**Capstone Project**

# Credit score classification- Interim Report

**Mentor: Mohit Sahu**

**Team:**

**Anusha C , Pranav Prashant, Souvik Roy, Sujan Shetty, Vishal, Vikas G.D**

**Table of Contents**

1. **Overview – Industry Background and Key Objectives**
2. **Data Dictionary**
3. **Data Pre-processing:**

* Imputation of missing values
* Grammatical rectification and other cosmetic improvements

1. **Initial Data Exploration**

* Analysing Relationships between Variables
* Univariate analysis
* Bi-variate analysis
* Multi-Variate analysis
* Treatment of Outliers
* Treatment of Imbalanced data

1. **Model Building and additional data treatments:**

* Splitting, Scaling and Encoding data
* Model training
* Model performance and scoring

1. **Future Actions**

**Overview**

**Industry Background and Key Objectives**

the company has collected basic bank details and gathered a lot of credit-related information. The management wants to build an intelligent system to segregate the people into credit score brackets to reduce the manual efforts.

**Task**  
Given a person’s credit-related information, build a machine learning model that can classify the credit score.

**Business problem statement:**

Credit scores are widely used in the lending industry for assessing the creditworthiness of individuals. Financial institutions and lenders use credit scores to evaluate the risk associated with lending money or extending credit to a particular individual. Here are some key reasons why credit scores are important in the lending business:

1. Risk Assessment: Credit scores provide a standardized way to assess the credit risk associated with borrowers. Lenders can use credit scores to evaluate the likelihood of a borrower defaulting on their credit obligations. Higher credit scores indicate lower risk, while lower credit scores suggest higher risk.
2. Loan Approval: Lenders often rely on credit scores to make decisions regarding loan approvals. A higher credit score increases the chances of loan approval, as it demonstrates a borrower's responsible credit behavior and ability to manage debts.
3. Interest Rates: Credit scores play a significant role in determining the interest rates offered to borrowers. Individuals with higher credit scores are likely to be offered lower interest rates, as they are considered less risky borrowers. On the other hand, individuals with lower credit scores may face higher interest rates or may even be denied credit altogether.
4. Credit Limits: Credit scores also influence the credit limits that lenders are willing to extend to borrowers. Individuals with higher credit scores may be eligible for higher credit limits, allowing them access to more credit resources.
5. Financial Risk Management: Credit scores help lenders in managing their overall financial risk. By analyzing the credit scores of their borrowers, lenders can make informed decisions about portfolio management, risk diversification, and setting appropriate risk management strategies.
6. Compliance and Regulation: Credit scoring systems often comply with regulatory requirements imposed by governing bodies. Lenders must adhere to these regulations and ensure fair and unbiased lending practices. Credit scores provide an objective and standardized measure to ensure compliance.

Top of Form

**Data Dictionary**

ID: Unique ID of the record

Month: Month of the year

Name: The name of the person

Age: The age of the person

SSN: Social Security Number of the person

Occupation: The occupation of the person

Annual\_Income: The Annual Income of the person

Monthly\_Inhand\_Salary: Monthly in-hand salary of the person

Num\_Bank\_Accounts: The number of bank accounts of the person

Num\_Credit\_Card: Number of credit cards the person is having

Interest\_Rate: The interest rate on the credit card of the person

Num\_of\_Loan: The number of loans taken by the person from the bank

Type\_of\_Loan: The types of loans taken by the person from the bank

Delay\_from\_due\_date: The average number of days delayed by the person from the date of payment

Num\_of\_Delayed\_Payment: Number of payments delayed by the person

Changed\_Credit\_Card: The percentage change in the credit card limit of the person

Num\_Credit\_Inquiries: The number of credit card inquiries by the person

Credit\_Mix: Classification of Credit Mix of the customer

Outstanding\_Debt: The outstanding balance of the person

Credit\_Utilization\_Ratio: The credit utilization ratio of the credit card of the customer

Credit\_History\_Age: The age of the credit history of the person

Payment\_of\_Min\_Amount: Yes if the person paid the minimum amount to be paid only, otherwise no.

Total\_EMI\_per\_month: The total EMI per month of the person

Amount\_invested\_monthly: The monthly amount invested by the person

Payment\_Behaviour: The payment behaviour of the person

Monthly\_Balance: The monthly balance left in the account of the person

Credit\_Score: The credit score of the person

**Data Pre-processing**

We begin by reading in the default csv file and getting a sense of the overall size and features we will be working with.

**Our default data file includes:**

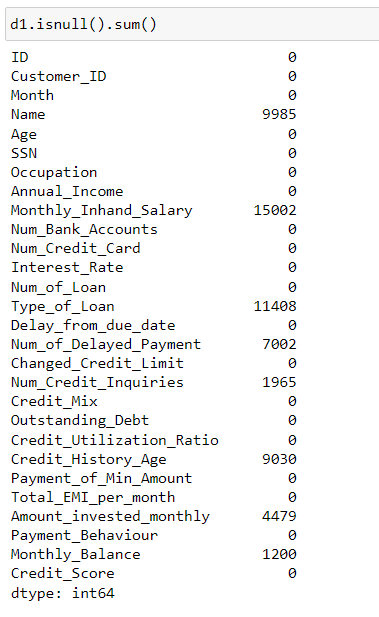
Number of Columns: 28

Total Number of Records: 100000

**Imputation of missing values:**

Our initial search helps us identify missing values. The columns ‘Occupation’, ‘Monthly In-hand Salary’, ‘Type of loan’ and ‘Num of Delayed Payment’, ’Num Credit Inquires’, ‘Credit history Age’, ‘Amount Invested Monthly’, ‘Monthly Balance’ have data missing.

We will think about treatment of null values on a feature-by-feature basis.

****

1. **Null values in ‘Amount Invested monthly’,’Monthly\_Inhand salary’, ‘Type of loan’., we replaced by b-fill ., as we don’t want to loose any data**
2. **Null values in ‘Monthly\_Balance’ , ‘Num\_Credit\_Inquiries’ are dropped ., as they have less rows**
3. **And column with less null values , we are dropping**
4. **Categorical Columns containing null values., we are replacing with mode., if there is significant mode**
5. **For numerical column we using b-fill, f-fill, and mean replacement**

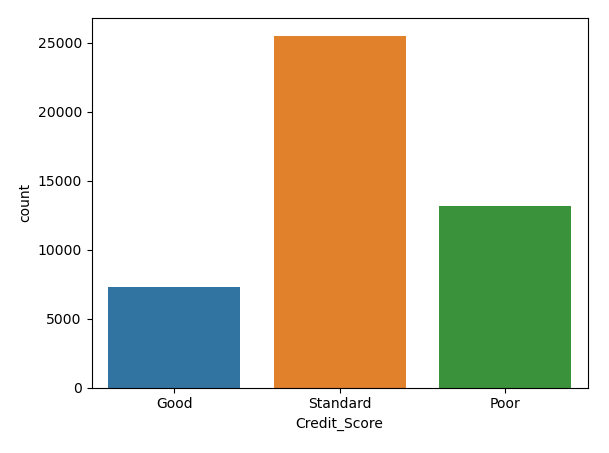
**I****nitial Data Exploration -Analysing Relationship between the Variables**

**FROM HERE , YOU CAN CONTINUE**

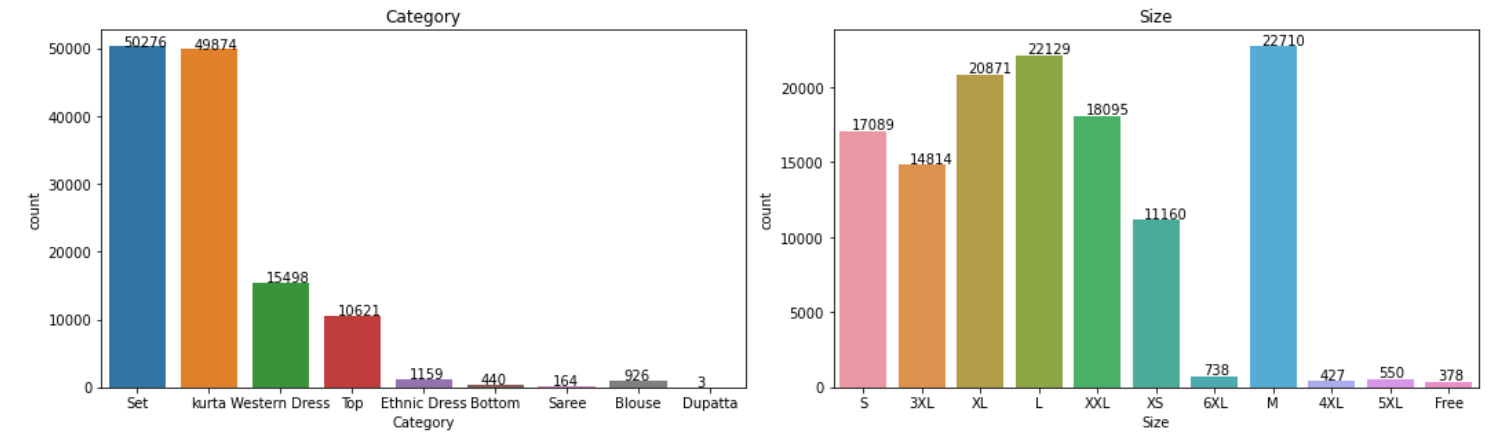
**Univariate analysis**

**Let us begin by charting the updated categorical columns**

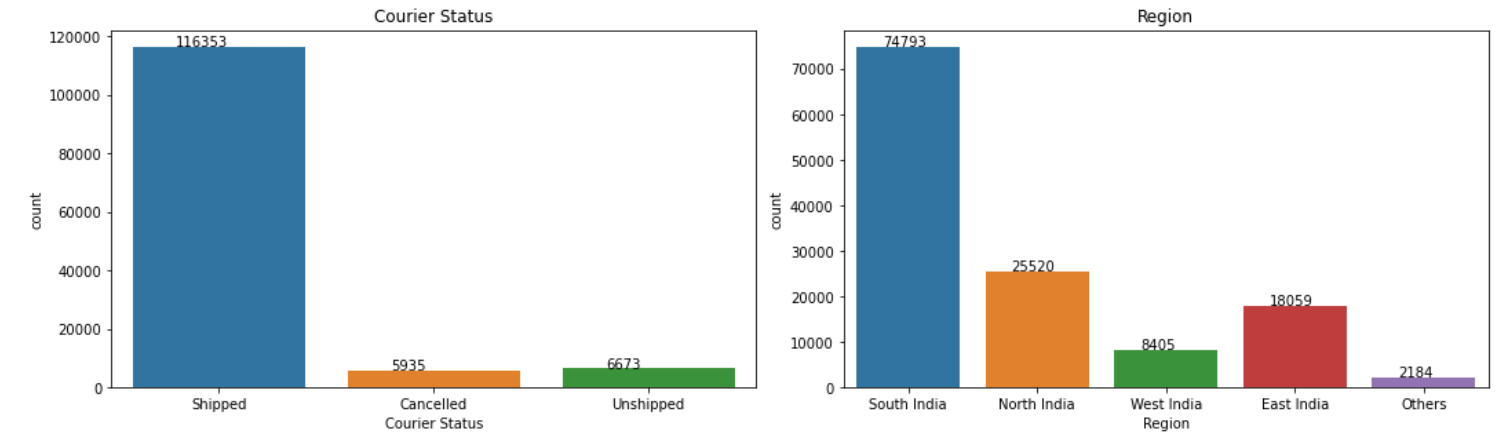
Below, we can see that a majority of the orders are delivered by Amazon itself and a majority of the orders are processed via expedited shipping, indicating that the customers for those are either in the prime/premium category or have paid additionally for faster shipping.



Next, visualising the ‘Category’ and ‘Size’ data points tells us that most of the orders placed have been either a set (combination of clothing types) or kurtas in particular; with a broad range of sizes being selected, all the way from small to extra-extra-large seeing robust demand.

****

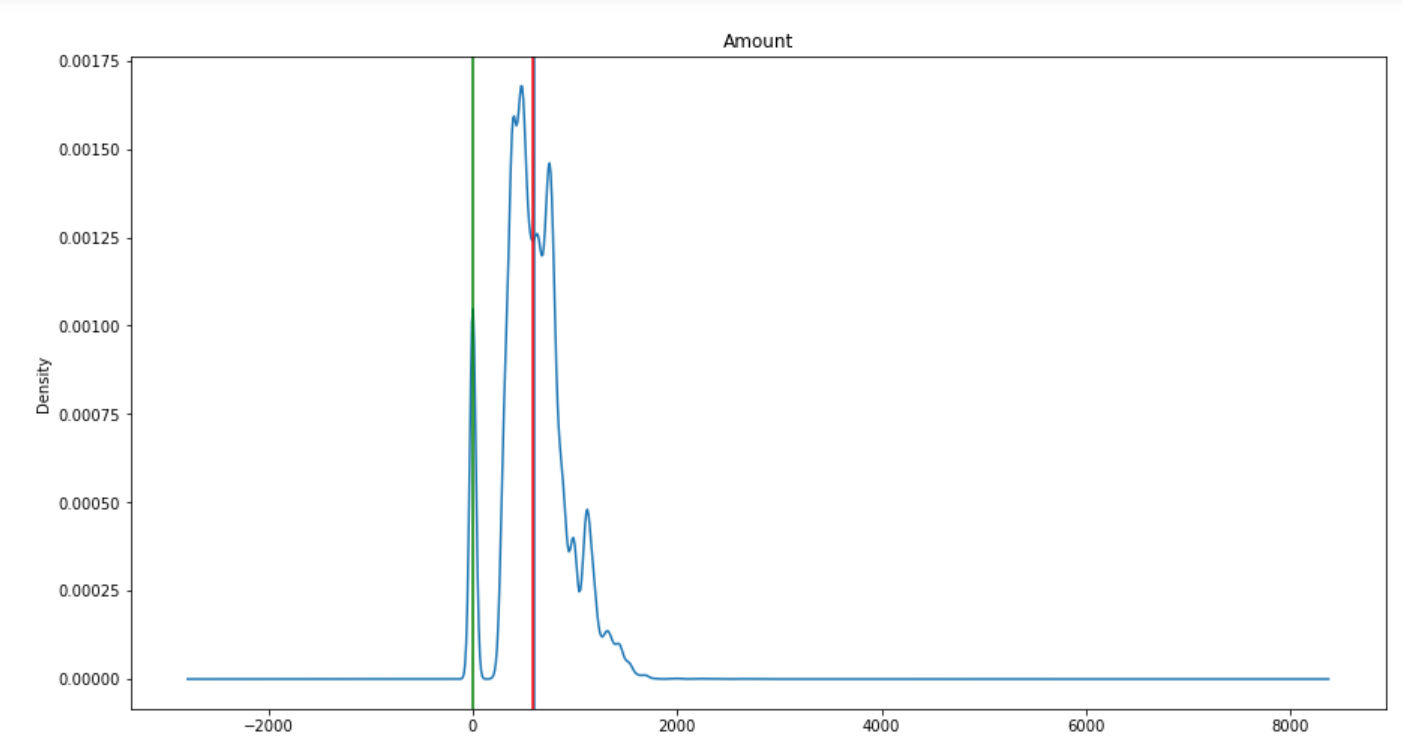
Finally, looking at the order status and comparing by region, we recognise that a majority of the orders are from southern Indian locations and most had been shipped at the point of data collection.

****

**Now, let us understand what the numerical columns are displaying**

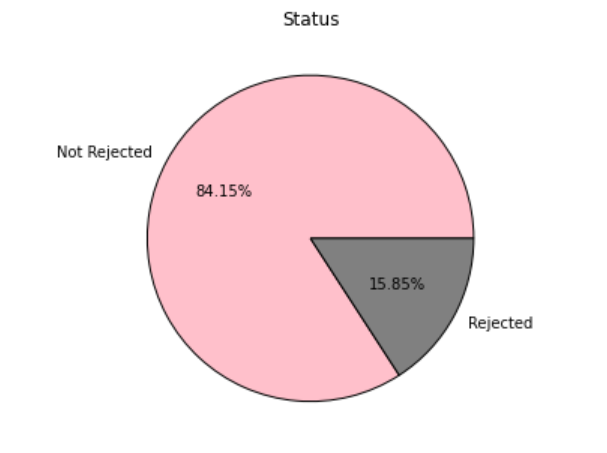
Most orders contain a single unit or a single set. These orders have an average value ofRs. 609.4, with Rs. 583 being the most common order value.

Additionally, as can be seen below, a smaller segment of the orders have a much higher price value, and it might make sense to focus on them particularly to solve pain points in order processing and delivery.



**Displaying the target variable**

We reconfirm the fact that most orders are categorised under the ‘Not Rejected’ section, with about 15.85% of the orders being categorised as ‘Rejected’. We hope to see a reasonable reduction in the % share for ‘Rejected’ post implementation of suggestions made via this analysis.

****

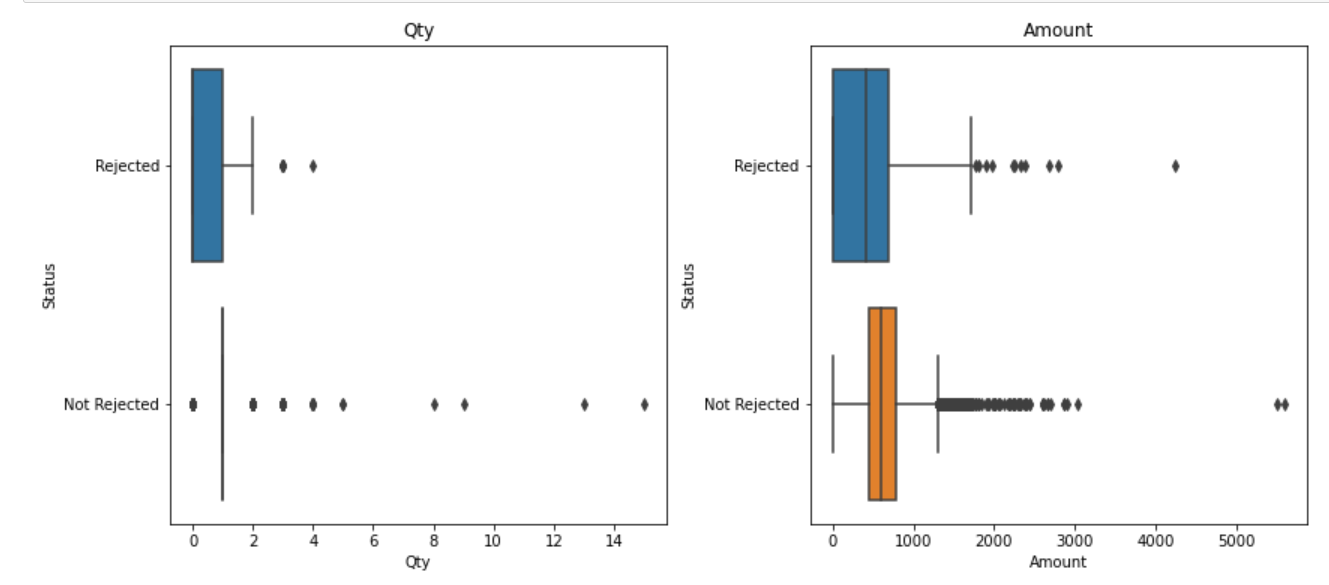
**Bi-variate analysis**

Here we will attempt to understand how the available columns are impacting each other on a one to one basis.

**First, we visualise the relationship between our target column (Status) and numerical features (Quantity, Amount)**

For quantity, larger proportion orders exist under ‘Not-Rejected’, with‘Rejected’ having mostly a single unit per order. Hence an initial inference can be made that most rejections have smaller volume per order.

Regarding the ‘Amount’ column, we see that both the rejected and non-rejected segment of orders are mostly in the ‘non-luxury’ segment though luxury segment orders can clearly be seen.

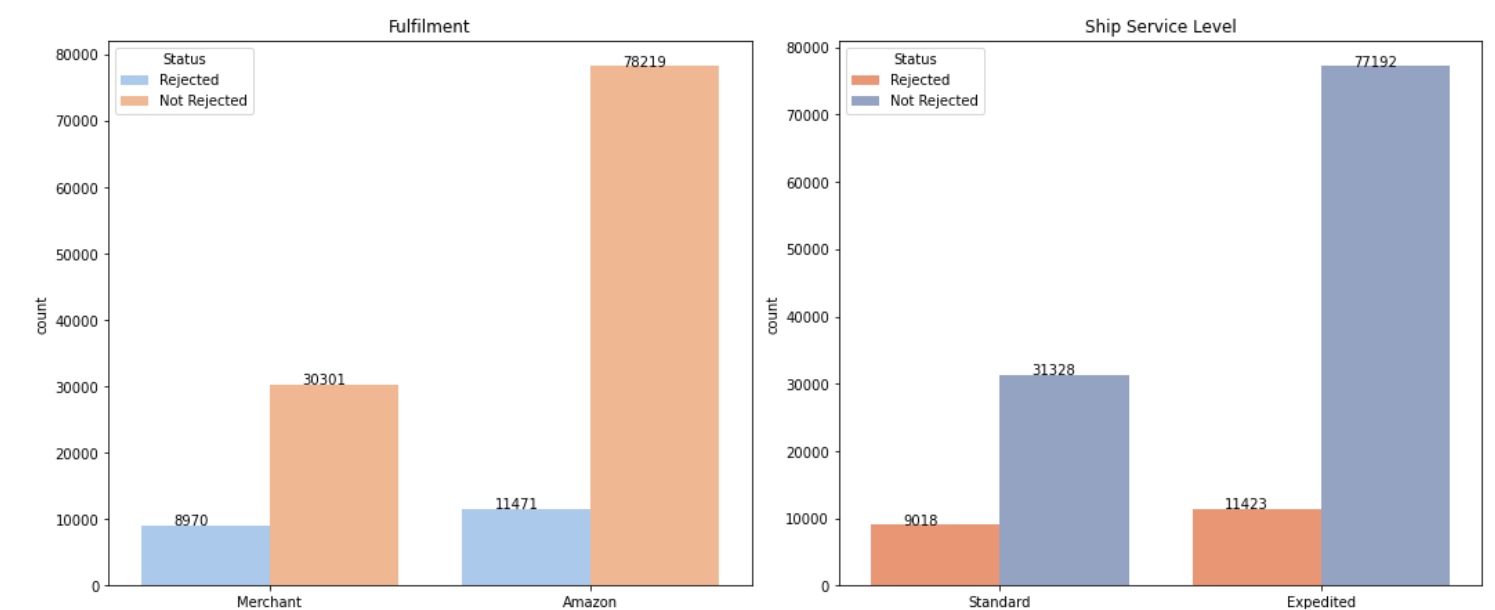
****

**Now, let us look at how categorical column data is distributed vis-à-vis ‘Status’**

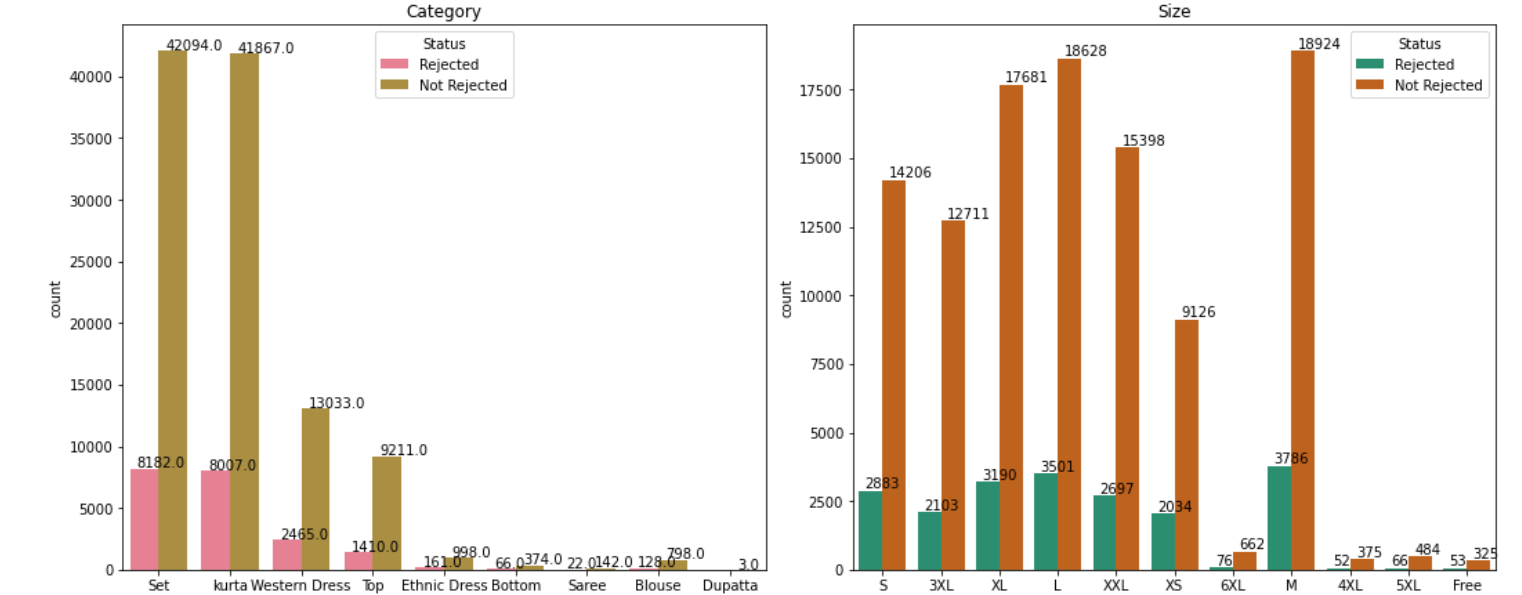
Broadly speaking, most orders come under the ‘Not Rejected’ segment, and that can be see across all visualisations.

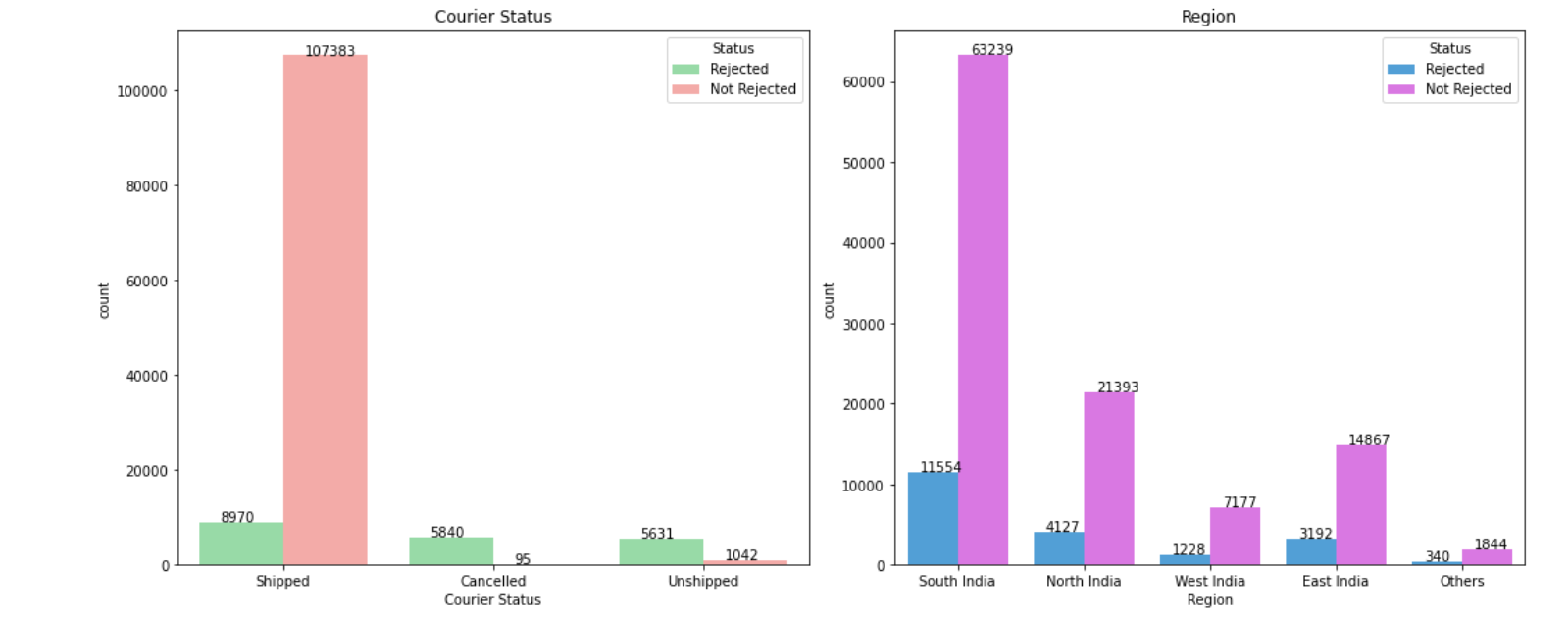
More specifically, for most orders, the delivery/fulfilment process has been handled by Amazon itself, with the rest being taken care of by third party organisations.

Most orders have been processed under the ‘Expedited’ segment, again indicating that these are orders from customers who have either paid an additional amount for faster delivery or are customer who have a premium/prime subscription of Amazon. It might be more important to check whether there are any unique pain points that occur under the delivery process for ‘Expedited’ shipping and work to solve those as a higher priority.

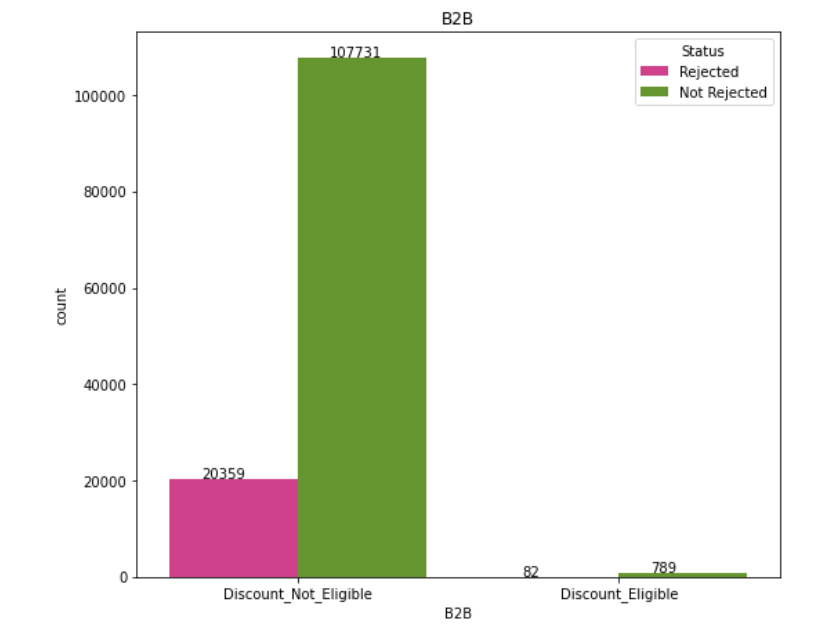


Further, we have plotted the ‘Category’, ’Size’,’ Courier Status’ and ‘Region’ individually vs ‘Status’.



****

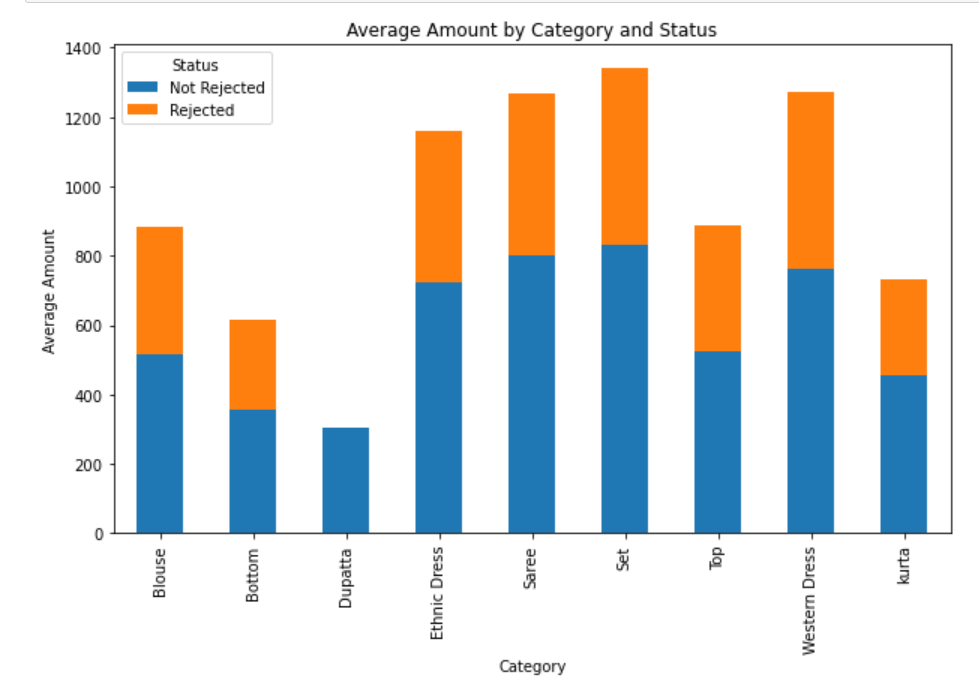
Finally, below we take a look at ‘B2B’ or the business-to-business segment of customers. An important portion of the client base considering that most of these orders have been placed without needing to provide any discounts, and keeping in mind that there might be potential to scale the relationship and thereby the order value for these customers to much higher volume than might be possible with individual customer type.

****

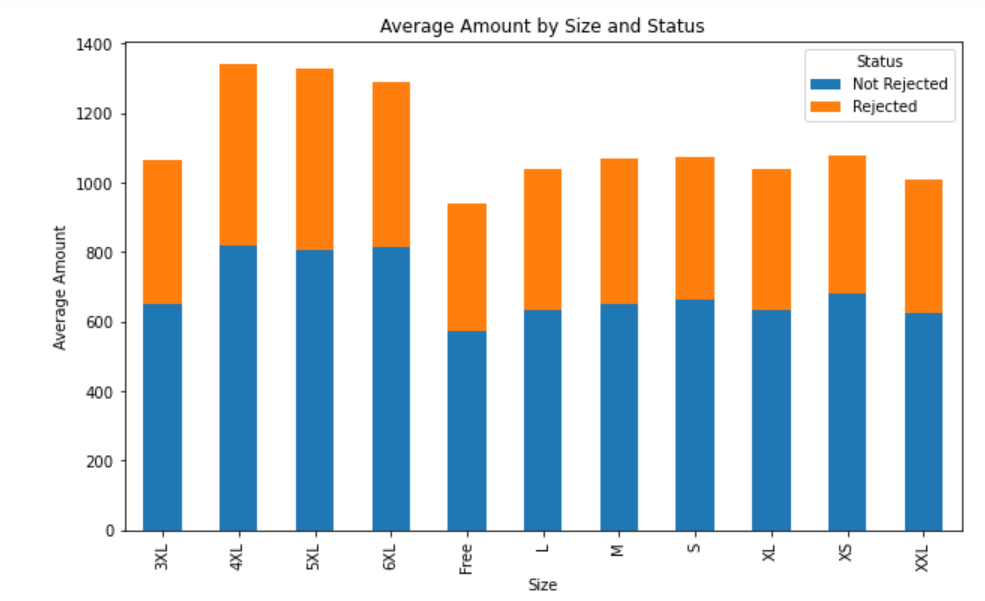
**Multi-Variate Analysis**

In this section we aim to visualise the relationship between our target variable and features more broadly.

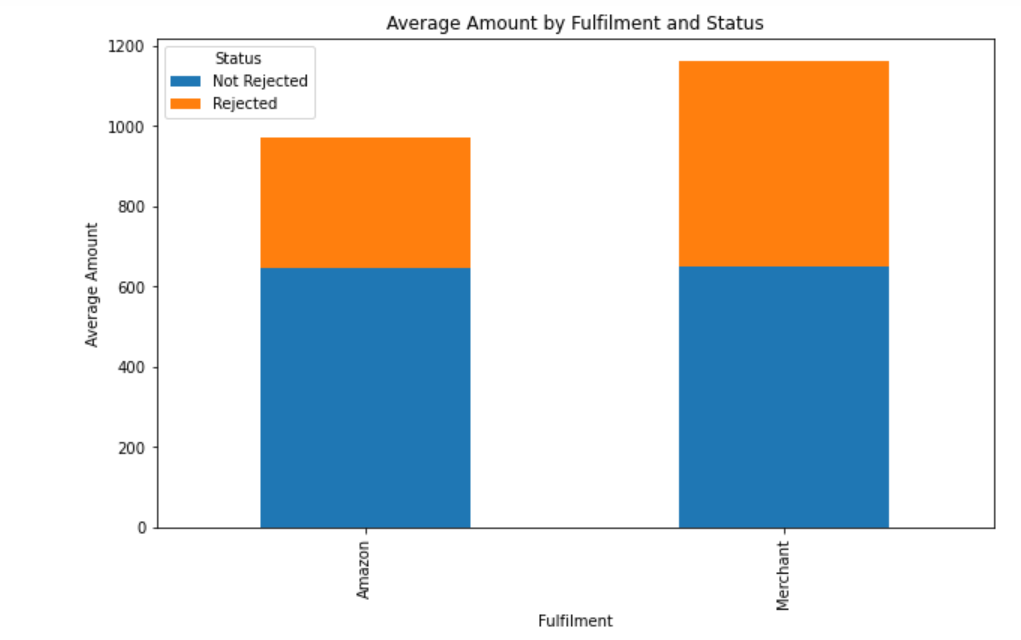
For instance, it might be useful to check how order rejection is affected when distributed by clothing-type/category and the average amount per order for the given clothing categories.



Similar to above, we visualise order rejection proportions vis-à-vis the various order sizes that have been selected by customers.



Similarly, we display how order rejection varies by whether the delivery process is handled by Amazon itself, or by a third party. We can see that the probability of rejection/failure seems to be higher when fulfilment is conducted by third party delivery providers. This might suggest manpower or other issues at this organisation. Certainly something to look into further and understand where the difficulties lie.



**Outlier Treatment**

While we have worked to treat missing values in our dataset, we have taken a different approach when it comes to outliers, given the type of data available and the scope of analysis that we aim to undertake.

For example, when it comes to extreme values in the ‘Amount’ column, it makes more sense to just leave them as is, rather than trying to filter them out or capping the values. This is because one of the areas of interest of us is to understand whether order rejection rates are affected by higher priced products. Later on we shall analyse rejections vs a luxury/non-luxury good perspective and make an inference accordingly.

**Treatment of Imbalanced Data**

Common consensus suggests that a proportion of atleast 30:70 should be present in a binary target variable. However in case of the ‘Status’ column, the share of values within ‘Rejected’ are of a lower proportion.

Similar to our decision on outliers, we have made certain considerations for our target variable, and not attempted to treat imbalanced immediately at the outset.

We have decided to use the data in its original proportion for our base model, and based on its performance will make decisions on whether and how to adjust the imbalance at a later stage.

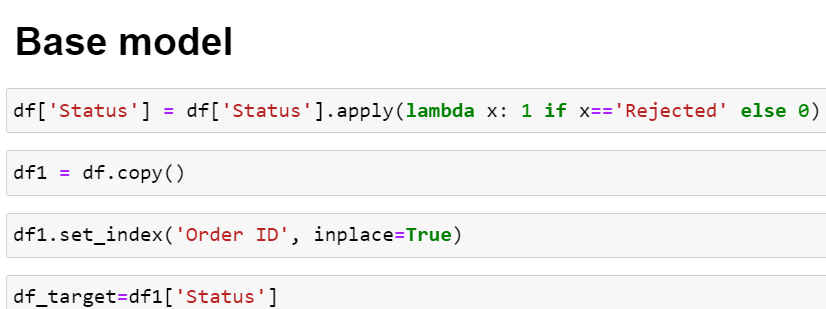
**Model Building and Additional Data Treatments**

**Classification Focussed Predictive Modeling**

As originally mentioned in our objective, our aim with this analysis is to try and understand trends in order rejections. Accordingly, we will build a model that will help identify which of the features are most likely to affect rejections and consequently be useful in predicting rejections in the future.

As a start, we will build a logistic regression based classifier. We shall use popular metrics such as precision and f1-score among other, to interpret the performance of our base model.

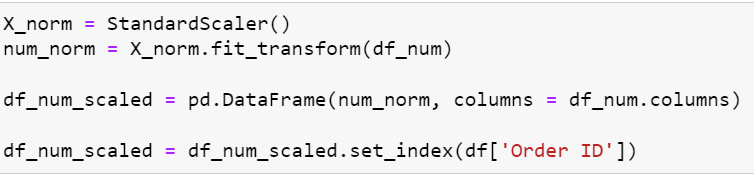
For our logistic model,we shall start with reclassifying our target variable ‘Status’ with values 0 (for ‘Not Rejected’) and 1(for ‘Rejected’).



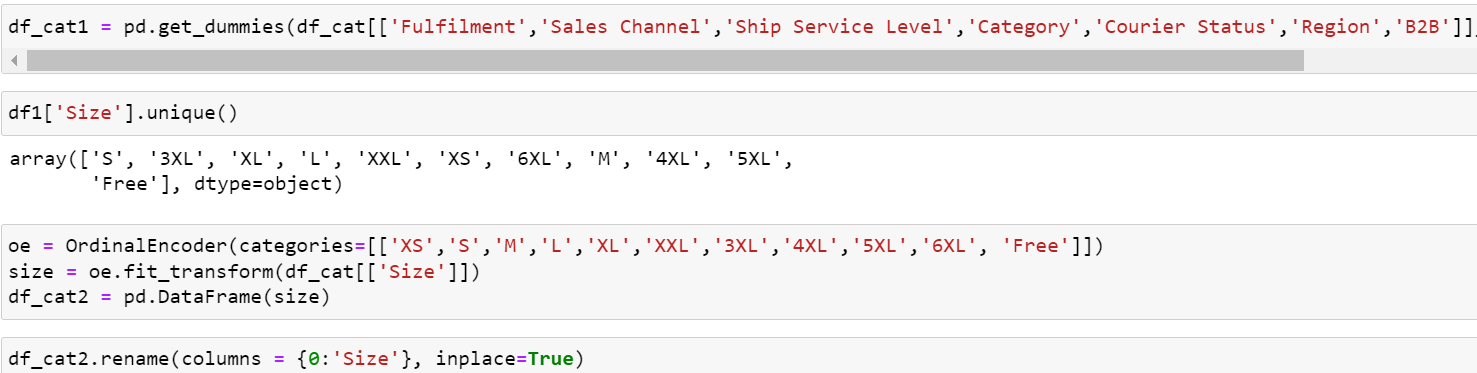
**Feature encoding and scaling**

We shall scale the numeric features of our dataset, and encode the categorical features. Numeric features are scaled to bring them into the same base level; and categorical features are encoded so that their numeric representations can be used to perform further analysis and be fed into out machine learning model.

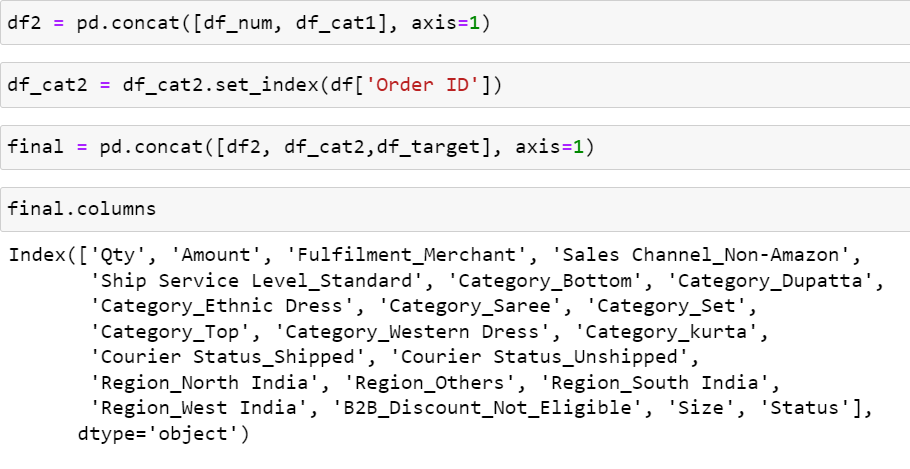
**Scaling the numerical columns using the Standard Scaler function**

****

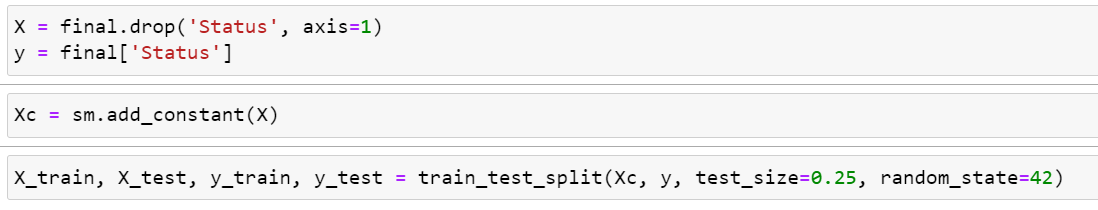
**Encoding the categorical columns using Ordinal Encoding**

****

**Concatenating the scaled numeric and encoded categorical columns to reform our dataset**

****

**Separating our data into target and features sections and using the Train\_Test\_Split function to create appropriate data partitions to be fed into the model**

****

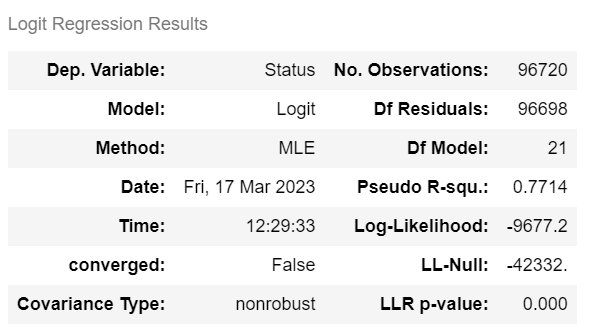
**Feeding the data in to train our model**

****

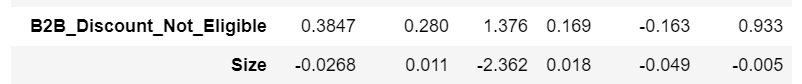
**Analysis and Scoring**

*Now that we have split our model in training and testing segments, and trained our model using the training segment, we shall use various methods to check performance.*

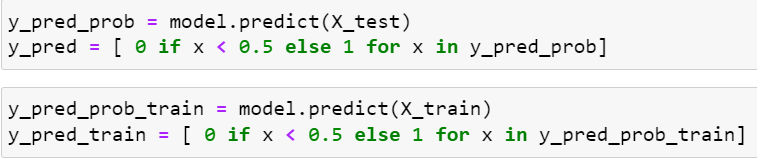
**First, we use the model.summary() function to output a table of various results from our Logistic Regression Classifier.**



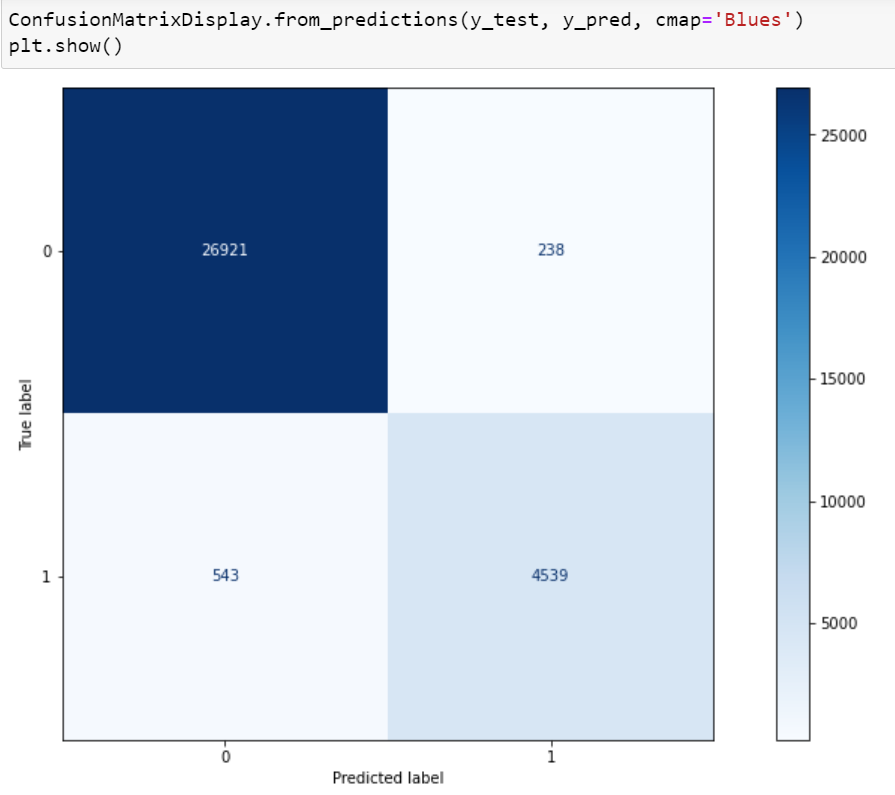


****

**Next, given the classification nature of this model, we shall use a standard cut-off of 0.5 to put predictions into the ‘Rejected’/’Not Rejected’ buckets**

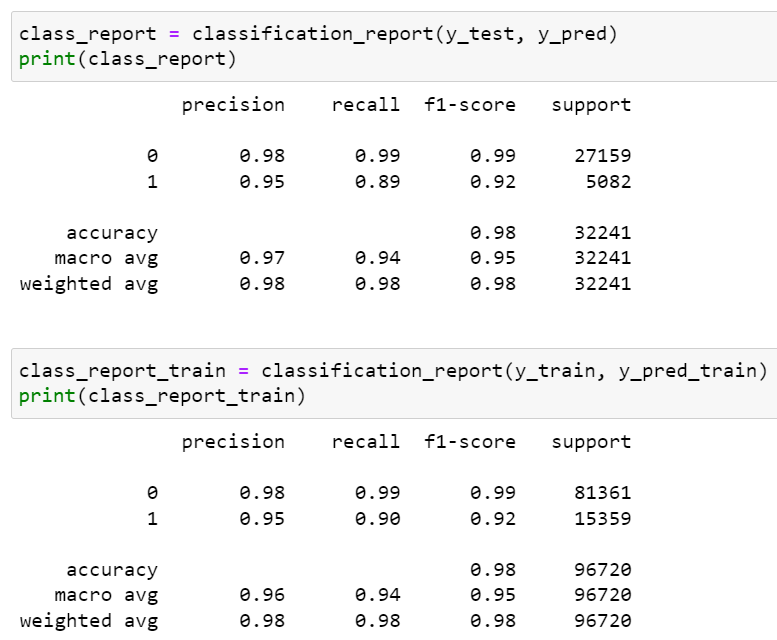


**As part of performance scoring, let us look at the confusion matrix, classification report and roc\_auc score.**

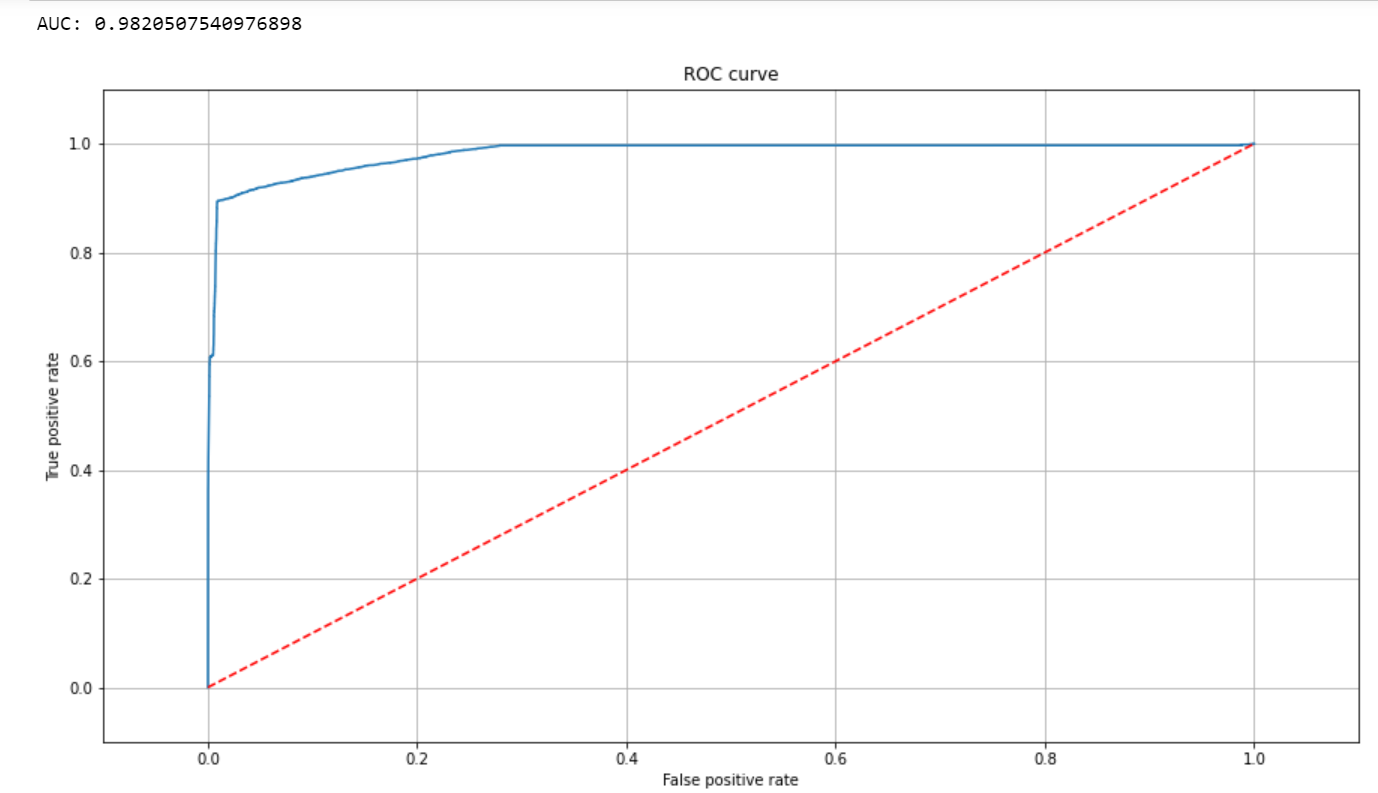


Below we can see that class wise and overall accuracy, precision and f1-score of the model on the training and testing set is broadly similar. However the model’s ability to accurately classify class 1 (‘Rejected’) values seems weaker.

**Our base model’s relative weakness in accurately predicting when order rejections might occur are a key reason for considering improvement techniques and alternative classifiers.**

****

**Finally, our model receives an AUC score of .9820. Given the range of AUC from 0 to 1, on this metric, we can conclude that our model has good performance.**

****

**5. Future Actions**

An important point to note at this stage is that while we now have various methods and related scores to judge our model, it would only be appropriate to do so on a relative basis, compared to how other supervised models perform on the given data.

Thus as next steps we just feed our dataset and build models using the following mentioned classifiers:

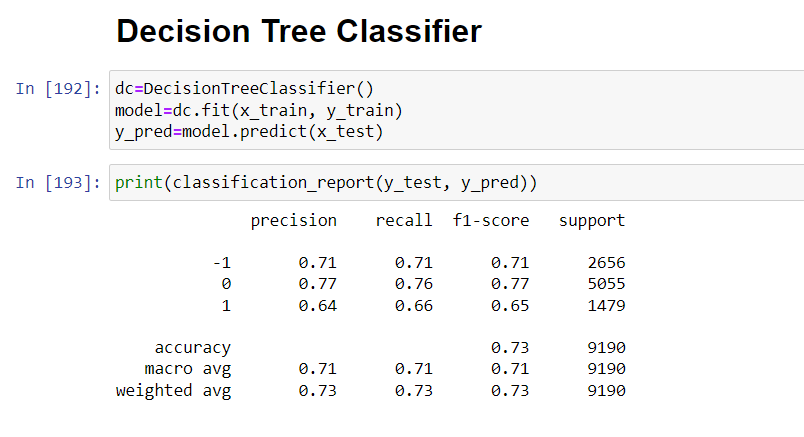
* Decision Tree
* Random Forest
* Naïve Bayes
* KNN

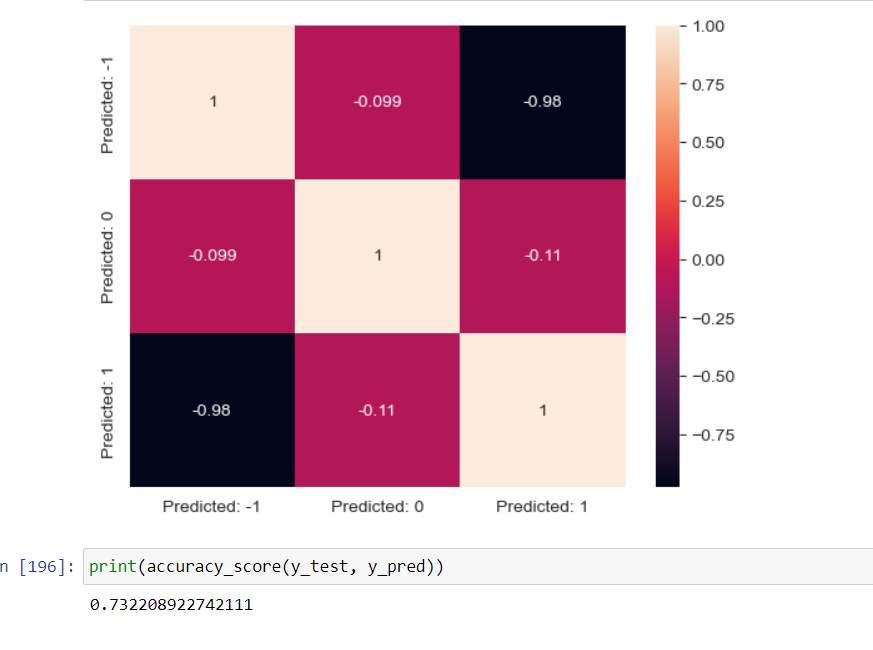
Additionally, tuning our hyperparameters for each model will be an important tool in helping improve the performance of the classifiers mentioned above.

Finally, based on the results achieved, in case there are indications of certain models being underfit, we may employ suitable techniques such as Adaboost, GradientBoost, and XGBoost. Similarly for overfit situations, we may employ methods such as regularisation, and potentially dimensionality techniques such as GLRM, and PCA more specifically.

# Decision Tree

considering decision tree method we could achieve a model with 71% accuracy





**Accuracy score is 73%**