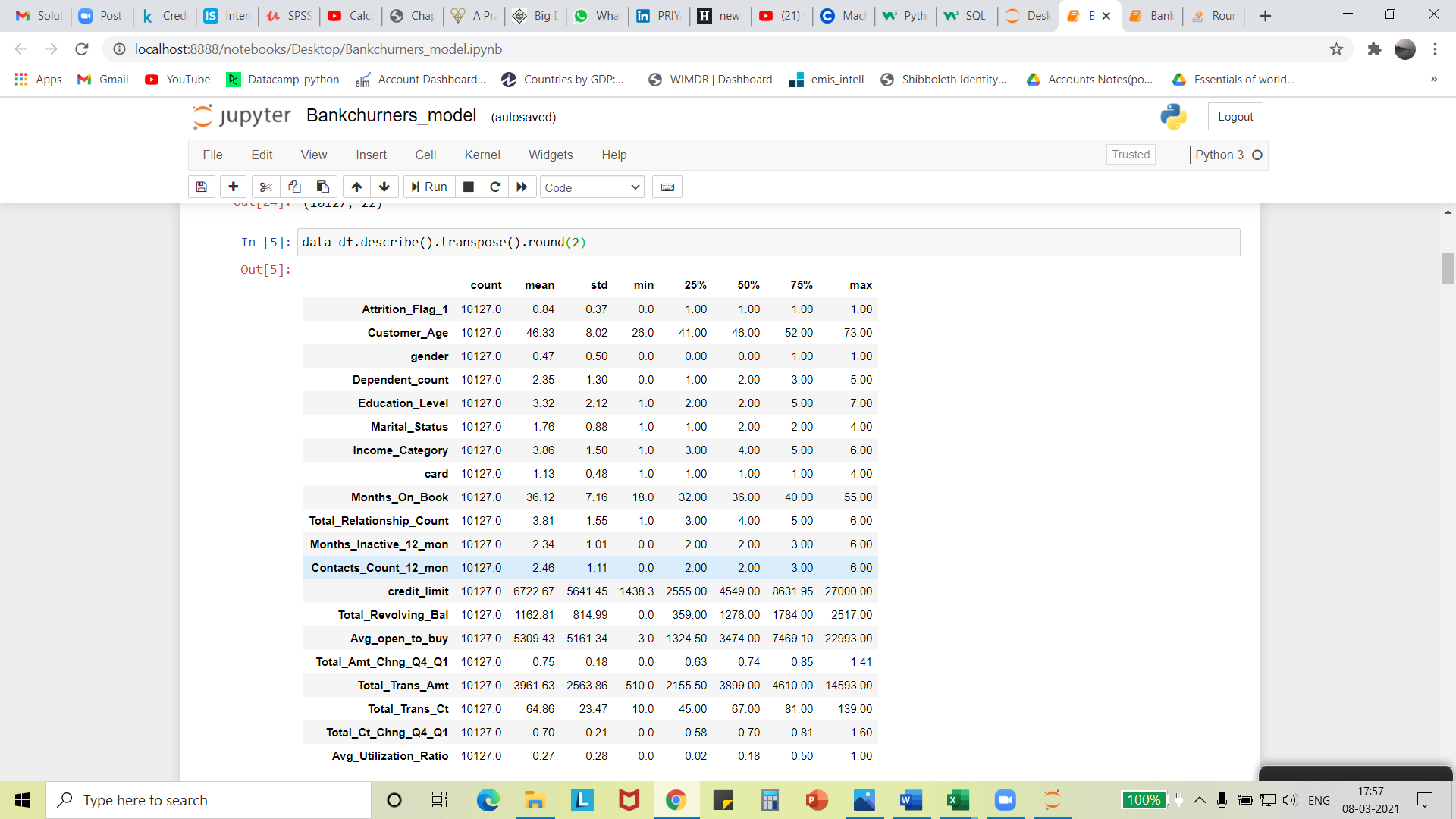
Months on book

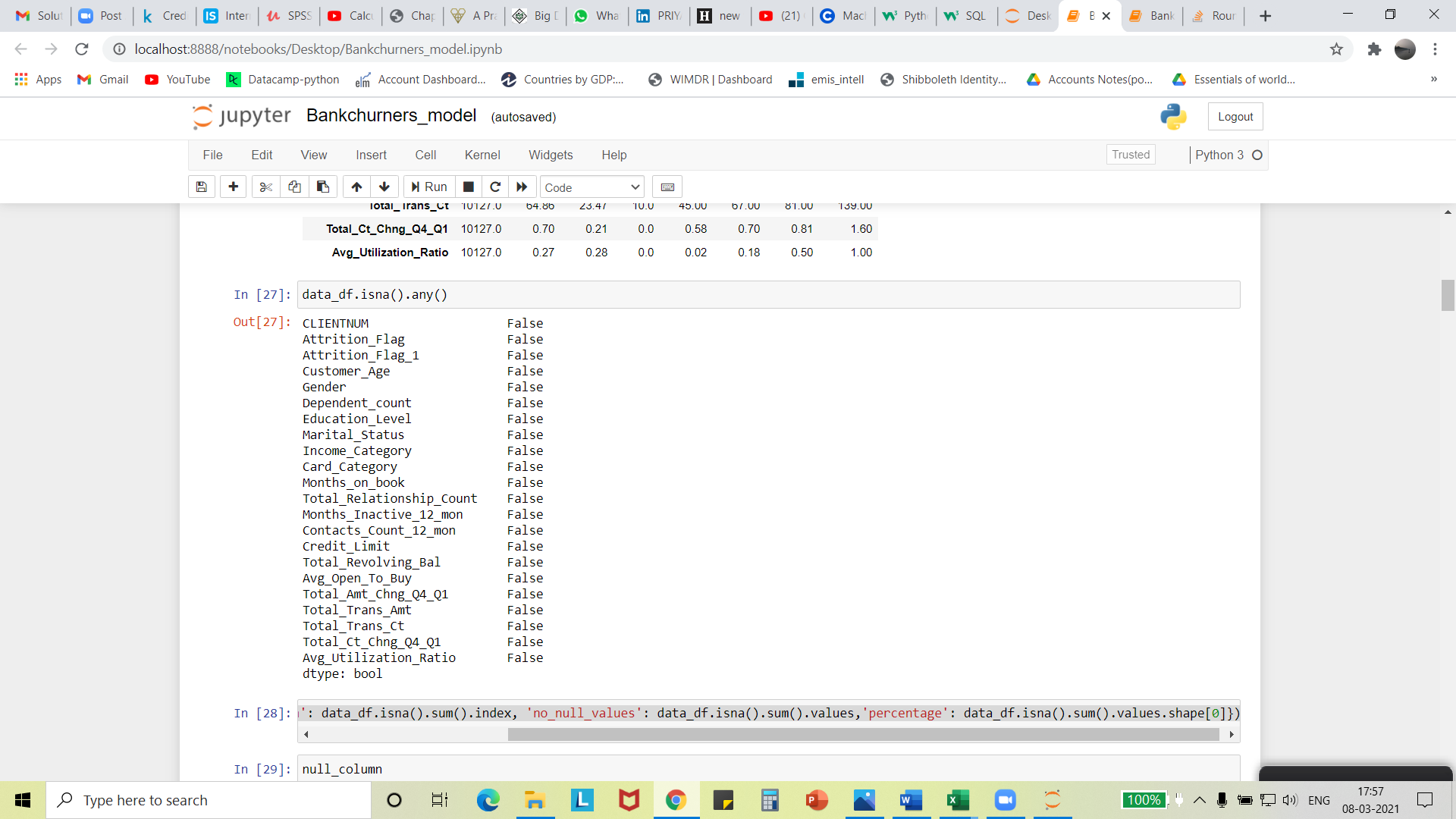


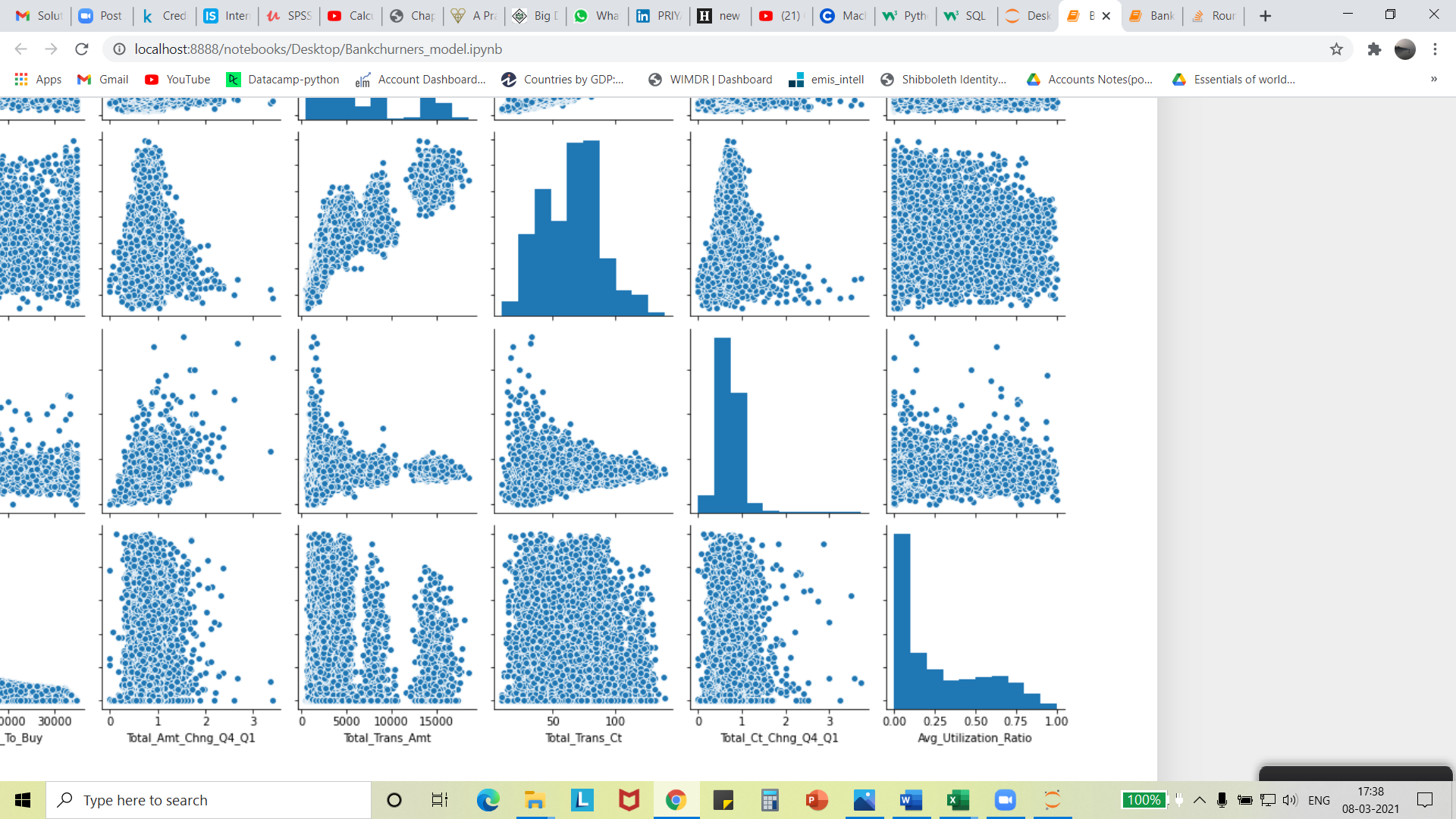
AGE



Total amt change Q4-Q1



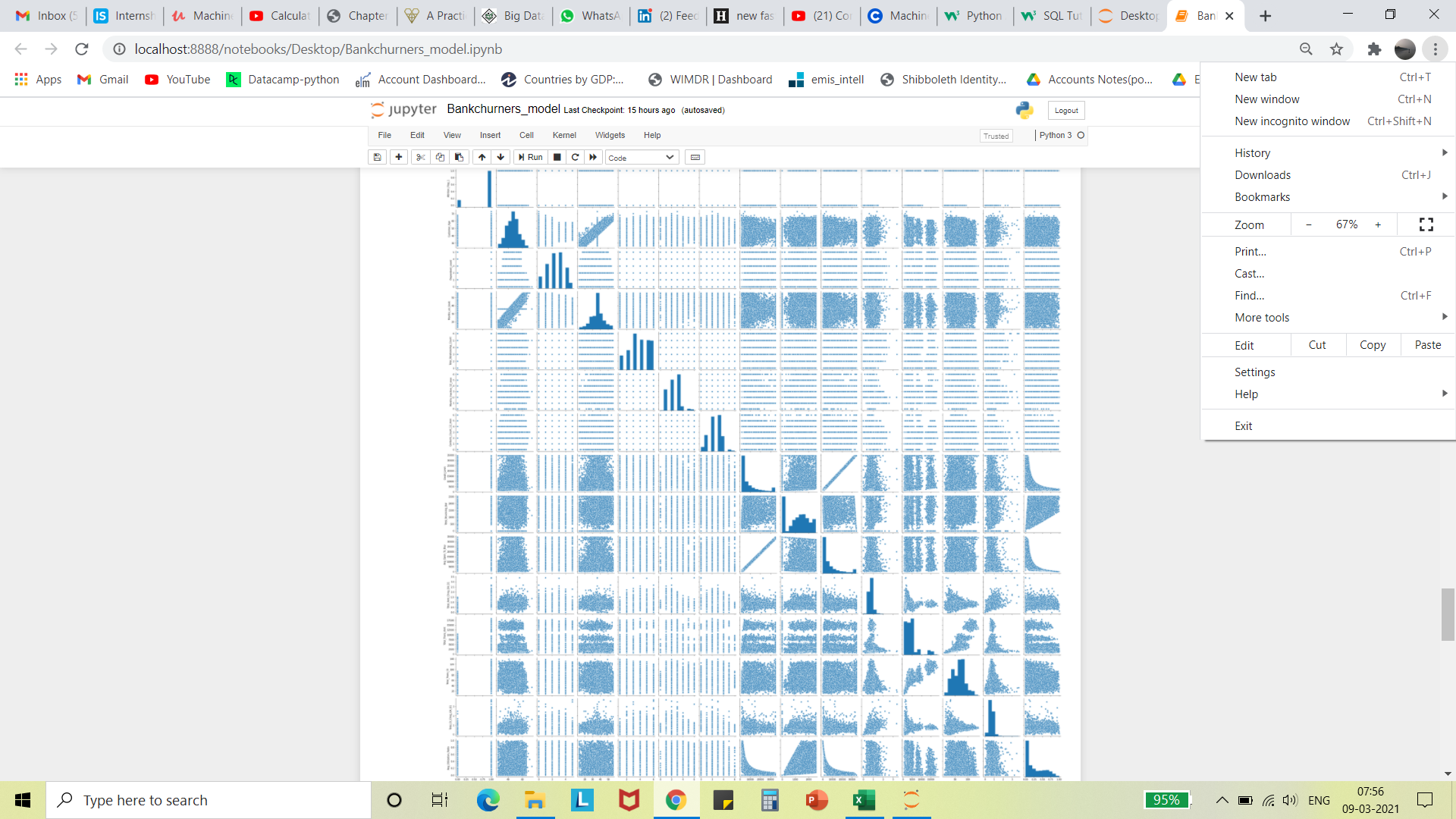
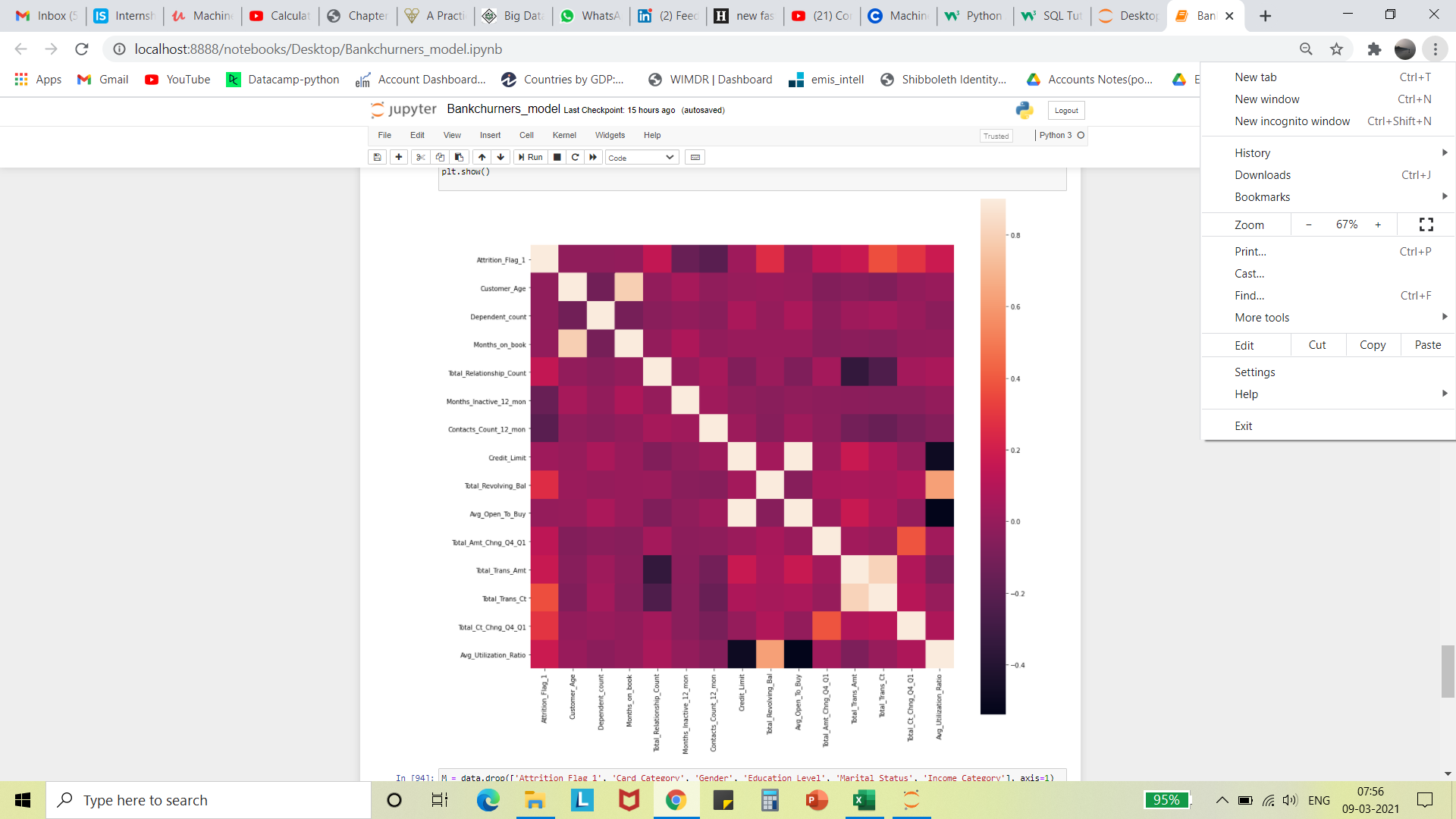




Utilization ratio



Avg open to buy



# Cross tabulation results.

# analysis and predictive ML model

**Steps** **undertaken for data preparation for model selection:**

1. Data preparation and feature engineering is done after data cleaning.
2. Splitting the data column or attribute-wise, i.e., dependent variable and independent variable. (18 independent variables and 1 dependent variable (attrition flag), X and Y respectively).
3. From python sklearn library, Using OneHotEncoder, encode categorical variables into 0’s and 1’s which helps better utilization and interpretation of independent values while modeling.
4. Splitting the whole dataset into a training set (75%) and test set (25%) with random state-1.
5. From sklearn using the StandardScaler module all attributes are standardized (taking mean as 0 and standard deviation from [1, -1]), which makes all the inputs in the same range for modeling. One of the important steps in feature engineering and done after splitting the data into train and test.

**Training** **datasets with following models application:**

1. **Decision tree.**
2. **Logistic regression**
3. **Random forest.**
4. **KNN classifier**
5. **Support vector machine (linear, RBF kernals)**
6. **Naïve bayes classifier**

Data preprocessing code

# Overfitting pondering

# 1. Generate real-time data and compare the results to get the accuracy, which would be a time constraint in the current scenario (customers must opt for bank services and then wait for them to display behavior), which is not sensible.

# 2. K-fold cross-validation process: the whole training dataset is randomly broken into 10 subsets (cv=10) and fed into a model to get 10 different accuracies and deviation of this accuracy conveys that accuracy obtained in general models are reliable or not. Lesser the deviation more the reliance on models.

# Generalisation

->import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

->dataset = pd.read\_csv('Bankchurners\_2.csv')

X = dataset.iloc[:, 1:-1].values

y = dataset.iloc[:, -1].values

->from sklearn.compose import ColumnTransformer

from sklearn.preprocessing import OneHotEncoder

ct = ColumnTransformer(transformers=[('encoder', OneHotEncoder(), [1,3,4,5,6])], remainder='passthrough')

X = np.array(ct.fit\_transform(X))

->from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.25, random\_state = 1)

->from sklearn.preprocessing import StandardScaler

sc = StandardScaler()

X\_train[:, 24:] = sc.fit\_transform(X\_train[:, 24:])

X\_test[:, 24:] = sc.transform(X\_test[:, 24:])

Accuracies of classification models before K-fold cross validation

Classification is the type of machine learning model which fits good for the problem statement and this type of problem-solving approach fits well in all sectors, where prediction of disease in a patient, debtors defaulting, student eligibility for selection, target customer selection, sports results prediction, spam emails, etc. scope of application is wide.

# Future studies

* Any classification problems mentioned in generalization can be approached with this method.
* Sentiment analysis: Inputs will be “keywords”, depending upon these keywords every comment, text body, tweets, and sales conversations in-mails (long term customer or short-term customer).
* Image and voice classification.
* Imbalanced classification: More attention to the minority class of the dataset, for example, Fraud detection. (In the current dataset considered, churned customers are minority class)

# Conclusion

In light of this information, organizations strive to acquire new customers, make value-adding decisions to retain existing customers, and turn them into committed customers of firms, for to attain companies collect a vast amount of data about customers, analyze this data so that management can take necessary steps to make transactions profitable with the helpful accurate predict about customers or situations.

# References

Machine Learning A-Z™: Hands-On Python & R in Data Science. https://www.udemy.com/course/machinelearning/

https://www.kaggle.com/sakshigoyal7/credit-card-customers/tasks

https://bradleyboehmke.github.io/HOML/linear-regression.html

https://heartbeat.fritz.ai/a-practical-guide-to-feature-engineering-in-python-8326e40747c8

https://www.usi.edu/endeavor/symposium/how-to-write-an-abstract/

https://content.wisestep.com/introduction-for-project/

<https://www.mytutor.co.uk/answers/3636/GCSE/Computing/What-is-the-difference-between-data-verification-and-data-validation/>

https://www.w3schools.com/python/python\_arrays.asp

https://machinelearningmastery.com/types-of-classification-in-machine-learning/

<https://machinelearningmastery.com/normalize-standardize-machine-learning-data-weka/#:~:text=Data%20standardization%20is%20the%20process,Gaussian%20(bell%20curve)%20distribution>

# acknowledgement

we would like to express our special thanks of gratitude to *Professor Ramanath Krishnamoorthy* who gave us the golden opportunity to do this wonderful project on the “Prediction of customers churners”, which also helped us in learning a good amount of “machine learning algorithms and extensive topics” and we came to know about so many new things we are thankful to them.  
Secondly, we would also like to thank our parents and friends who helped us a lot in finalizing this project within the limited time frame.

# Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Author-2

Prasanna Kulkarni

Welingkar Institute of management, Bangalore

8050392220 / 9945904229

[Prasanna.kulkarni.we@gmail.com](mailto:Prasanna.kulkarni.we@gmail.com)

Author-1

Muskan Agarwal

Welingkar Institute of management, Bangalore

9198784248

[muskan.agarwal.we@gmail.com](mailto:muskan.agarwal.we@gmail.com)