Home Credit Loan Default Prediction Project

Introduction

Problem Definition:

- This project is about accurately predicting the loan repayment abilities of customers (In this case they are people with insufficient or non- existent credit histories) to ensure that the clients capable of repayment are not rejected.

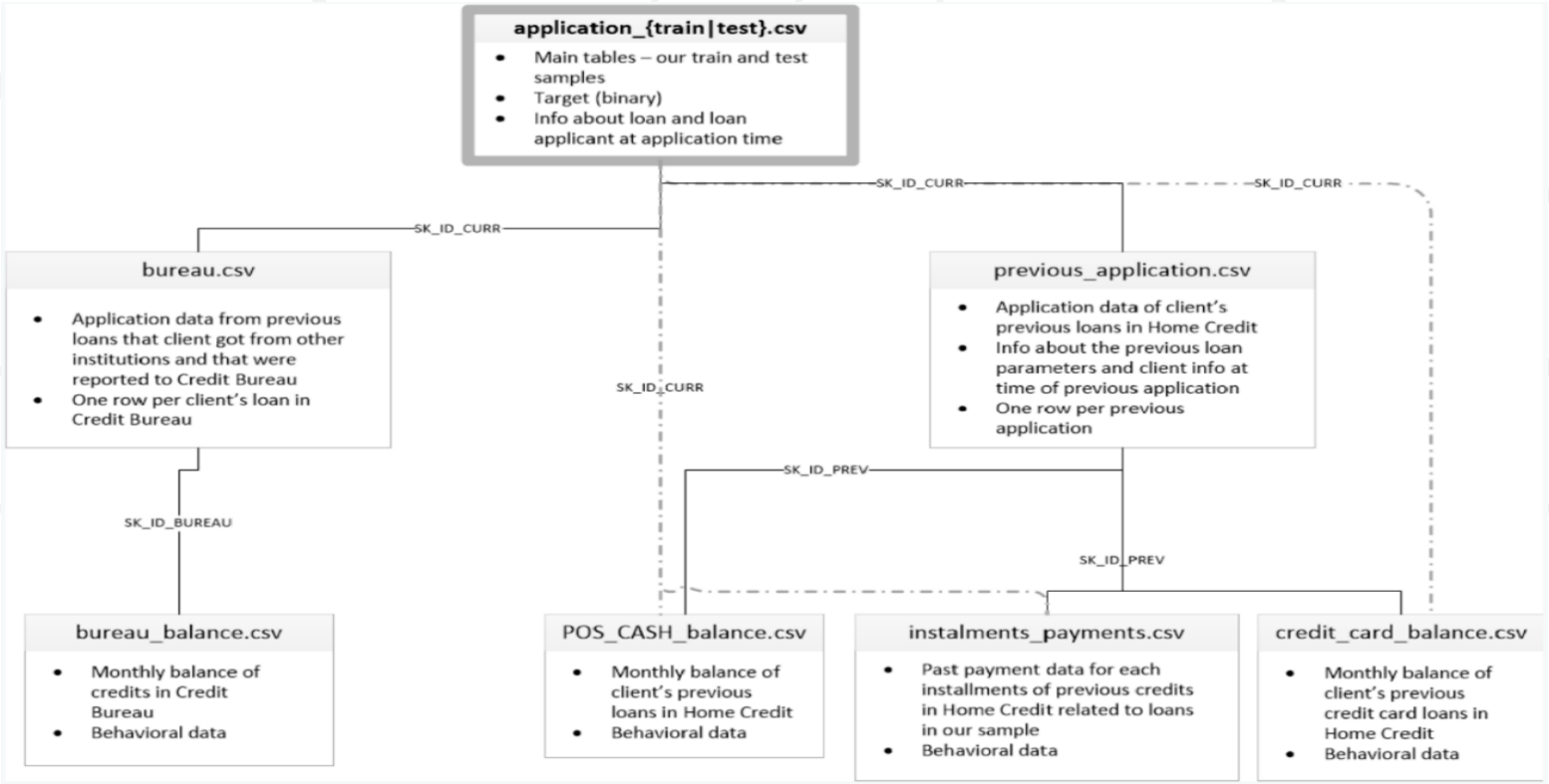
Business Client and Objective:

- My client is Home Credit, an international consumer finance provider with operations in 9 countries. They focus on responsible lending primarily to people with little or no credit history.

- An optimum solution will ensure that home credit clients capable of repayment are not rejected and at the same time will also ensure minimum loss to home credit because of giving loans to future defaulters.

Data Overview

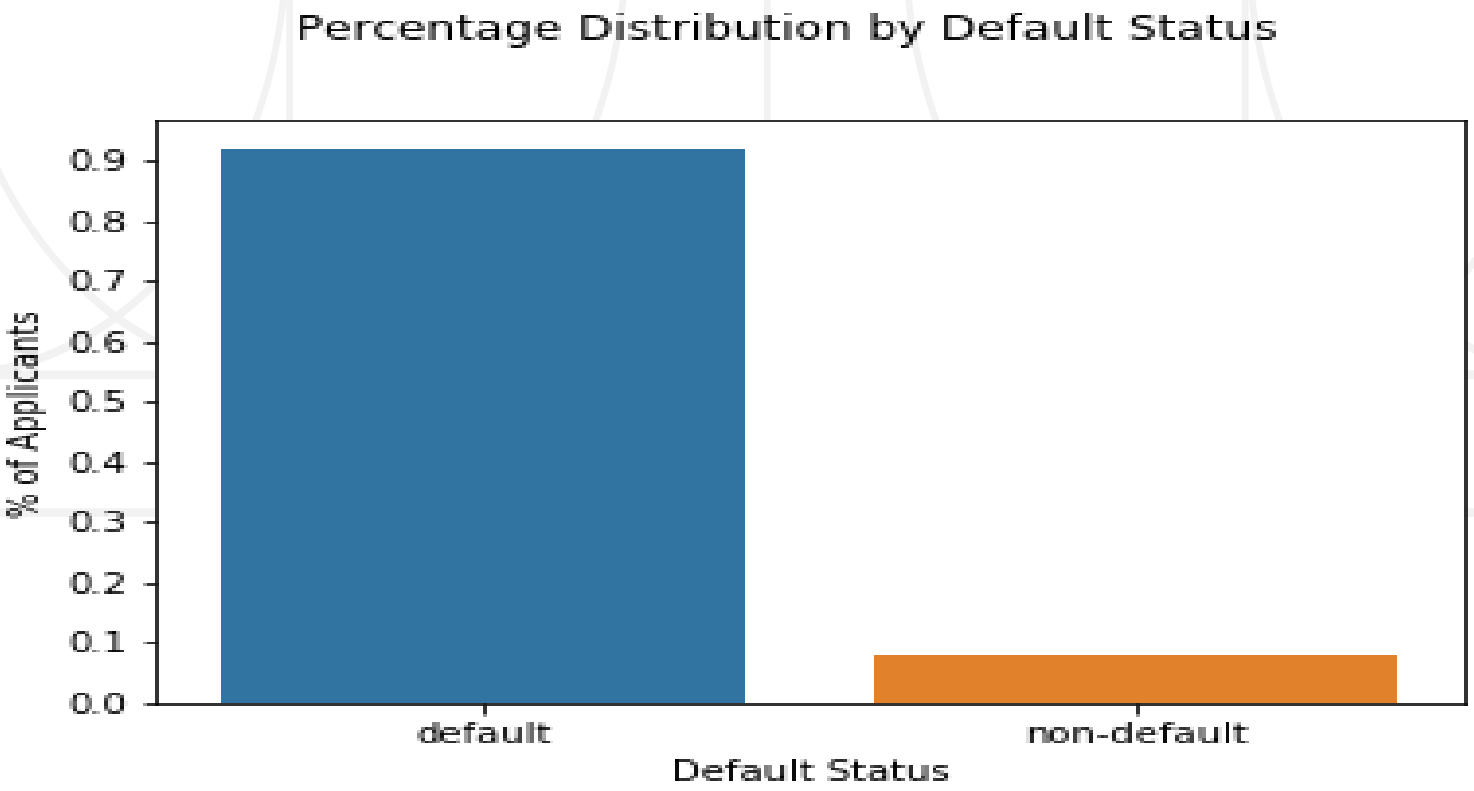
- **Data Description**: There are **seven** main files that are available as part of this problem and I have tried using information from all these files as part of the solution.
- A combined data set was prepared after feature engineering and using data from all the above files.
- The combined data set was used for EDA and as the input to the Machine Learning models.



Exploratory Data Analysis : Distribution of Defaults vs Non-Defaults

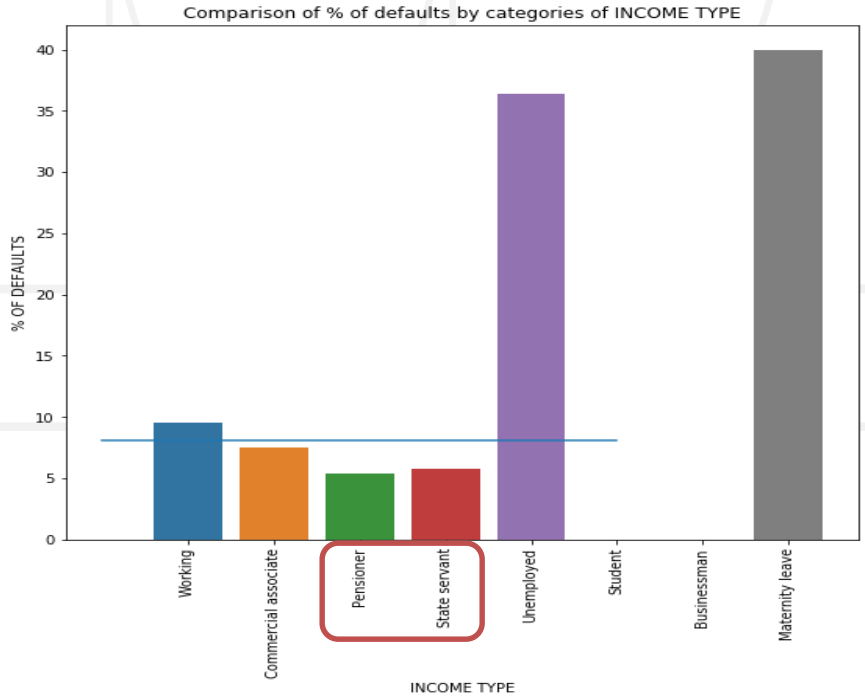
TARGET	
default	8.07
non-default	91.93

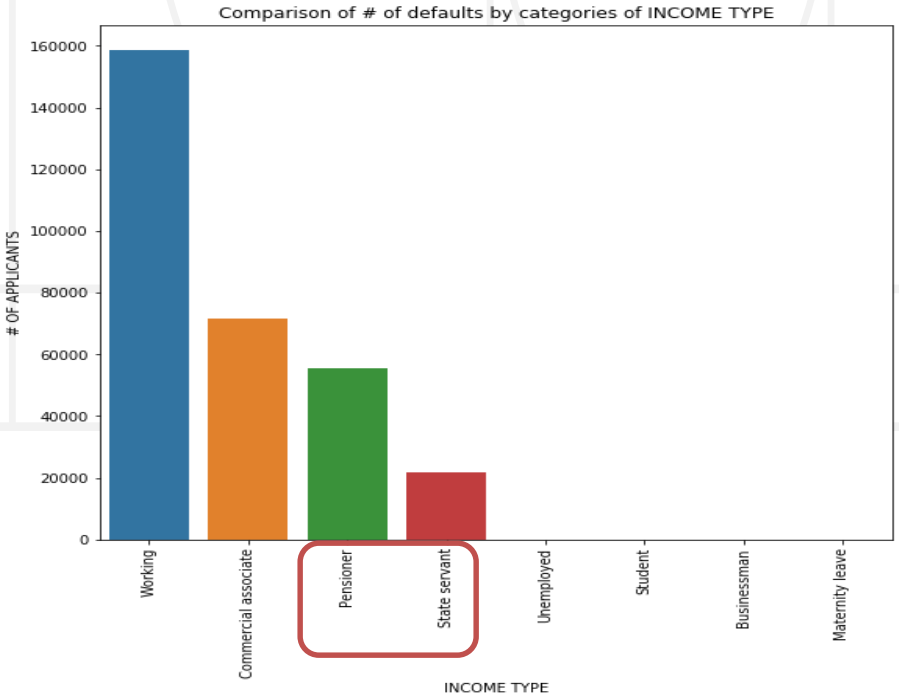
- The below graph shows the distribution of defaults vs non-defaults in our data set. We can see that the **rate of default in the data set** is **8.07%.**
- In the upcoming slides we are going to visualize some of the categorical features and investigate whether certain groups within that feature could be an indicator of a group which is relatively riskier or safer to provide loan/credit for home credit.



Exploratory Data Analysis Rate Of Default Comparison by INCOME TYPE

	NAME_INCOME_TYPE	# OF APPLICANTS	# OF DEFAULTS	% OF DEFAULTS
0	Working	158774	15224	9.588
1	Commercial associate	71617	5360	7.484
2	Pensioner	55362	2982	5.386
3	State servant	21703	1249	5.755
4	Unemployed	22	8	36.364
5	Student	18	0	0.000
6	Businessman	10	0	0.000
7	Maternity leave	5	2	40.000



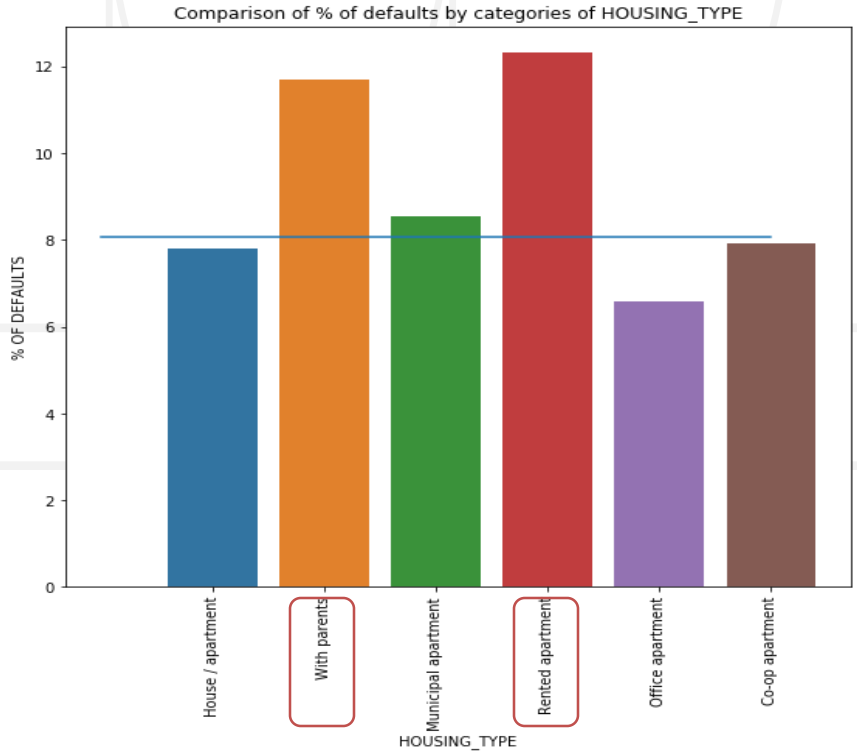


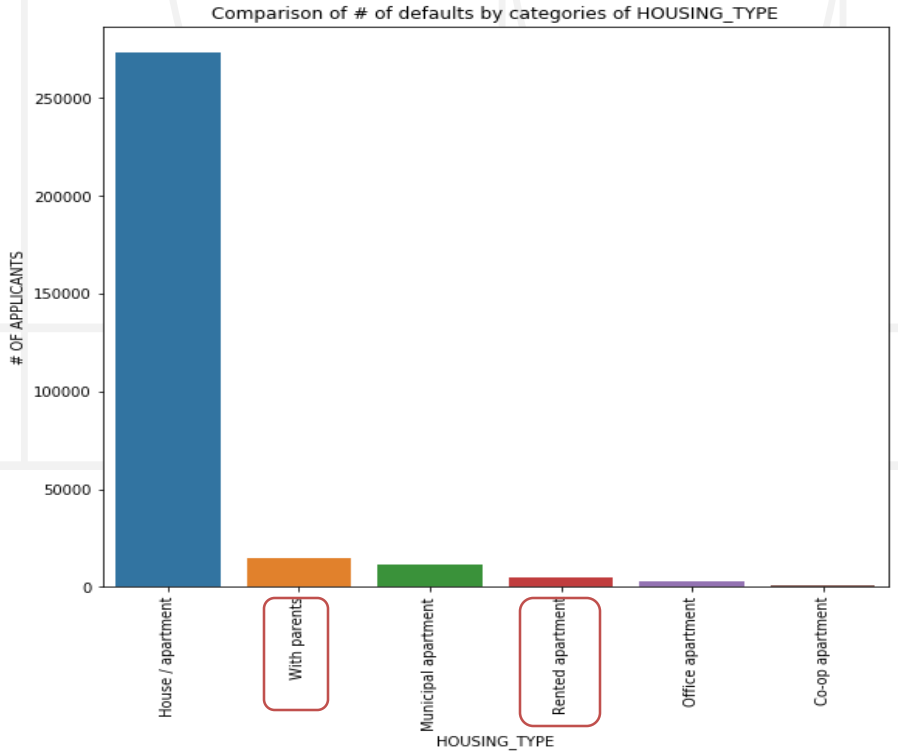
- The default % by **INCOME TYPE** indicates that the rate of default among state servants (**5.75%**) and pensioners(**5.38%)** is lower than the average default % of **8.07**%.
- As the total # of applicants in the state servant (**21703**) and pensioner (**55362**) categories is substantial, the lower rate of default by **2.5%** from the overall default average rate indicates that people from these categories are a safer bet to provide loan/credit.

Exploratory Data Analysis Rate Of Default Comparison by HOUSING TYPE

	NAME_HOUSING_TYPE	# OF APPLICANTS	# OF DEFAULTS	% OF DEFAULTS
0	House / apartment	272868	21272	7.796
1	With parents	14840	1736	11.698
2	Municipal apartment	11183	955	8.540
3	Rented apartment	4881	601	12.313
4	Office apartment	2617	172	6.572
5	Co-op apartment	1122	89	7.932

- The default % by **HOUSING TYPE shows** that the **rate of default** among applicants **staying with parents** (**11.69%**) and those **staying in rented apartments** (**12.31%)** is **higher than** the average default % of **8.07**%.
- As **the total # of applicants** staying with parents(**14840**) and those staying in rented apartments (**4881**) is sizeable, the higher rate of default by around **3.5%** from the overall default average rate indicates that people from these categories are riskier to provide loan/credit.

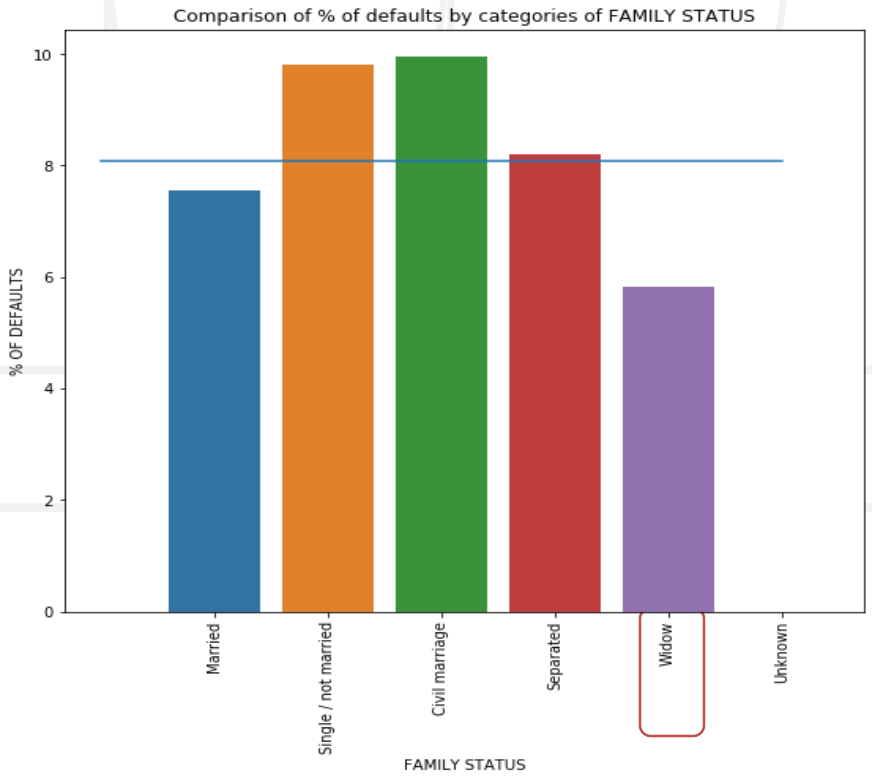


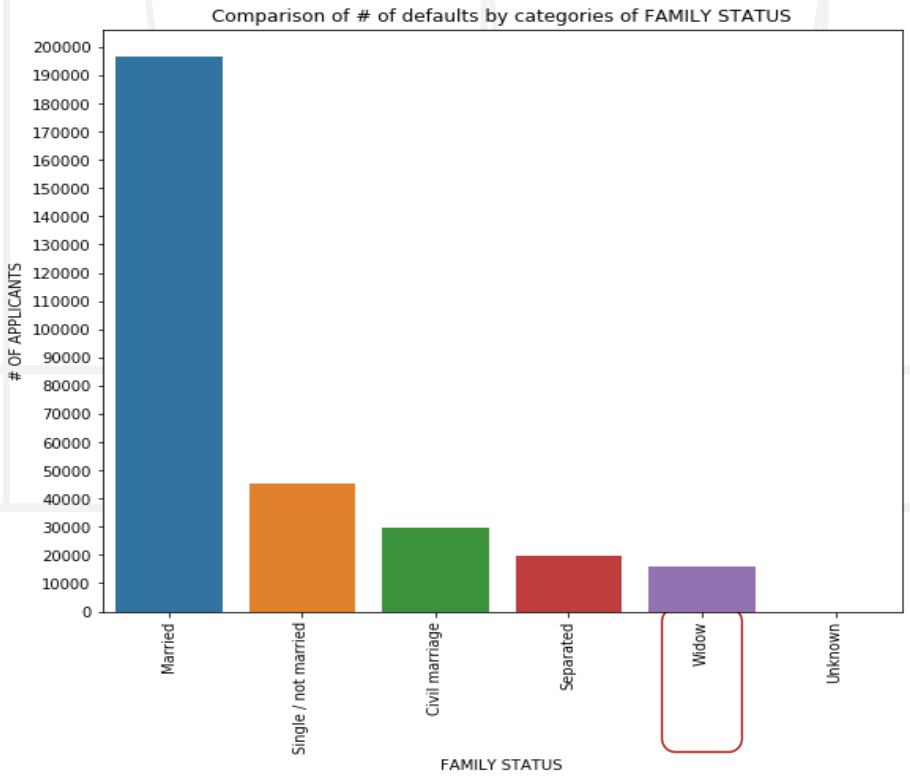


Exploratory Data Analysis Rate Of Default Comparison by FAMILY STATUS

	NAME_FAMILY_STATUS	# OF APPLICANTS	# OF DEFAULTS	% OF DEFAULTS
0	Married	196432	14850	7.560
1	Single / not married	45444	4457	9.808
2	Civil marriage	29775	2961	9.945
3	Separated	19770	1620	8.194
4	Widow	16088	937	5.824
5	Unknown	2	0	0.000

- The default % by **FAMILY STATUS shows** that the rate of default among widows (**5.82%**) is lower than the average default rate of **8.07**%.
- The total # of applicants who are widow(**16808**) is also sizeable, hence the lower rate of default by around **3.3%** from the overall default average rate indicates that **people from this category are safer to provide loan/credit to**.

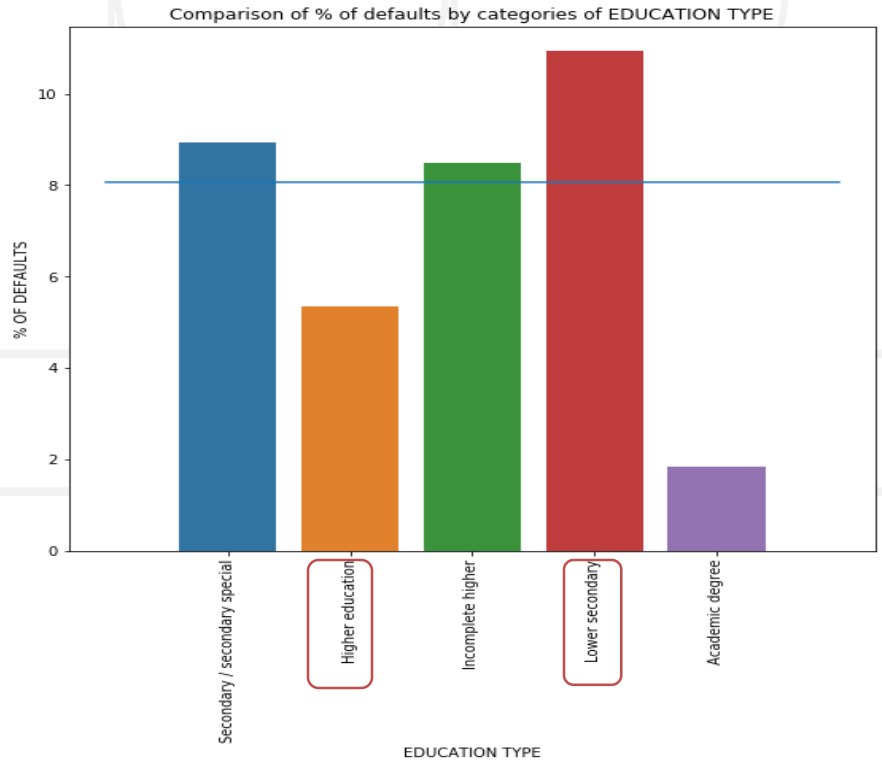


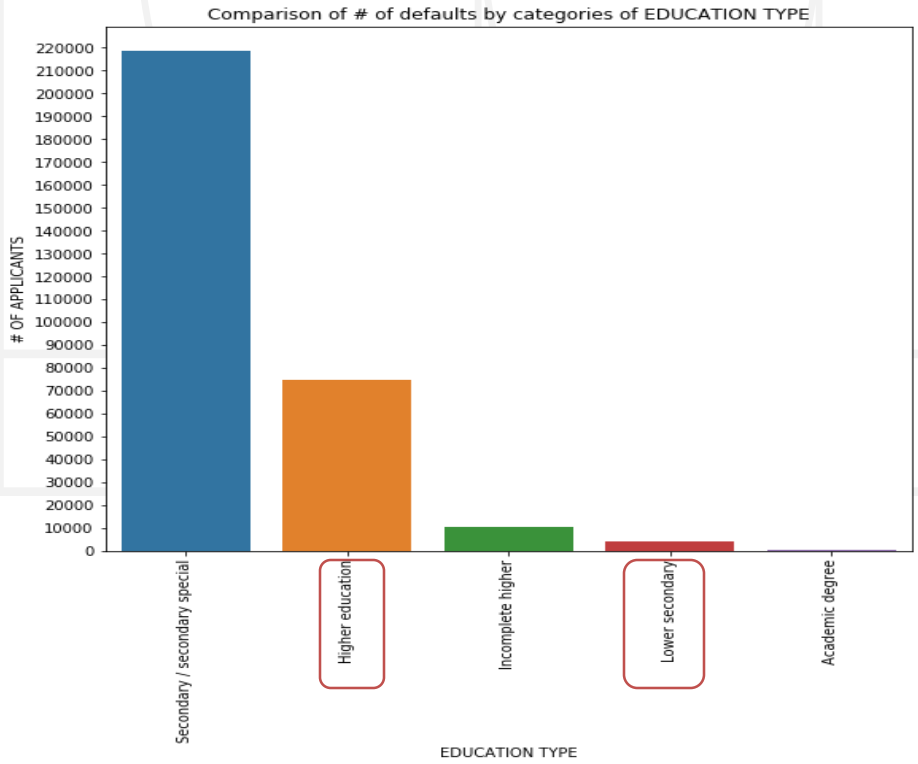


Exploratory Data Analysis
Rate Of Default Comparison by EDUCATION TYPE

	NAME_EDUCATION_TYPE	# OF APPLICANTS	# OF DEFAULTS	% OF DEFAULTS
0	Secondary / secondary special	218391	19524	8.940
1	Higher education	74863	4009	5.355
2	Incomplete higher	10277	872	8.485
3	Lower secondary	3816	417	10.928
4	Academic degree	164	3	1.829

- The default % by **EDUCATION TYPE shows** that the rate of default among applicants **having higher education** (**5.35%**) is **lower than** the **average default rate** of **8.07**%, whereas among those **having lower secondary education**(**10.9**%) is **higher than the average**.
- As the **total # of applicants** of both these categories is **sizeable** we can say that the **applicants with higher education are safer** where as those with **lower secondary education are riskier** population to give credit/loan to.

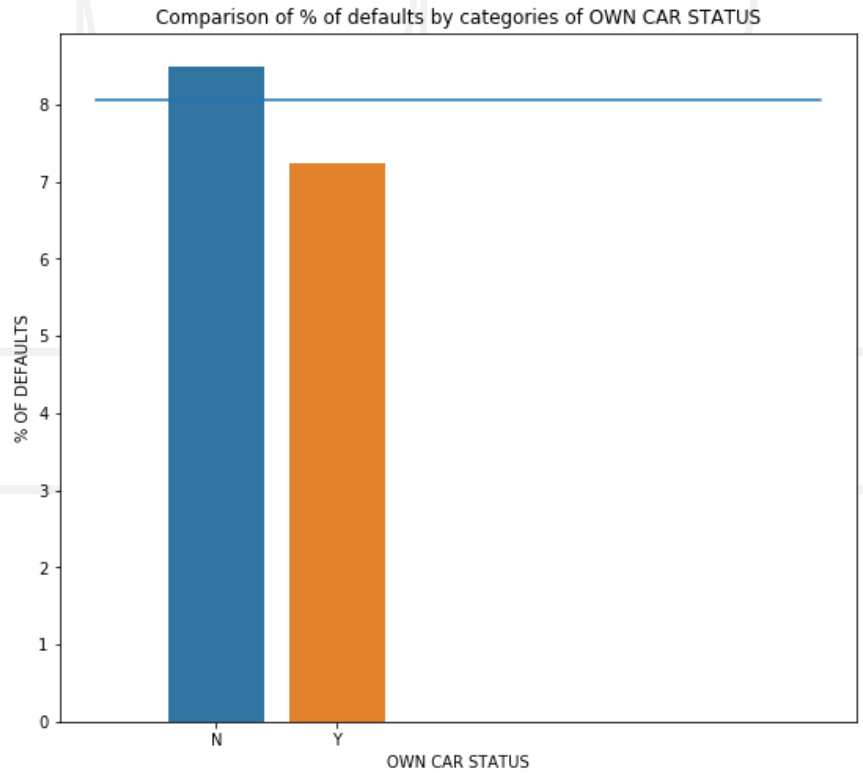


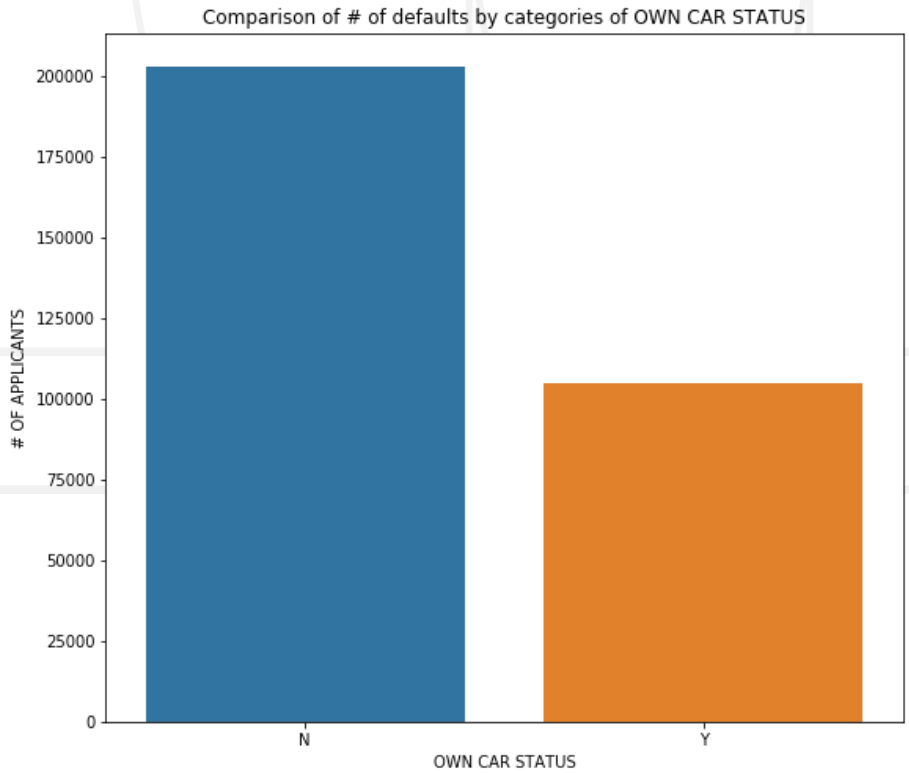


Exploratory Data Analysis Rate Of Default Comparison by OWN CAR STATUS

	FLAG_OWN_CAR	# OF APPLICANTS	# OF DEFAULTS	% OF DEFAULTS
0	N	202924	17249	8.500
1	Y	104587	7576	7.244

- The default % by **OWN CAR STATUS shows** that the rate of default among applicants **having own car** (**7.24%**) is **slightly lower** than the **average default rate** of **8.07**%, whereas **those not having their own car** (**8.5%**) has a default rate **slightly higher** than the average.
- But, **as the difference of rates from the average rate for both the groups is low(0.5% – 1.0%)**, apparently, **this feature can't conclusively determine the safer/riskier group to give a credit/loan to.**

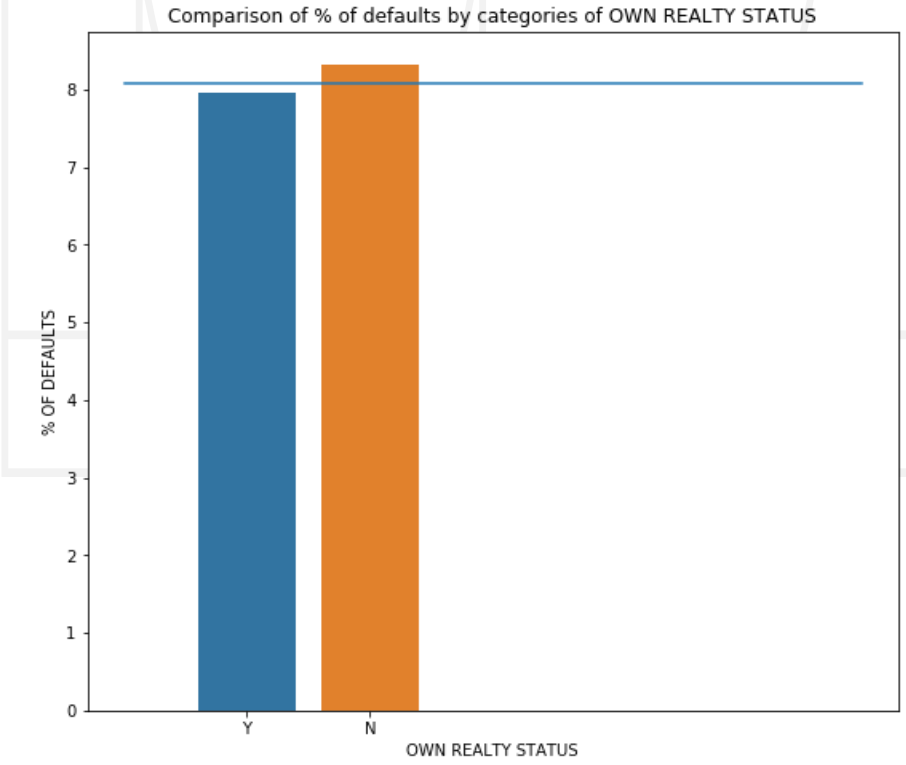


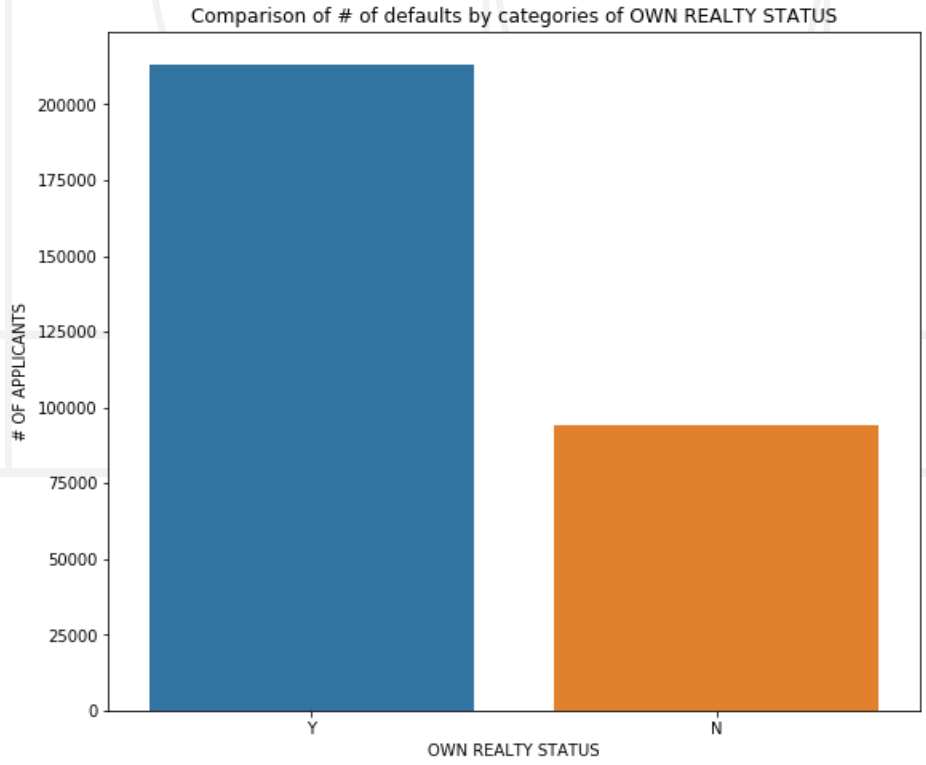


Exploratory Data Analysis
Rate Of Default Comparison by OWN REALTY STATUS

	FLAG_OWN_REALTY	# OF APPLICANTS	# OF NON-DEFAULTS	% OF NONDEFAULTS	# OF DEFAULTS	% OF DEFAULTS
0	Y	213312	196329	92.038	16983	7.962
1	N	94199	86357	91.675	7842	8.325

- The default % by **OWN REALTY STATUS shows** that the rate of default among applicants **having own realty** (**7.96%**) is **slightly lower** than the average default rate of **8.07**%, whereas those not **having their own realty** (**8.32%**) has a default rate **slightly higher** than the average.
- But, as the **difference of rates from the average rate for both the groups is low(0.1% – 0.25%)**, apparently, this **feature can't conclusively determine** the safer/riskier group to give a credit/loan to.



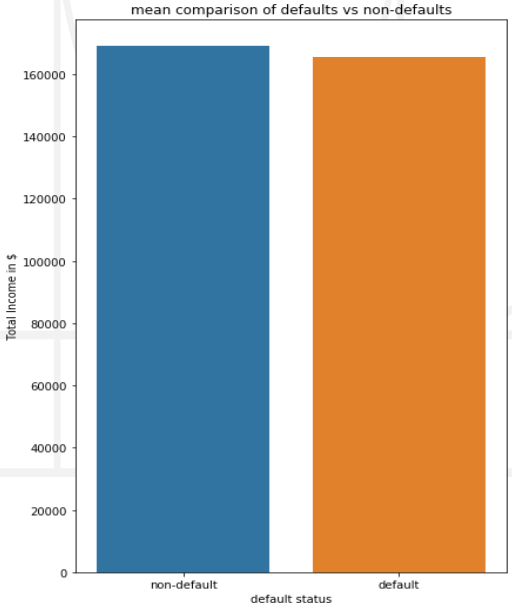


Exploratory Data Analysis : Mean Total Income - Defaults Vs Non-Defaults

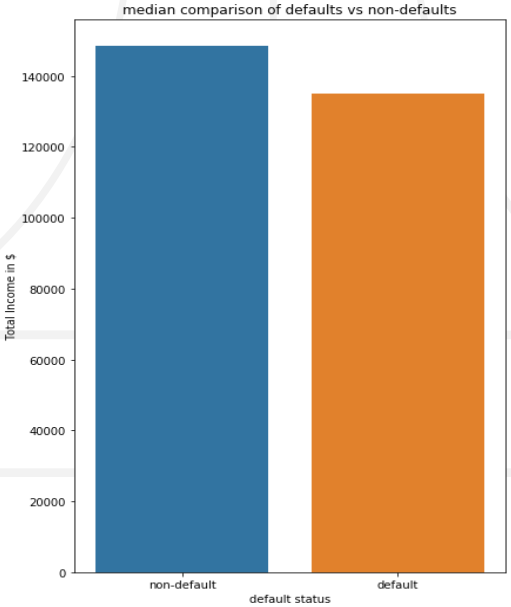
- In this and couple of upcoming slides we are going to compare the means of certain features **with a goal of finding out whether these features are good contenders for differentiating applicants who default from those who don't**.

- The mean total income amount for the non-default group is **169077.722** whereas the total income amount for the default group is **165611.76**.

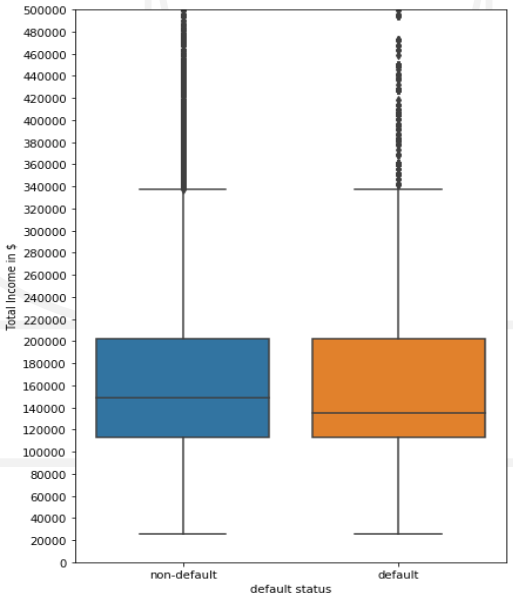
- In conclusion, the difference doesn't appear to be substantial and consequently, **total income amount somewhat surprisingly doesn't appear to be a good indicator to differentiate the default and non-default groups.**



Mean



Median

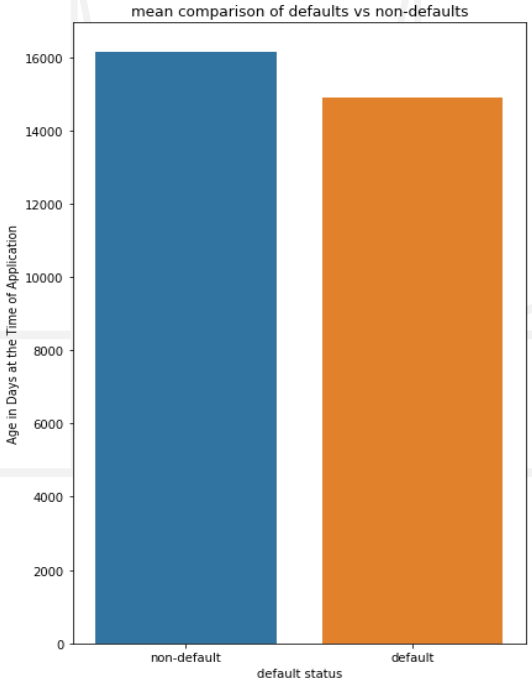


Box Plot

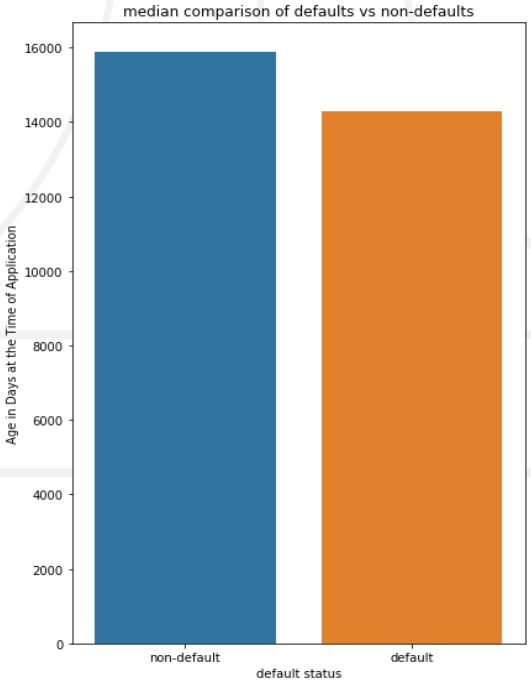
Exploratory Data Analysis : Mean Age - Defaults Vs Non-Defaults

- The mean age of the non-default group is **16138 days (around 44 years)** whereas the mean age of the default group is **14884 days (around 40 years)**. This means that on average the **people who don't default are elder by 1254 days (around 3.5 years) than the people who default.**

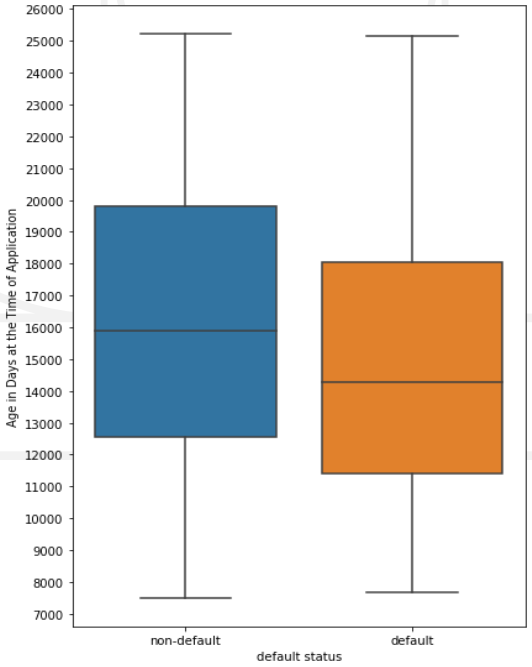
- All these findings indicate that on average the **people who default are younger than the people who don't default**.
- There is also substantial difference in age of the two groups in their 50th (median) and 75th percentiles respectively.
- In Summary, **the age appears to be a good indicator to differentiate a default from non-default.**



Mean



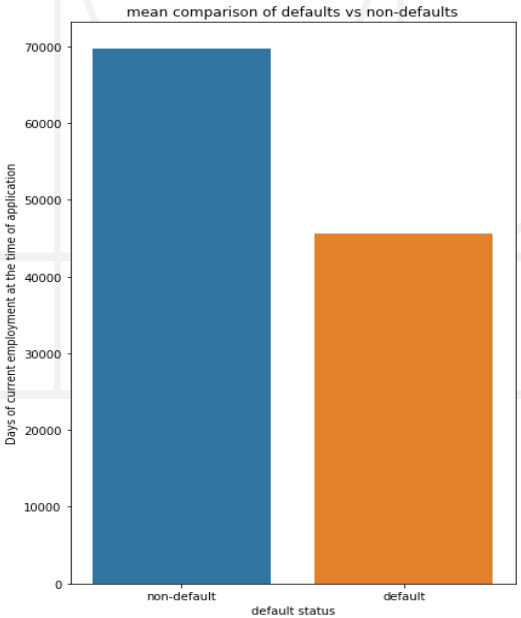
Median



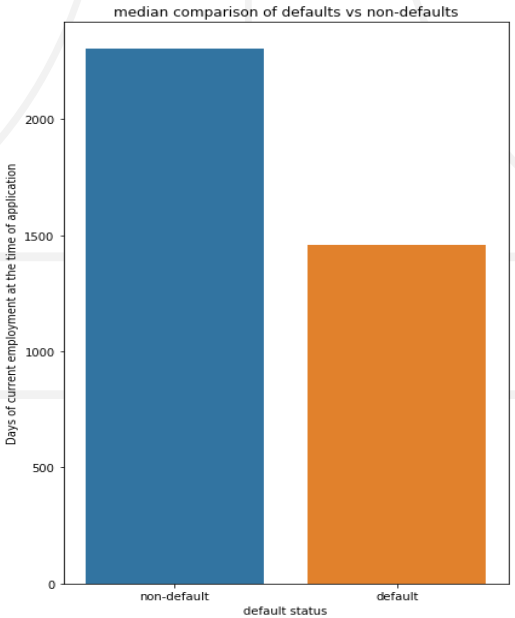
Box Plot

Exploratory Data Analysis : Employment Duration - Defaults Vs Non-Defaults

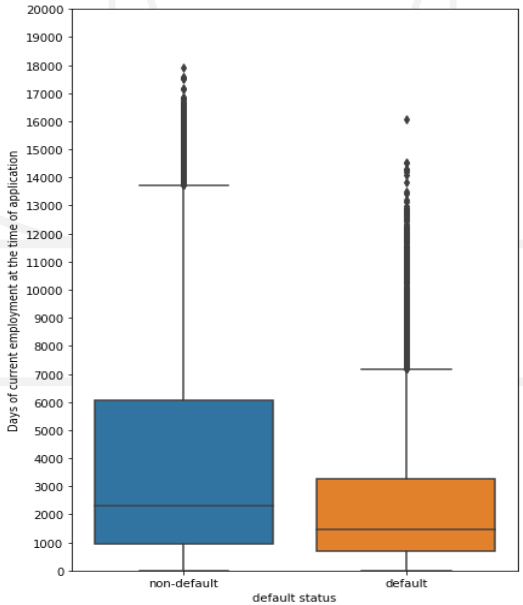
- The mean **employment duration for the non-default group is 69668 days** whereas the **employment duration for the default group is 45587 days**.
- There is also **substantial difference** in **employment duration** of the two groups in their **50th (median) and 75th percentiles** respectively.
- The **75th percentile of the non-default group is 6074 days** whereas for the **default group it is 3280 days.**
- **In summary**, all these findings indicate that there is a **substantial difference in employment duration between the default and non-default groups** and thus the employment duration appears to be a **good indicator to differentiate a default from non-default**.



Mean



Median



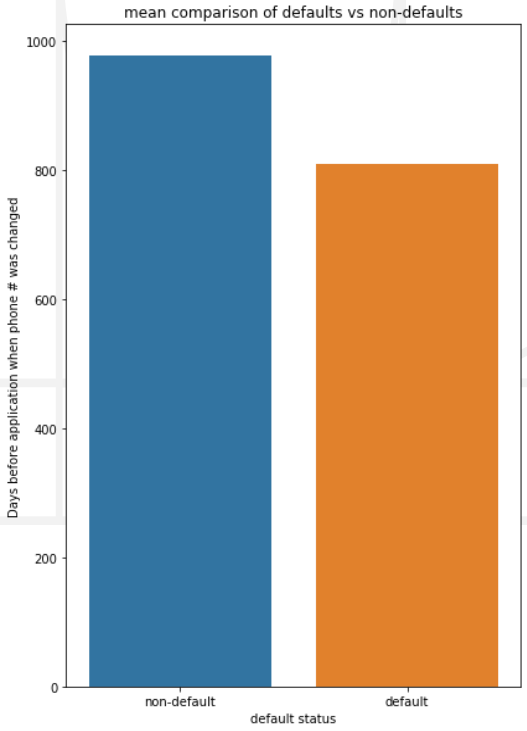
Box Plot

Exploratory Data Analysis : Mean Days Before The Last Phone Number Change/Update - Defaults Vs Non-Defaults

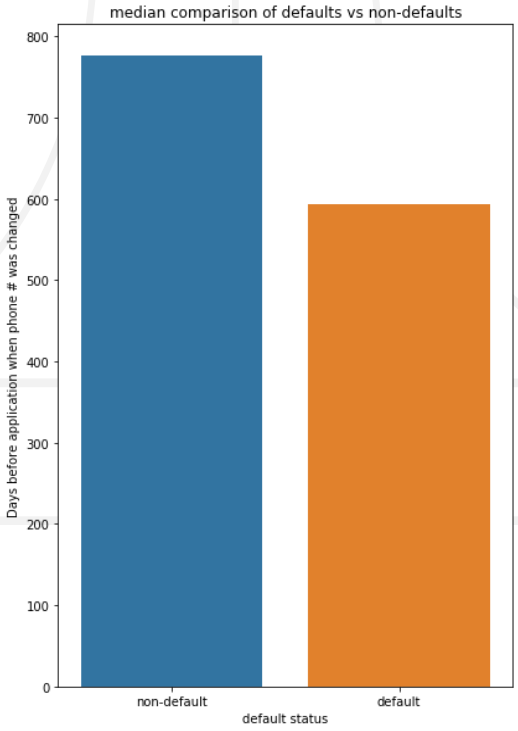
- The mean of **days since last phone number change** for the **non-default group** is **976 days** whereas that for the **default group** is **808 days**.

- There is also **substantial difference** in days since last phone number change of the **two groups** in their **50th (median) and 75th percentiles** respectively.

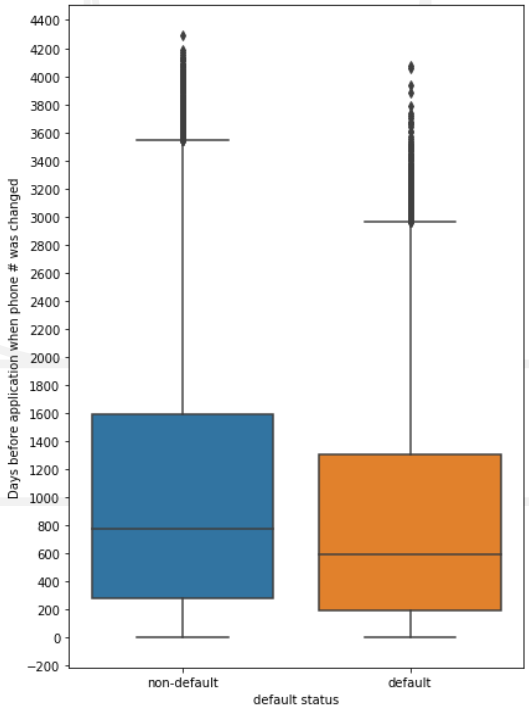
- **In conclusion**, the findings from the bar graph indicate that **on average the days since last phone number change for the people who defaulted is earlier than that of the people who didn't default**.



Mean



Median



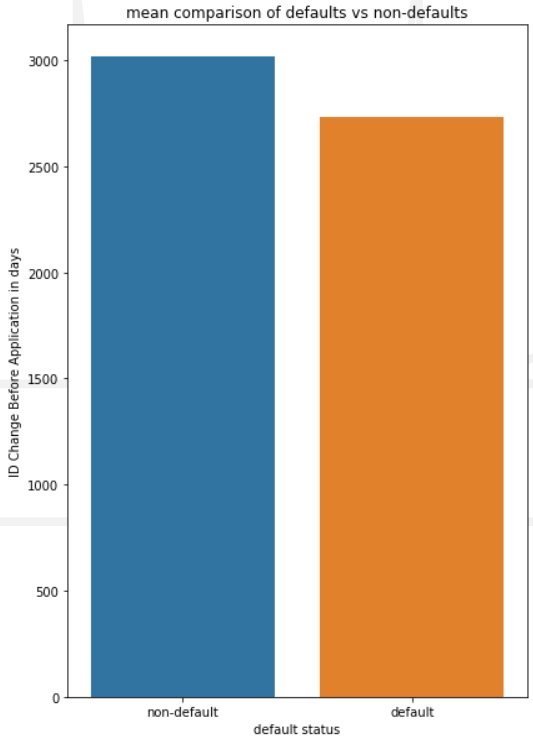
Box Plot

Exploratory Data Analysis : Mean Days Before The Last ID Change/Update - Defaults Vs Non-Defaults

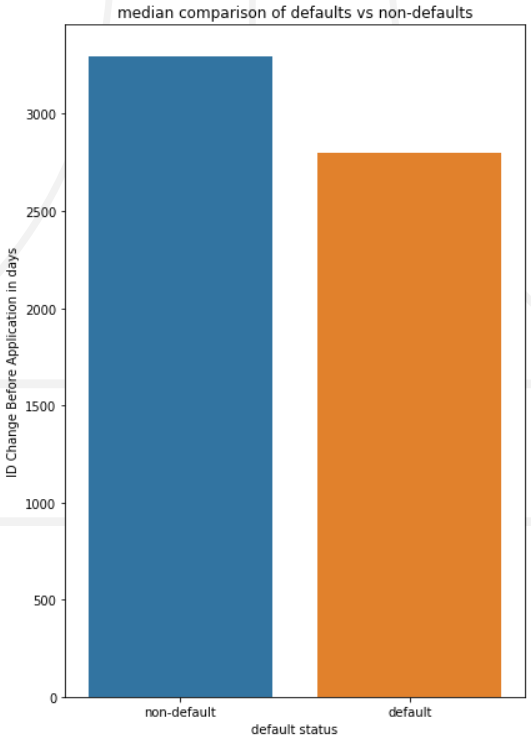
- The mean of **days before the id change/update** (id with which the client applied for the loan) for the **non-default group is 3017 days** whereas for the **default group the value is 2732 days**.

- There is also **substantial difference** in days before the id change/update of the **two groups** in their **50th (median) and 75th percentiles** respectively.

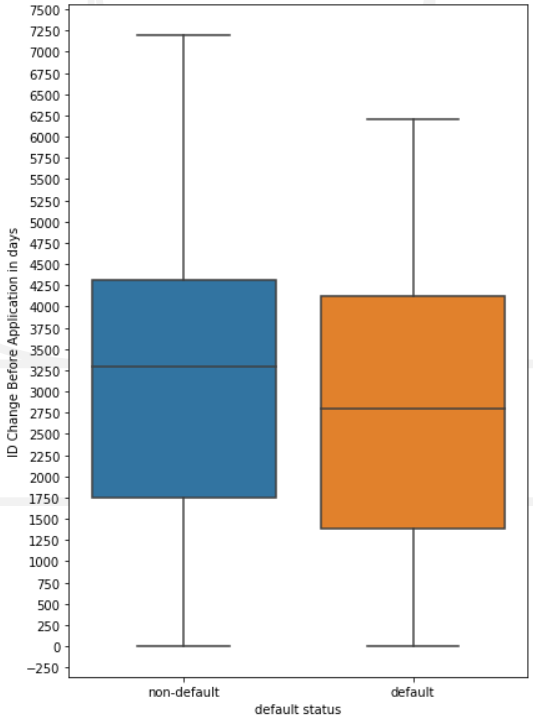
- **In conclusion**, the findings from the bar graph indicate that on average **the days before the id change/update for the people who defaulted is earlier than that of the people who didn't default**.



Mean



Median



Box Plot

Exploratory Data Analysis : Conclusion

- It is **relatively safer to provide loan/credit to** people who are **pensioners or state servants** when compared to other income groups.
- It is **riskier to provide loan/credit to people staying with parents or those staying in rented apartments** when compared to people staying in other housing types.
- It is **safer to provide loans to people with higher education**, **but riskier to people with lower secondary education** when compared to people with other types of education.
- It is safer to **provide loan to widows** than people with other family status.
- The data visualization of total income reveals that **income doesn't appear to be a good differentiator of defaults from non-defaults**.
- But the variables age, employment duration, days since last phone number change and days before the id change/update **shows a pattern that the non-defaults on average appear to be elder in age and are more stable in their behavior in changing employment, changing phone numbers and changing/updating ids when compared to defaults.**
- **This relative stability also reflects in their behavior in paying back the loan without defaulting**.

Machine Learning

Broadly below are the 8 steps used in building the machine learning pipeline

- **Building of single source of data by combining and feature engineering the 7 available data sources.**
- **Selecting a set of most important features to be used for model training**
- **Choosing a set of models to be tried on the problem**
- **Choosing a sample that is used for hyperparameter tuning for contender models using grid search**
- **Splitting the combined data set into training and test data sets.**
- **Using the sample data set in deciding the best hyperparameters for each model and then training each model on the training data set using the set of most important features and these hyperparameters.**
- **Evaluating the performance of each model on the test data set using the evaluation metric (AUC-ROC score).**
- **Finalizing the best model depending on the above score and explaining how the model can be possibly used in the business context.**

Machine Learning : Feature Engineering and Selection

- **Building of single source of data by combining and feature engineering the 7 available data sources.**
    - The details of the steps performed for building the single data source can be accessed here Data Wrangling Doc.

- **Selecting a set of most important features to be used for model training**
    - Feature selection is performed to look at the possibility whether using a limited set of features (the most important ones) can give almost the same performance when compared to the scenario in which we use all the features. If there is such possibility then it will be desirable as this will let the model be faster and use less computing resources.
    - The reason for which Random Forest was chosen as the model for the feature selection is that while determining feature importance it takes into consideration the effect of all the features at the same time rather than considering them individually. Not only this, it also automatically considers the interaction between feature variables.

Machine Learning : Choice of Model

- **Points to Highlight before proceeding to choose the model**
  - **Time constraint (Latency)** : As the problem definition doesn't mention any time constraint on time taken to classify an applicant who is likely to default, I have assumed that choice of model is flexible in terms of time taken by the model to predict the output
  - **Model Interpretability** : I assume in this project that the prediction power of the model is lot more important than the model interpretability in terms of features. The significance of different features were covered in detail as part of the EDA.
  - **Evaluation Metric** : The **ROC-AUC score** is chosen as the metric for evaluating the performance of the classification model as this metric checks the error rate for each threshold in contrast to misclassification rate which checks the same for only one threshold (0.5). The Business would most probably like to see the performance of the model at different thresholds so that they can decide their optimum loan providing strategy.
- **Choosing a sample used for hyperparameter tuning for contender models using grid search**
  - This step was taken since running the GridseachCV function to find the best hypermeter values is a heavily CPU/ resource intensive task and hence using a sample helps in reducing the running time of this function without compromising much on the result.

Machine Learning : Choice of Model

- Depending on the points discussed in the previous slide I tried the below models from different families.

  Parametric Models : Logistic Regression

  Non-Parametric Models (Bagging) : Random Forest

  Non-Parametric Models (Boosting) : XGBoost and Light GBM

- In the upcoming slides the result in terms of the confusion Matrix and the ROC-AUC score has been discussed in detail

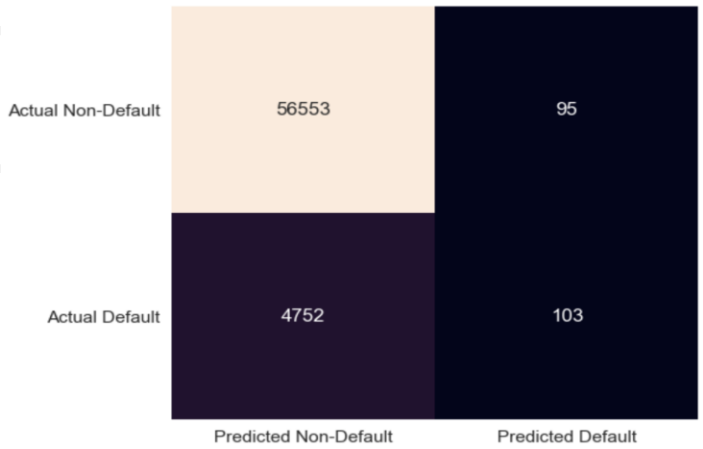
Machine Learning : Model Training

- **Splitting the combined data set into training and test data sets**
  - The training data is used to train the model on the seen data and the test data will be used to test the performance of the trained model on the unseen data.
- **Using the sample data set in deciding the best hyperparameter values for each model and training each model on the training data using the set of most important features and hyperparameter values.**
  - Once the best hyperparameter values are determined and the data is split into train and test set, each model was trained on the training data by selecting different number of most important features (e.g., top 100, top 200 features, all the 364 features etc.) along with using the chosen hyperparameter values.

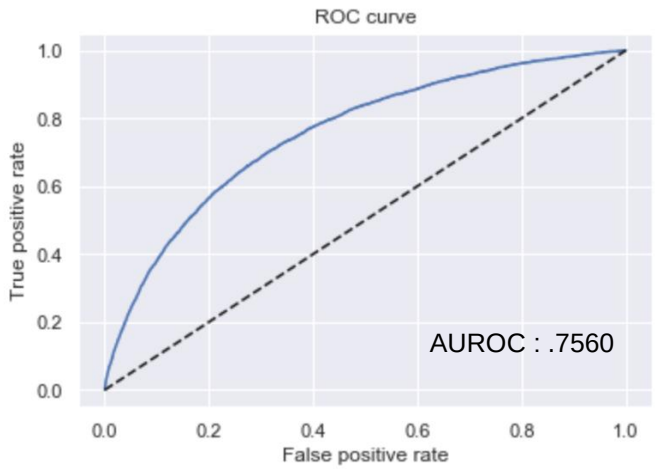
Machine Learning : Model Result and Evaluation

- For each predictive model a confusion matrix was generated on the unseen (test) data, which shows the actual default/non-default applicants and the corresponding ones predicted by the model.
- Each model assigns a probability score to each applicant indicating his/her likelihood of default. The confusion matrix shows the result of actual versus predicted default under the assumption that any applicant with probability score more than 0.5 will default.
- The ROC curve gives a more comprehensive picture showing the ratio of **false positives (predicted as defaults but actually didn't default)** vs **true positives (predicted as defaults and actually defaulted)** rate at every threshold and not only at 0.5. Home Credit can decide on the threshold that would more suite their business objective.
- The **best model** was decided based on the best **AUC-ROC** score returned on the test data set.

Confusion Matrix



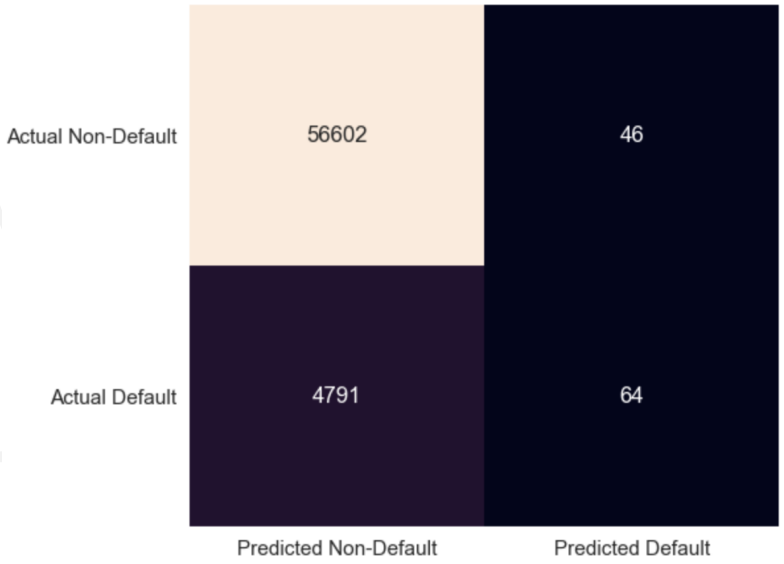
ROC Curve



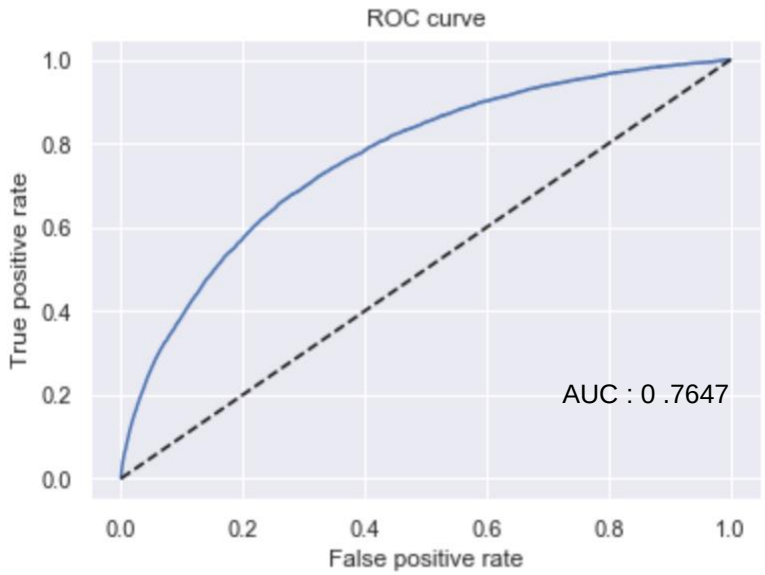
Machine Learning : Best Model - XGBoost

- After Comparing all the models i.e., Logistic Regression, Random Forest, XGBoost and LGBM, XGBoost returned the best AUC-ROC score of **.7655,** which was my evaluation metric. Hence I chose XGBoost to be my final model for model prediction.

Confusion Matrix of test data For XGB



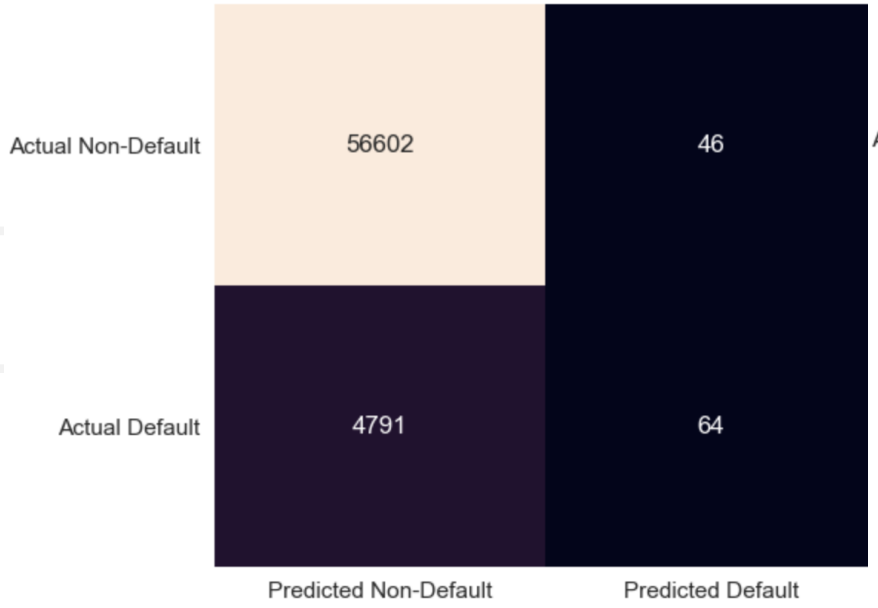
ROC Curve For XGB



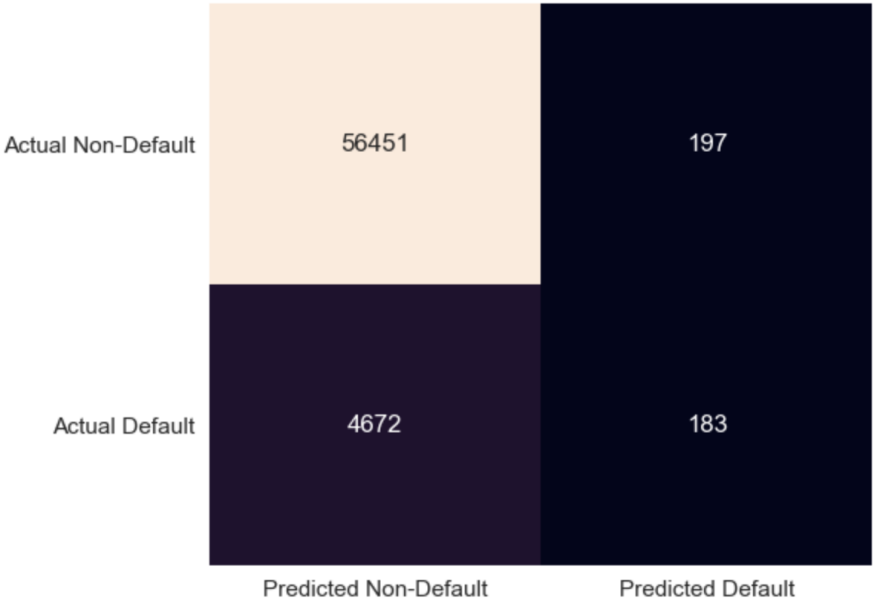
Machine Learning : XGBoost Model – Business Usage

- As the final model is now ready, we need some numbers to decide the business usage of the model by deciding the correct threshold for predicting the default status of an applicant.
- As this data is not handy let's assume some numbers to understand how this model can be used to achieve the business objective which is **ensuring that deserving home credit clients are not rejected and at the same time ensuring minimum loss to home credit because of giving loans to future defaulters.**

Confusion Matrix of test data
For XGB
@ a threshold of 0.5



Confusion Matrix of test data
For XGB
@ a threshold of 0.4



Machine Learning : XGBoost – Business Usage – Example 1

- **As discussed in the previous slide, let assume some numbers in order to understand how to use the model I created for predicting defaults**
  - On average, it costs home credit **$ 20,000** if credit is given to a future defaulter.
  - On average, it costs home credit **$ 5000** in revenue if it refuses a loan to a non-defaulter.
  - Cost to Home Credit when the model used with a threshold of 0.5
    - Cost (0.5) = 46 * 5000 + 4791 * 20000 = 230000 + 95820000 = **960,50,000** $
  - Cost to Home Credit when the model used with a threshold of 0.4
    - Cost (0.4) = 197 * 5000 + 4672 * 20000 = 985000 + 93440000 = **944,25,000** $
  - **Under the above assumption a threshold of 0.4 would make a better choice for home credit as it will reduce the cost for home credit.**

Machine Learning : XGBoost – Business Usage – Example 2

- **In this example let's change the assumed numbers a little and see whether it changes the choice of the threshold**
  - On average, it costs home credit **5,000$** if credit is given to a future defaulter.
  - On average, it costs home credit **$ 10,000** in revenue if it refuses a loan to a non-defaulter.
  - Cost to Home Credit when the model used with a threshold of 0.5
    - Cost (0.5) = 46 * 10000 + 4791 * 5000 = 460000 + 40380000 = **244,15,000** $
  - Cost to Home Credit when the model used with a threshold of 0.4
    - Cost (0.4) = 197 * 10000 + 4672 * 5000 = 1970000 + 23360000 = **253,30,000** $
  - **Under the above assumption a threshold of 0.5 would make a better choice for home credit as using this threshold will cost less to the company.**

Machine Learning : Conclusion

- We chose **XGBoost** with certain values of hyperparameters as our final machine learning model as it returned the best score of **.7647** for our chosen evaluation metric of **AUC-ROC**.
- The Business usage of model will depend on the **actual cost to home credit for giving a future defaulter a loan and also on the cost of not giving the loan to a future non-defaulter**.
- These numbers will help **decide the actual probability threshold home credit shall use for the model to optimize giving loans** to deserving candidates and at the same time minimize its loss because of giving credit to a future defaulter.
- Home credit can also use this model to understand the likelihood of default for a particular applicant and take calculated risk by adjusting his/her loan interest rate accordingly for these candidates to minimize the loss.