University of New Haven
Tagliatela college of Engineering
Data Science, Computer Engineering and Computer Science

Credit Card Approval Prediction

MACHINE LEARNING

DSCI-6003-02
Date: 04/24/2022

By,
Boddu Upender

Supervisor,
Muhammad Aminul Islam

Introduction

Machine Learning is an application that provides computers the ability to learn and improve from the experience without being explicitly programmed. Machine Learning is classified into different types. They are as follows:

1)Supervised Learning:

Supervised learning is the task of machine learning that maps the input to an output based on input-output patterns to the system.
Supervised learning is classified into two types of regression and classification.

A) Regression: Regression is classified under the same branch of supervised learning. In machine learning regression algorithm try's to estimate the mapping function(f) from the variable (x) to numerical or continuous output variable(y).

Example: To predict the prices of the houses, that is the regression task because price
of the houses will be the continuous output.

B) Classification: Classification is classified under the same branch of supervised learning. In machine learning classification algorithm try's to calculate the mapping function(f) from the input (x) to discrete or categorial output variables(y).

Example: To predict the prices of the houses "sell more or less than the recommended retail price." Here, the houses will be classified whether their prices fall into two discrete categories: above or below the said price.

2)Unsupervised Learning:

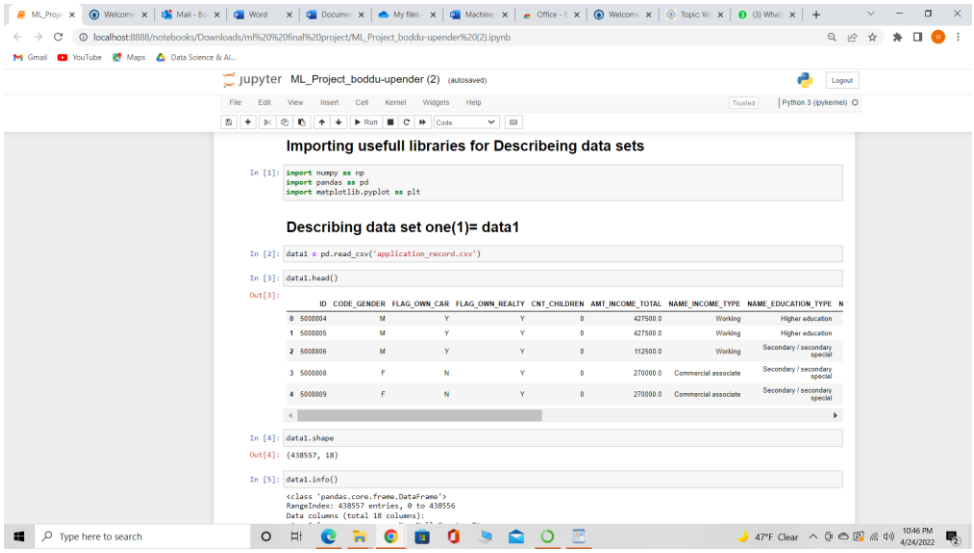
Unsupervised learning is a self-learning technique in which system has to discover the features of the input population by its own and no prior set of categories are used.
Linear Regression is the example of Regression.
This project aims at developing two supervised Machine Learning Model- Linear Regression from scratch using Python and NumPy. The developed code has various parameters. From the given models Bias-Variance Trade-off and Prediction the best model can be selected from the given data. This project enables the users to use the parameters and select the best model in Linear Regression.

Motivation

I have selected this project because it is one of the main aspects of life. It is important to know if your credit card gets approved or not well before in time to plan life and other situations. It is very challenging to predict this information by humans as there are several dependent variables. My aim is to Build a machine learning model which is used predict whether an applicant is a good customer or a bad customer. So many people apply for credit card, and we need to understand which applicant is good enough to get the approval. A machine can predict the approval if well trained. In this project several methods, models, and parameters are tested to get highest prediction accuracy. My aim is to Build a machine learning model which is used predict whether an applicant is a good customer or a bad customer. So many people apply for credit card, and we need to understand which applicant is good enough to get the approval

Describing Data Set:

I have downloaded the data from Kaggle. We know that the shape of the data set 4 lakhs 38000 rows and 18 columns. All the columns are either integer type float type or object type. For both test data and train data checked for outliers & null values now the data is balanced and cleaned.



- **ID- It shows the customer number.**
- **CODE_GENDER: It shows the gender of the customer.**
- **FLAG_OWN_CAR: It shows whether the customer owns a car or not.**
- **FLAG_OWN_REALTY: It shows whether the customer owns a property or not.**

CNT_CHILDREN: It shows how many children does customer have.

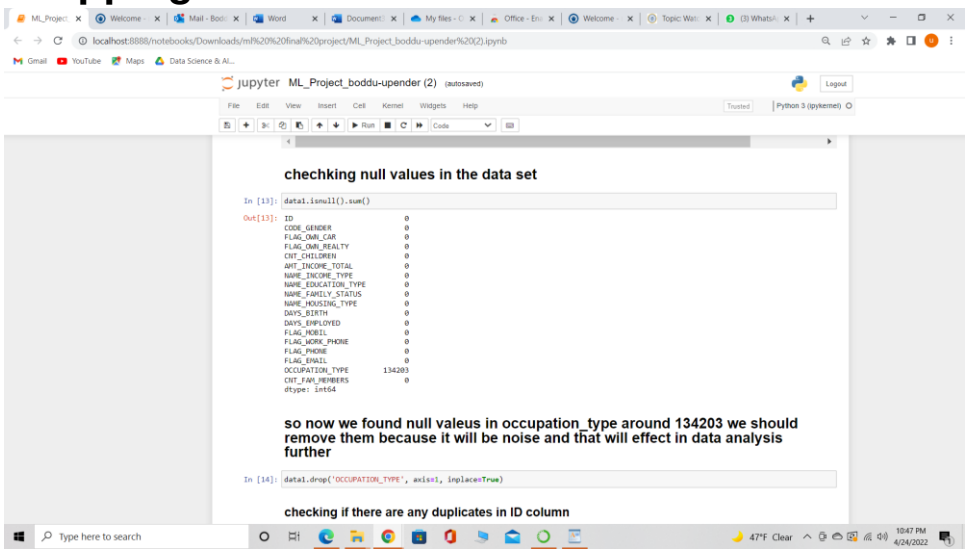
- **AMT_INCOME_TOTAL: it shows the yearly income of the customer.**
- **NAME_INCOME_TYPE: It shows what type of income does the customer get.**

NAME_EDUCATION_TYPE: It shows the qualification of the customer.
• MARITALSTATUS: It shows whether the customer is single/married/widow/divorcee.
• NAME_HOUSING_TYPE: It shows the way of living of the customer.
• DAYS_BIRTH: It shows the date of birth of the customer.
• DAYS-EMPLOYED: It shows from when the customer is working.
• FLAG_MOBILE: Does the customer have mobile.
• FLAG_WORK_PHONE: Does the customer have work phone.
• FLAG_PHONE: Does the customer have the phone.

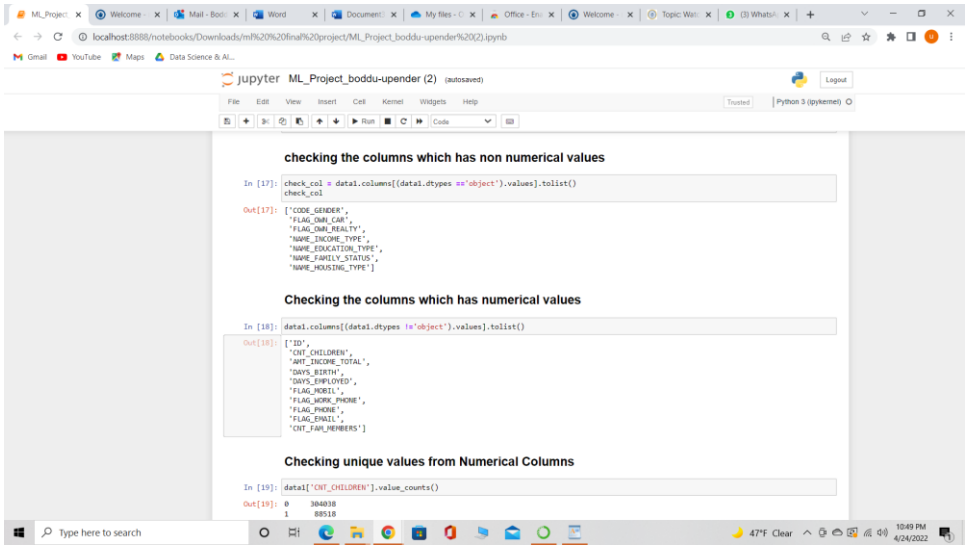
Exploring Data Analysis:

After describing data set, we should find which data is useful and which data is not useful so for that we should describe the data and find null values and duplicates and we should drop them from data

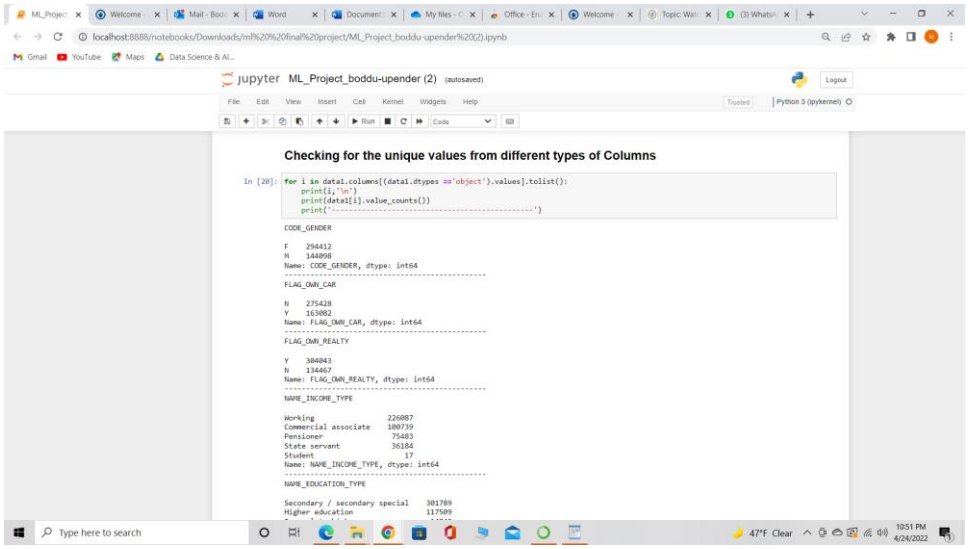
Dropping Null values -



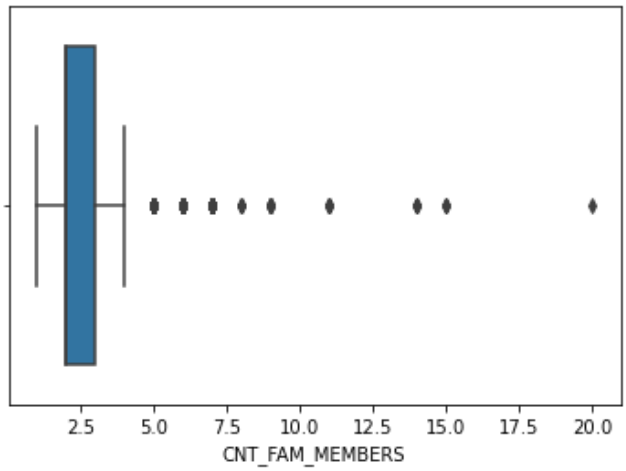
Now we should check the columns which has non numerical values, numeric values and unique values.

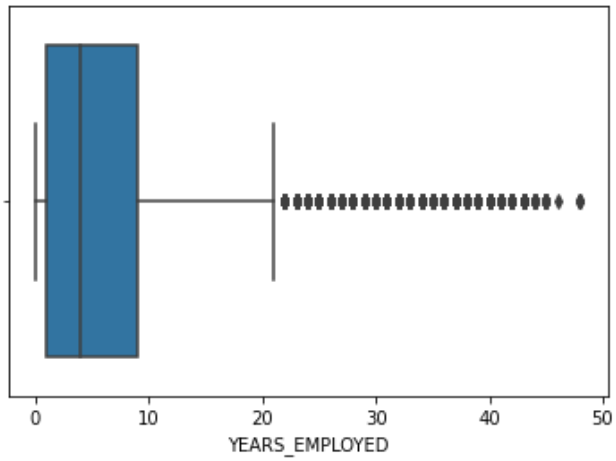


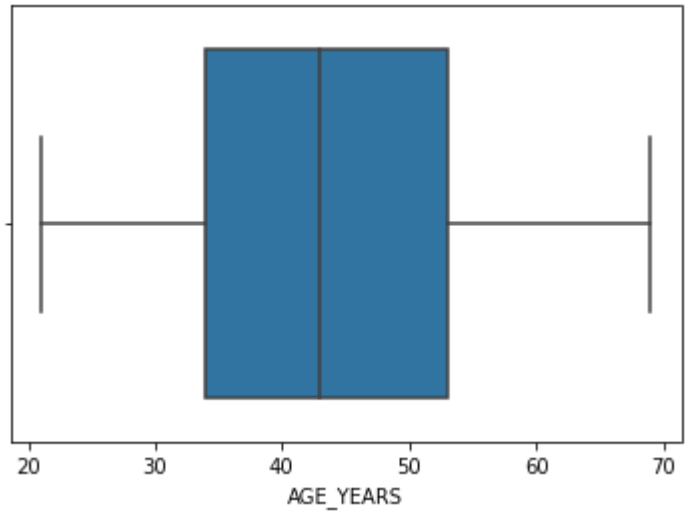
Checking the unique values for different types of Columns

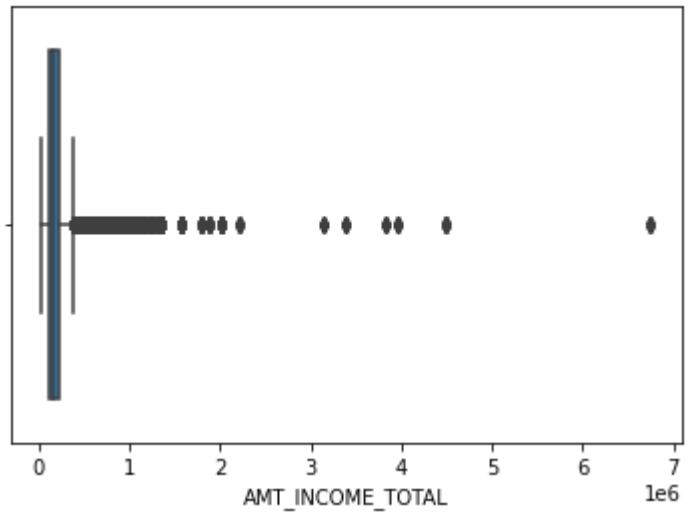


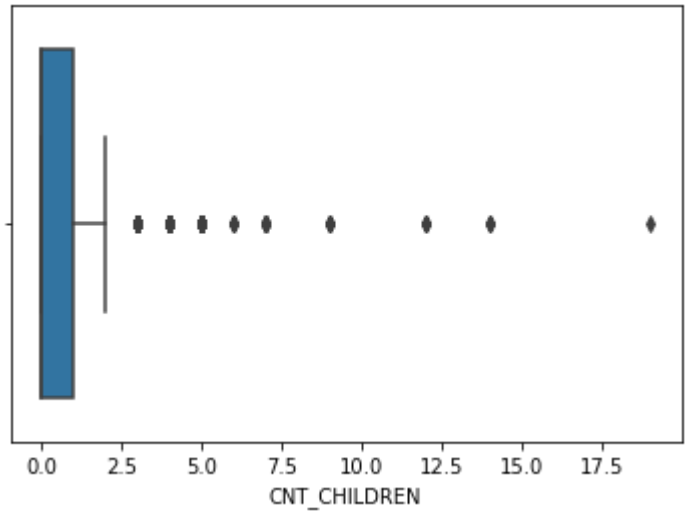
- **Removing outliers**





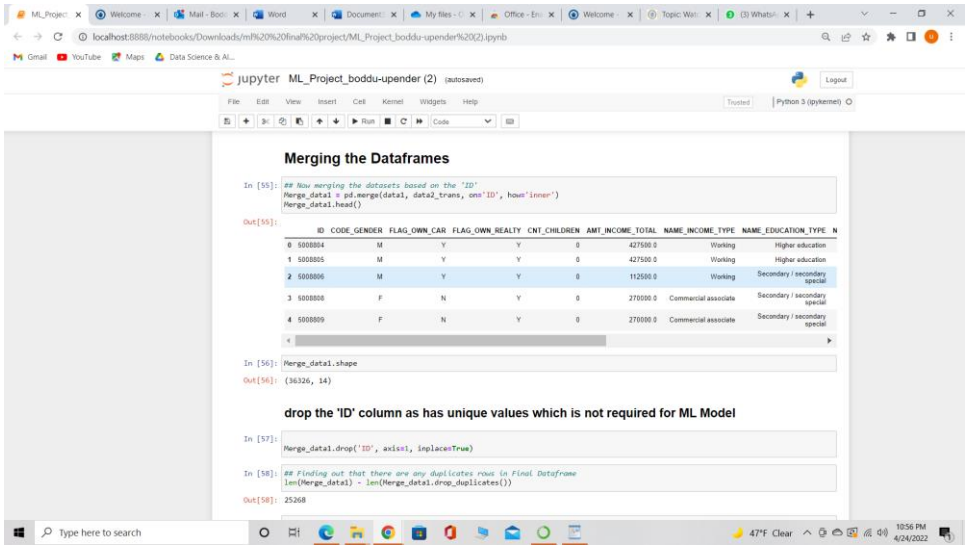

20
30
40
50
60
70
AGE_YEARS


0
1
2
3
4
5
6
7
AMT_INCOME_TOTAL
1e6


0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
CNT_CHILDREN

Merging Data frames

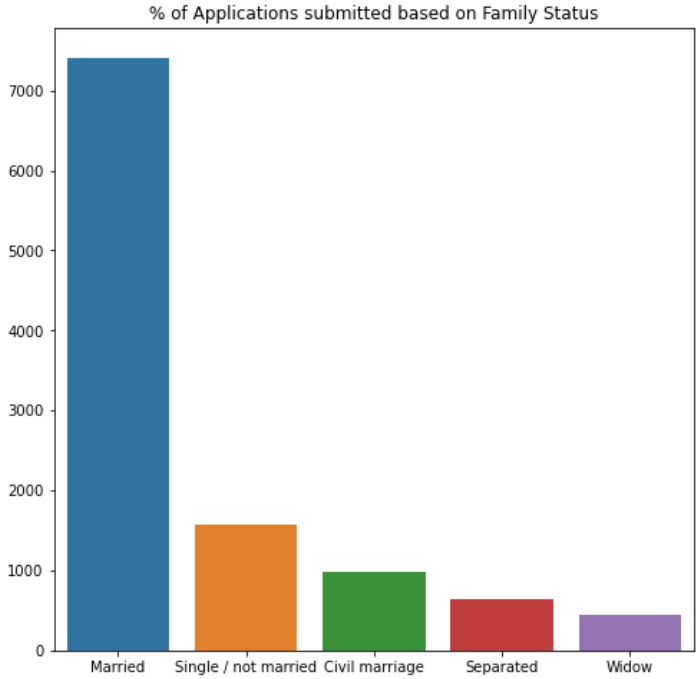
Now we should merge the data frames by dropping nulls and noise values and only taking data which is required so that we can get required and usefull features from that data which can be used for training and testing data and building the ML models.

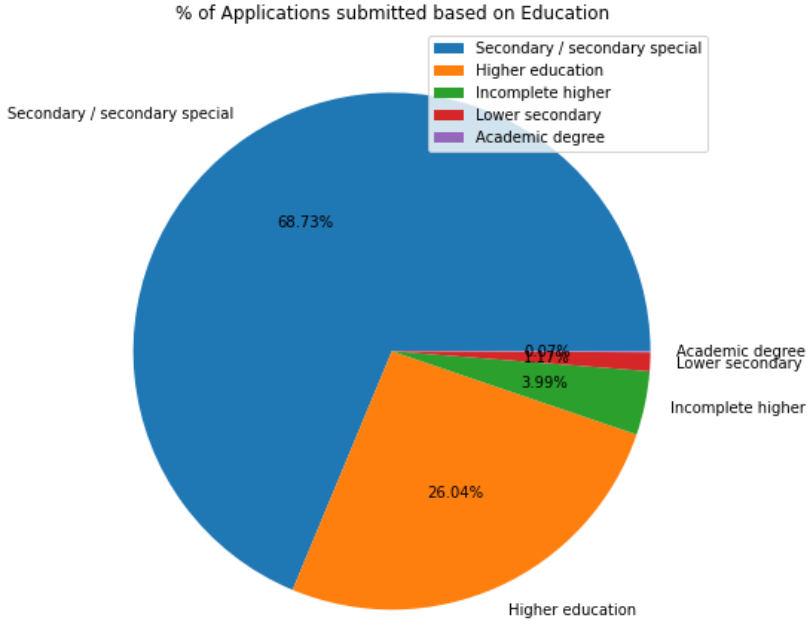


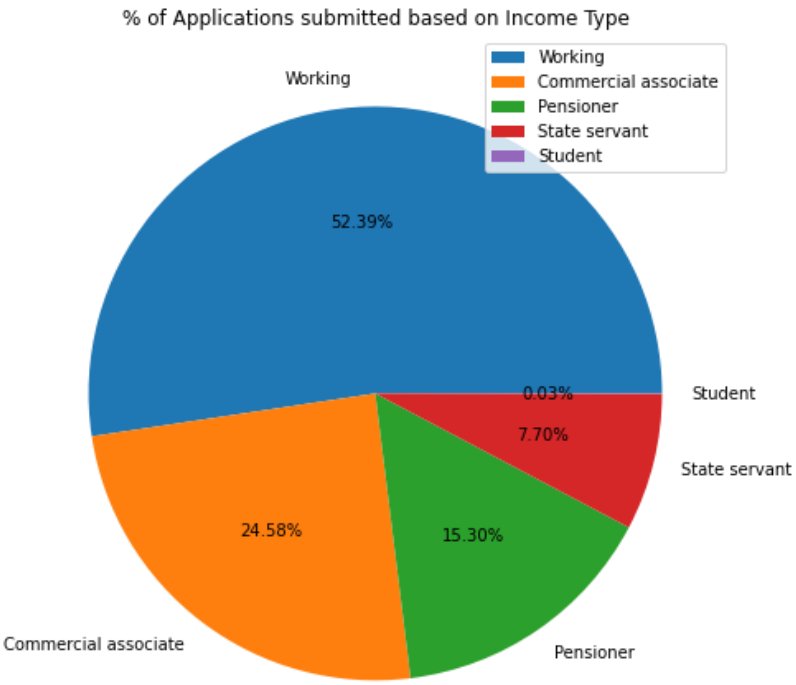
Drop the 'ID' column as has unique values which is not required for ML Mode

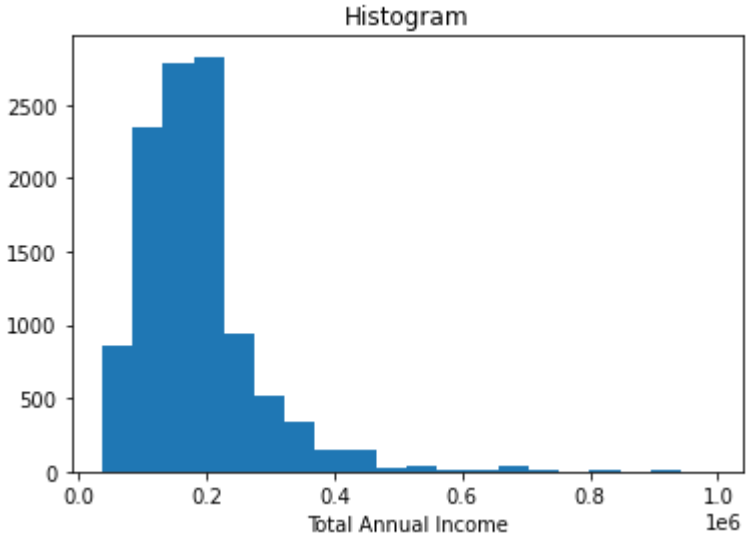
Merge Visualization

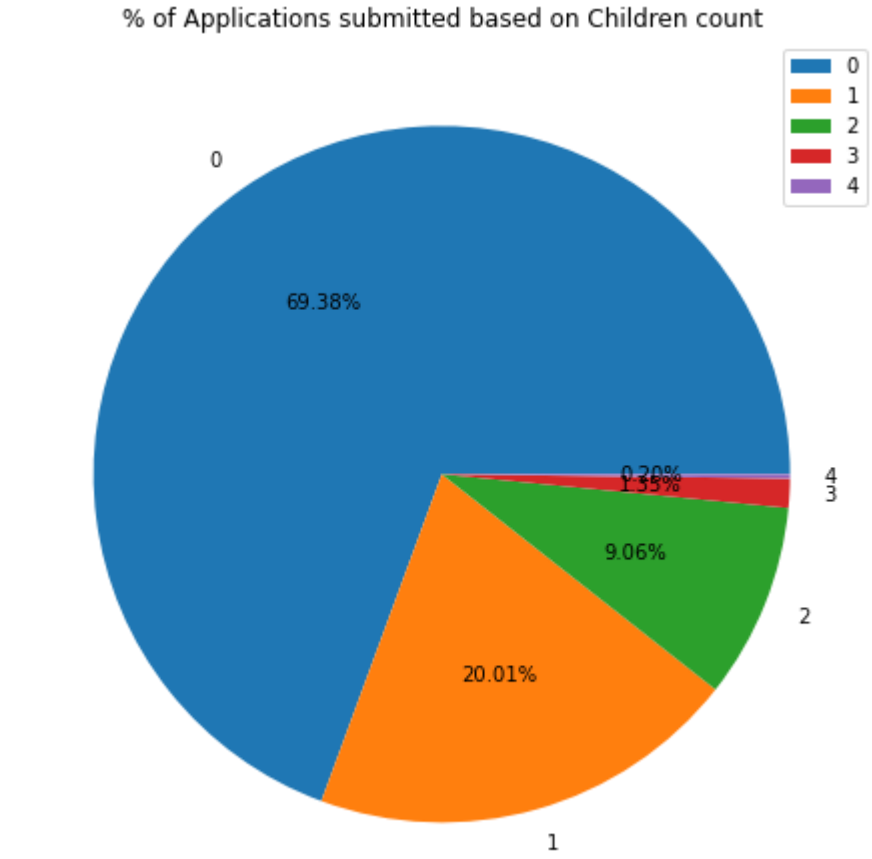
The below heatmap shows that, there are no such column / Feature which is highly co-related with 'Status'
Wherever there is 1 or light colors, there is high coorelation between the variables

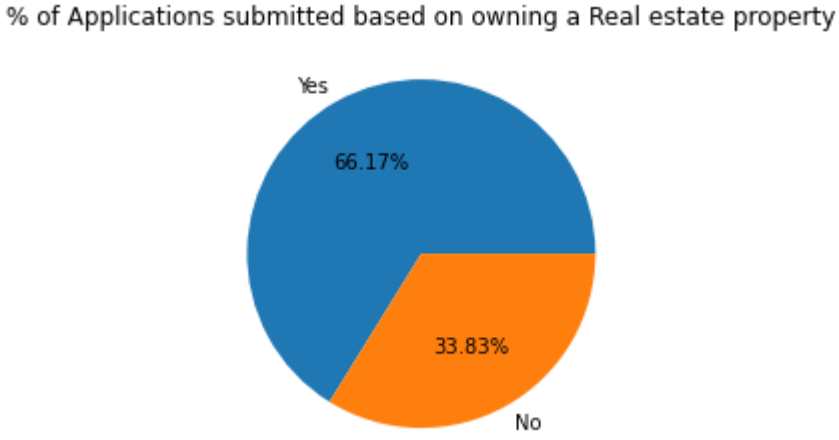

% of Applications submitted based on Family Status
7000
6000
5000
4000
3000
2000
1000
0
Married
Single / not married
Civil marriage
Separated
Widow


% of Applications submitted based on Education
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
Secondary / secondary special
68.73%
0.07%
1.17%
3.99%
26.04%
Academic degree
Lower secondary
Incomplete higher
Higher education

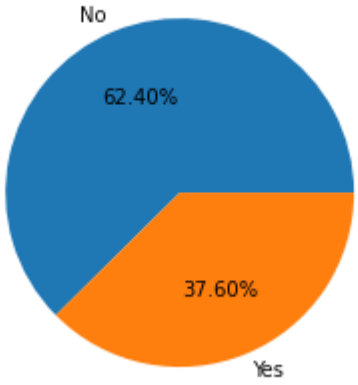

% of Applications submitted based on Income Type
Working
Commercial associate
Pensioner
State servant
Student
52.39%
0.03%
7.70%
24.58%
15.30%
Student
State servant
Commercial associate
Pensioner


Histogram
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Total Annual Income

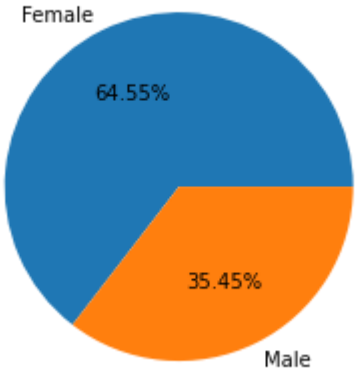

% of Applications submitted based on Children count
0
1
2
3
4
0
69.38%
4
0.20%
3
1.55%
2
9.06%
1
20.01%


% of Applications submitted based on owning a Real estate property
Yes
66.17%
No
33.83%

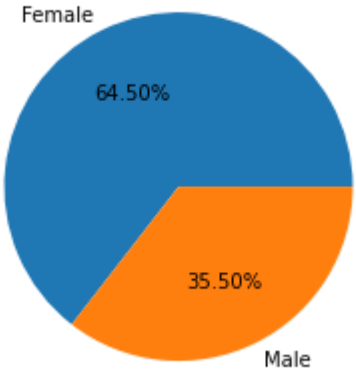
% of Applications submitted based on owning a Car

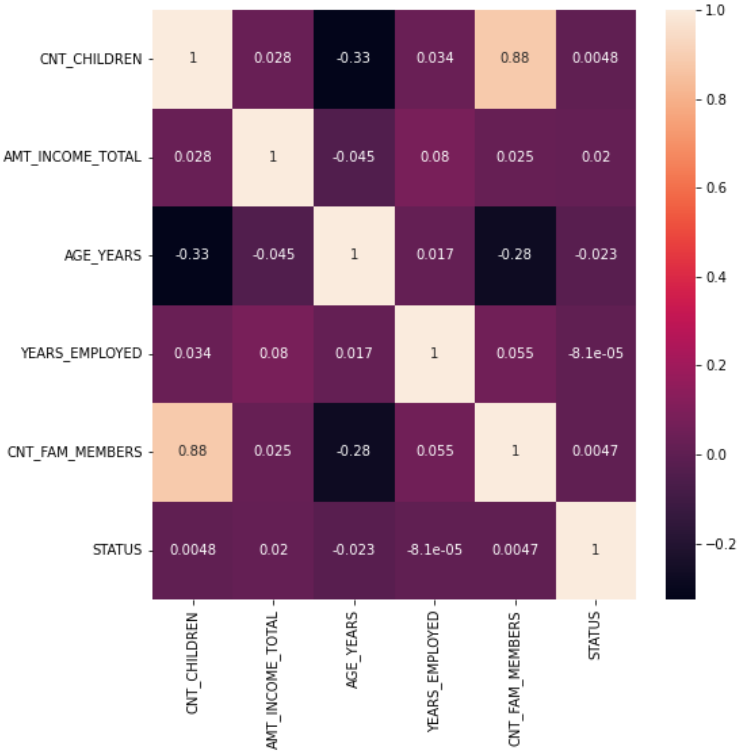


% of Applications Approved based on Gender



% of Applications submitted based on Gender



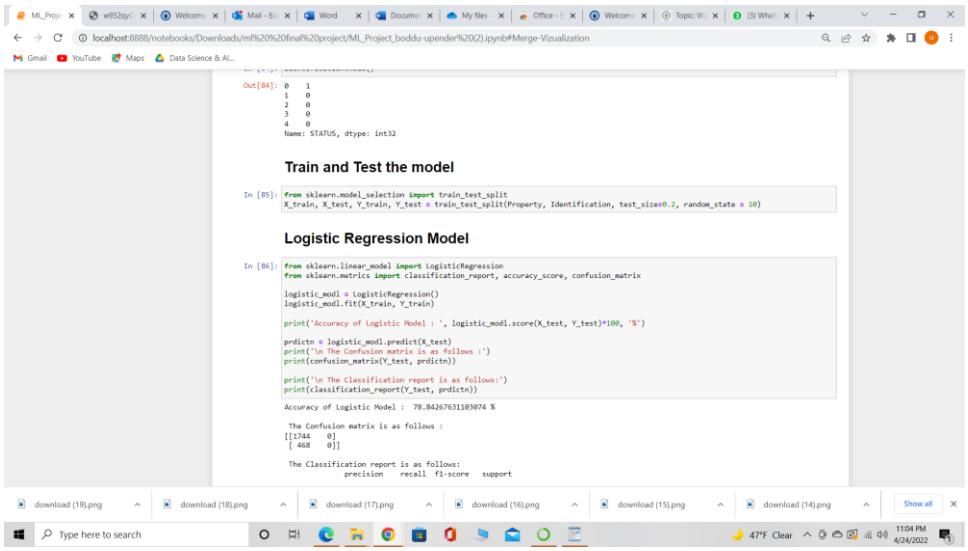


Feature Selection

```
'CODE_GENDER',
 'FLAG_OWN_CAR',
 'FLAG_OWN_REALTY',
 'NAME_INCOME_TYPE',
 'NAME_EDUCATION_TYPE',
 'NAME_FAMILY_STATUS',
 'NAME_HOUSING_TYPE
```

Converting all the Non-Numerical Columns into Numerical columns
Checking for correlation between variables

Then we should train and test the final data which we got from pre-processing by selecting features which plays a crucial role.



Modelling

In this project I am modelling 2 different types of algorithms and they are as follows:

- Logistic Regression Model.
- Decision tree classification.

Deliverables

Jupyter notebook presentation outlining

- Data cleaning
- Data exploration
- Implementation of machine learning algorithms
- Results of different methods
- Comparison of different methods
- Evaluation methods
- Selection of best methods

Result

The below are the methods used to evaluate the models.
Test accuracy.
Classification report.
Confusion matrix.

- Logistic Regression Model (78.84%) have more accuracy than Decision tree classification(73.50%).