**Introduction:**

In the recent times, one of the most important transformations taking place is the shift towards the digitalization. One of the sustaining areas where digitalization has become a trend is in the cashless transaction activities. This has become very prominent method and more people are inclined towards the cashless transactions as this reduces the risk of misplacing the physical cash. So, many financial institutions across the world are providing cashless means for their users like debit and credit cards. One of the most prominent options is having a credit card. Most people rely on credit cards to perform their transaction activities as it is a very easy way of making their payments. The private banks rely on consumer information like their basic info, living standards, salary, yearly and monthly returns, their current livelihood income source etc. All this info is reviewed before processing the application forms with the bank. This complete check and analysis can help the banks get through the hassle of enduring the technical / non-technical losses and the loss of time to both the customers and the banks so that the concentration is more towards the productive outcomes of both parties. A proper analysis is required as we see tremendous growth in the financial sector to avoid any kind of potential risk related to the unethical consumer.

With this scope in our minds, we have considered to take an initiative with the project in analysing how the financial institutions can achieve the productivity by taking many variables into our account. The main objective of the project is to analyse the credit risk criteria by taking many consumer factors into the consideration. Our goal is to apply different methodologies to form a strategy which determines if a consumer is eligible for a loan or not. To achieve this, we have considered taking a credit risk proposal dataset from **Kaggle**.

<https://www.kaggle.com/code/advaithmenon14/credit-card-approval-prediction-using-ml/data>

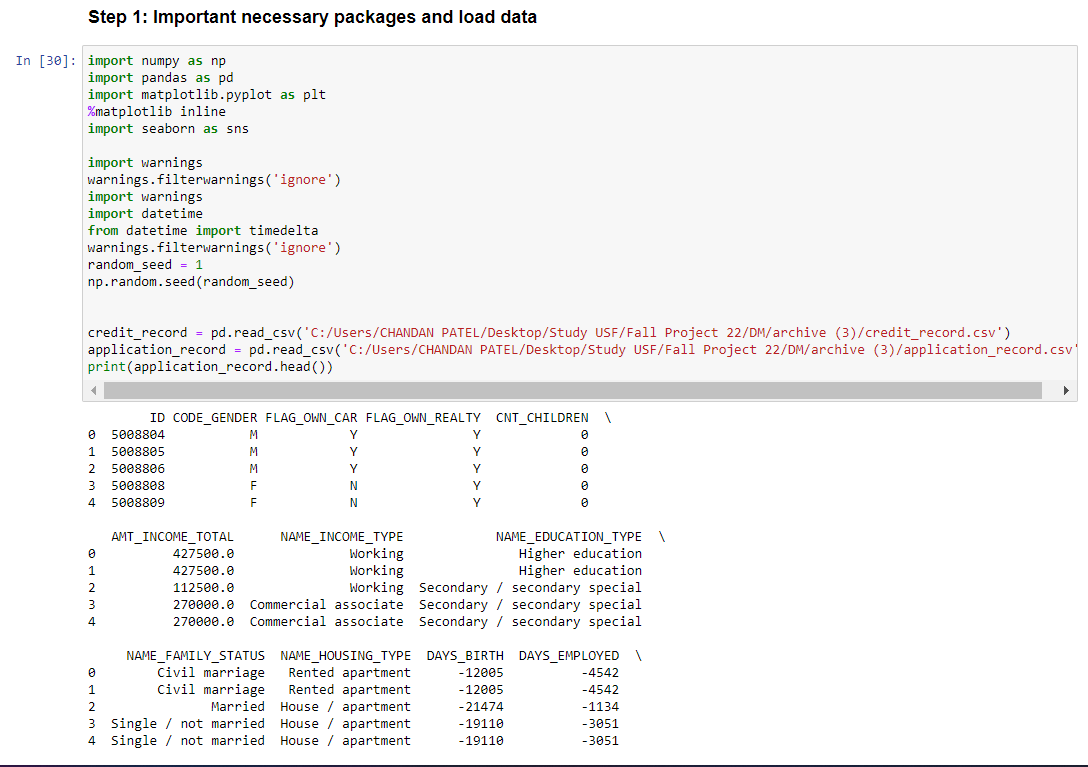
As already being mentioned before, Credit risk is associated with the possibility of a client failing to meet contractual obligations, such as mortgages, credit card debts, and other types of loans. Our datasets from Kaggle contains of two files, application\_record.csv containing of over 43000 rows and 18 columns, and the credit\_record.csv detailing the Dependent variables. Firstly, by understanding the dependent variables given in the credit\_record.csv, we could see that under the STATUS column, the **STATUS showing as 0** describes that the consumer is 1-29 days past due the payment for their credit card. The **STATUS 1** describes the consumer is 30-59 days past due the payment for their credit card. **STATUS 2** describes the consumer is 60-89 days overdue, **STATUS 3** describes 90-119 days overdue, **STATUS 4** shows 120-149 days overdue and **STATUS 5** shows the overdue or bad debts, write-offs for more than 150 days. The **STATUS C** details about the paid off that month and the **STATUS X** describes there is no loan for the month. In the **MONTHS\_BALANCE** column, assuming the month of the extracted data as the starting point, the value 0 is considered to be of the current month, -1 is of the previous month, and so on.

Hence based on the details mentioned above, our task is to implement certain methodologies and the machine learning algorithms on our Independent Variables in the application\_record.csv dataset to validate how efficiently our machine learning algorithms perform in accessing the credit risk of the consumers while various factors are taken into consideration. Our assessment is done based on the IV’s such as the consumer’s gender, if they own a car, if they own a property, what is the housing type, their annual income etc. For more details regarding each IV, please review it under the Kaggle link above.

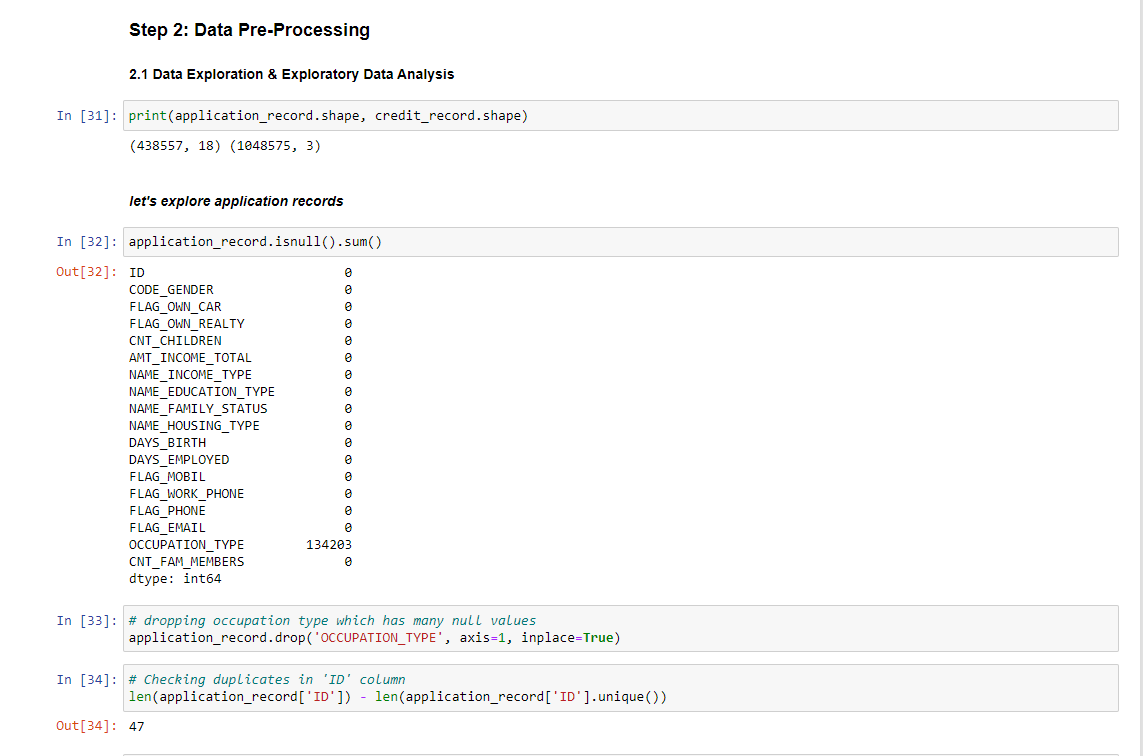
**Data Pre-processing:**

As we validated the values under each IV, we could see that the dataset contained a mixture of both numerical and non- numerical features. The values are also of different ranges and the dataset had a number of missing values. Hence, it is a liability that the data has to be pre-processed before running different machine learning algorithms; so that the resulting predictions are accurate. After the data is in good shape, we will perform some explanatory data analysis on this data and build our intuitions on it. In order to do so, we have installed some necessary packages as below:

Note: we have accomplished the data pre-processing and processing using the Jupyter Notebook.



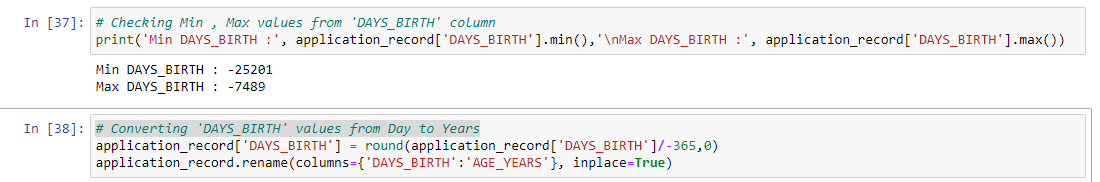
Once the files are imported and the packages are installed, we will start with our initial pre-processing of the data.

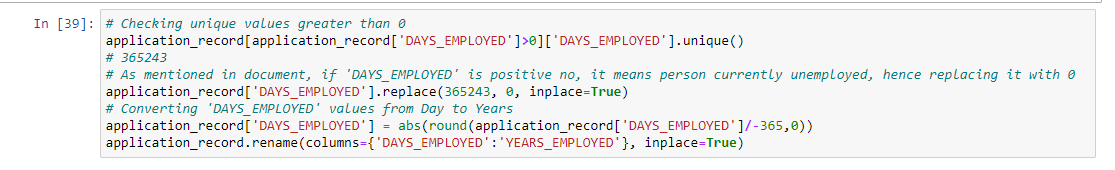


From the above screen capture, it can be seen that we are initially validating the null values and then are dropping the occupation type record which contains many null values in it. Then we have looked for the duplicate data in the ID column and have dropped the duplicates.

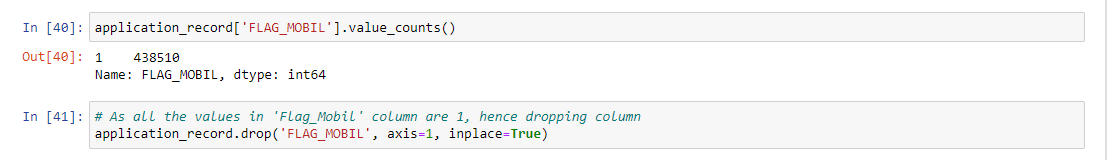


Later we have changed the DoB and record containing dates in the form of days to years for easier assessment.

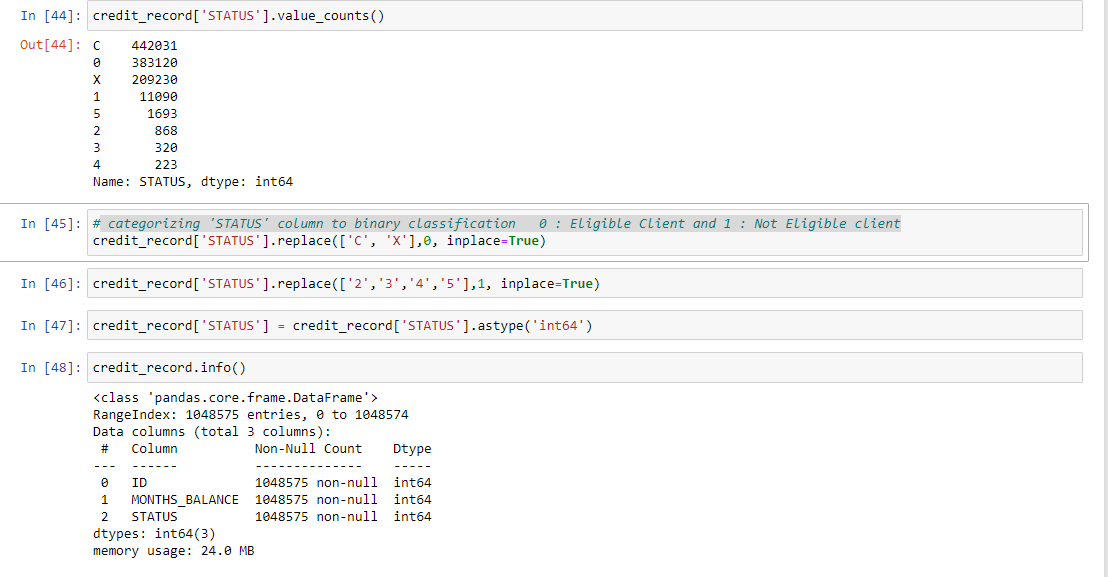




Also, notice all values in the FLAG\_MOBIL are the same, hence dropping the column.

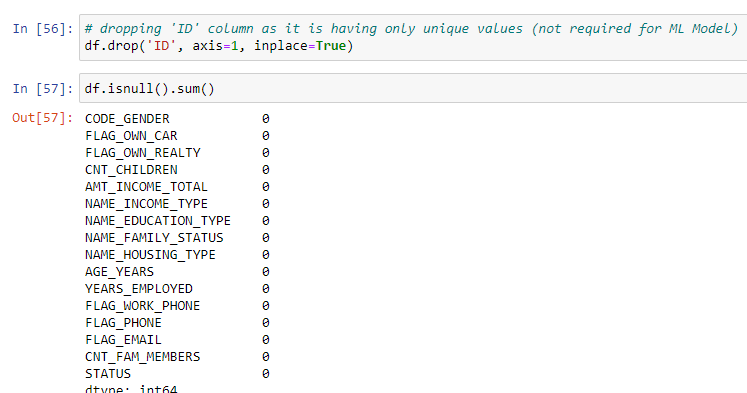


Now, we are done cleaning the IV dataset. Hence, we move forward by pre-processing the DV dataset, credit\_record.csv. Firstly, we categorize the 'STATUS' column to binary classification where 0 means the client is **Eligible** and 1 means the client is **Not Eligible**.



Later we have merged both the datasets.



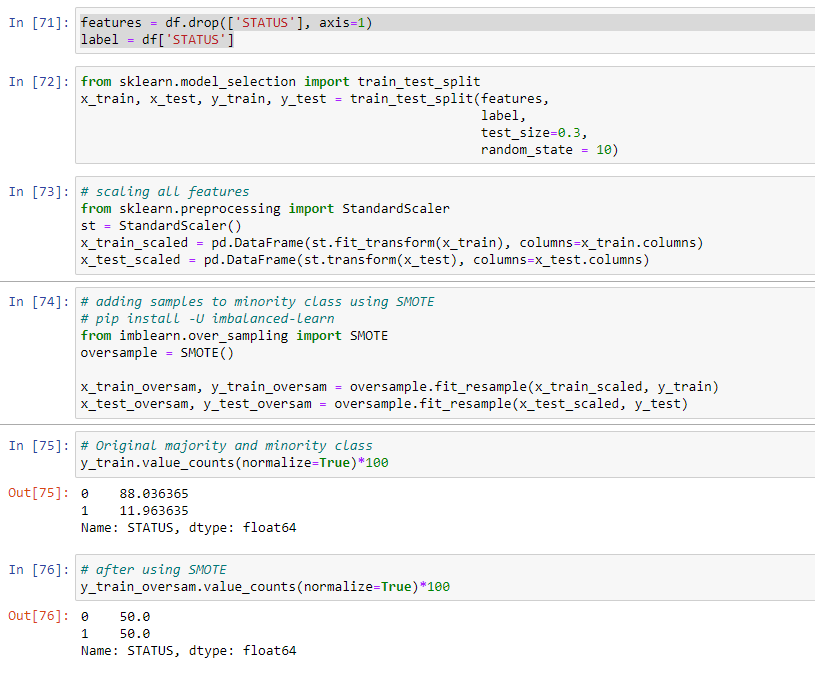
We have found it is best to drop the ID column as it had unique values and is not of much use.

After doing our analysis of the data using the corelation, we have found the below:

1. Majority of the applications submitted are by the females.
2. Majority of the approved applications are of females.
3. Majority of the applicants don’t own a car.
4. Majority of the applicants own a properly/house.
5. Majority of the applicants don't have any children.
6. Majority of the applicant’s income lies between 100,000 to 300,000.
7. Majority of the applicants are working professional.
8. Majority of the applicants are married.
9. Majority of the applicants lives in House / Apartment.

After converting all the non-numeric columns to numeric, we were finally able to get the cleaned data file.

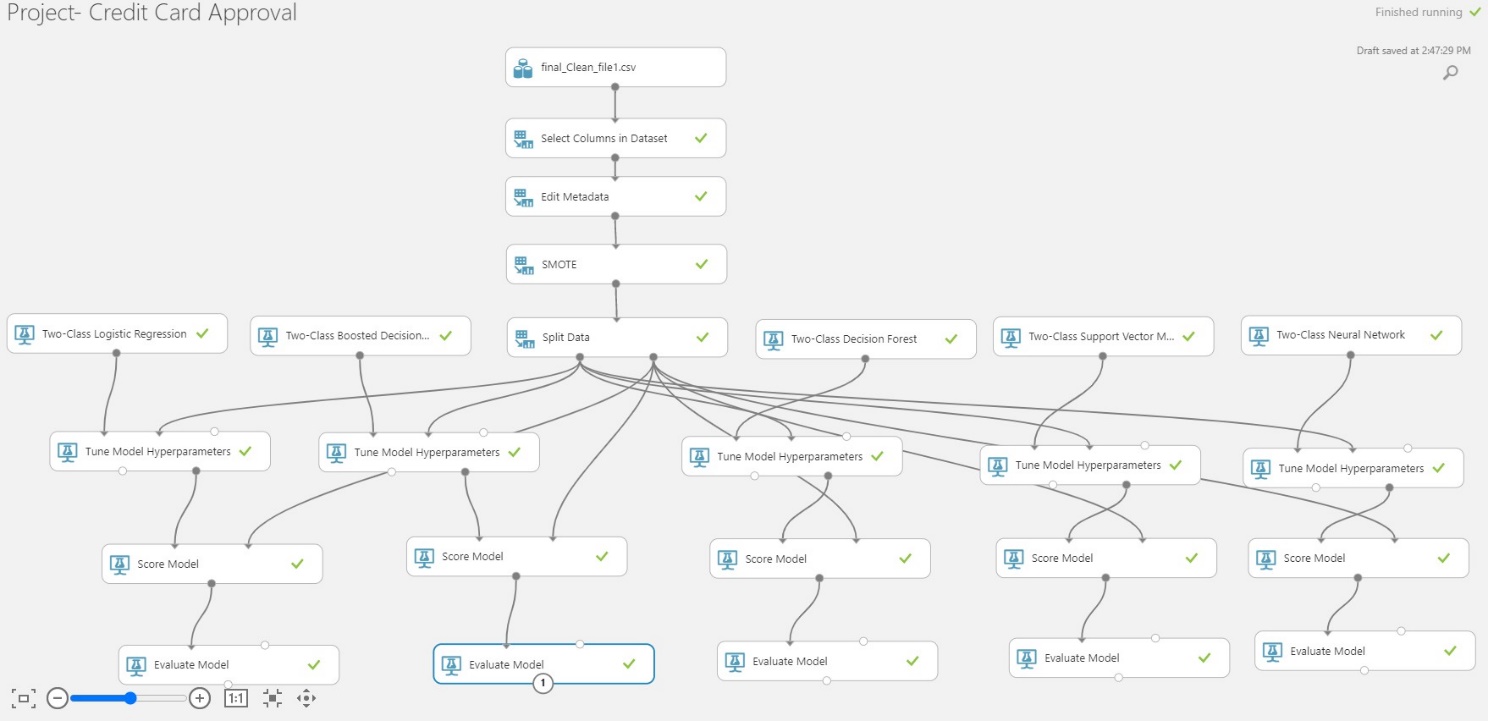
It is also very important to deal with the imbalances in the datasets after we are merging the data.



**Analysis of Credit Risk using Machine Learning:**

In order to make predictions whether the consumer falls under the risk category or not by using different machine learning algorithms and selecting the best performing model, we are choosing the Azure Machine Learning Studio to perform the operations.

Firstly, we have created a pipeline where the plan is to use five best machine learning algorithms to make predictions. In order to do so, we used the cleaned file. Now the data is heavily imbalance after the merging the IV dataset and the DV dataset. In order to deal with the imbalances, we have used **SMOTE**. SMOTE is used when the class that we are dealing with is unprecedented. It increases the number of cases in your dataset in a balanced way. The component works by generating new instances from existing minority cases that you supply as input. This implementation of SMOTE does not change the number of majority cases. The new instances are not just copies of existing minority cases. Instead, the algorithm takes samples of the feature space for each target class and its nearest neighbours. The algorithm then generates new examples that combine features of the target case with features of its neighbours. This approach increases the features available to each class and makes the samples more general.



Splitting data:

The data was split into test and train. 70% is considered to be the train and the rest as test. Then we applied the machine learning algorithms.

Algorithms:

We have used the below five Machine learning algorithms to make predictions:

1. Two Class Logistic Regression
2. Two Class Boosted Decision Tree Model
3. Two Class Decision Forest
4. Two Class Support Vector Machine Model
5. Two Class Neural Network

Tuning MODEL Hyperparameters:

We use hyper tunning module because it let us control the model training process. For example, with neural networks, you can decide on the number of hidden layers and the number of nodes in each layer. Model performance depends heavily on hyperparameters. Hyperparameter tuning, also called hyperparameter optimization, is the process of finding the configuration of hyperparameters that results in the best performance. The process is typically computationally expensive and manual. Azure Machine Learning lets you automate hyperparameter tuning and run experiments in parallel to efficiently optimize hyperparameters.

We used Random sweep mode in the module and the Maximum number of runs on the random sweep was 5. We used this because we want to increase model performance by using the metrics of our choice and simultaneously conserve computing resources also.

**RESULTS:**