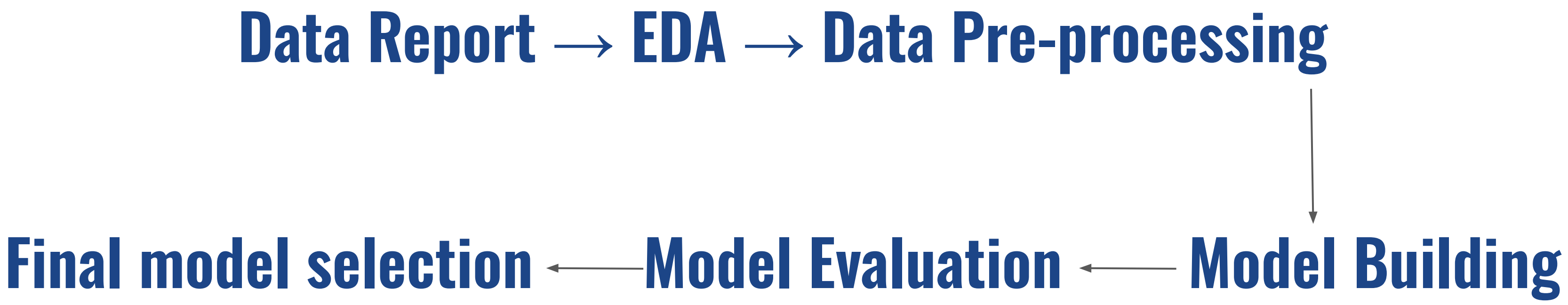
LOAN DEFAULT PREDICTION

GROUP 6

Solution Approach



Data Report → EDA → Data Pre-processing

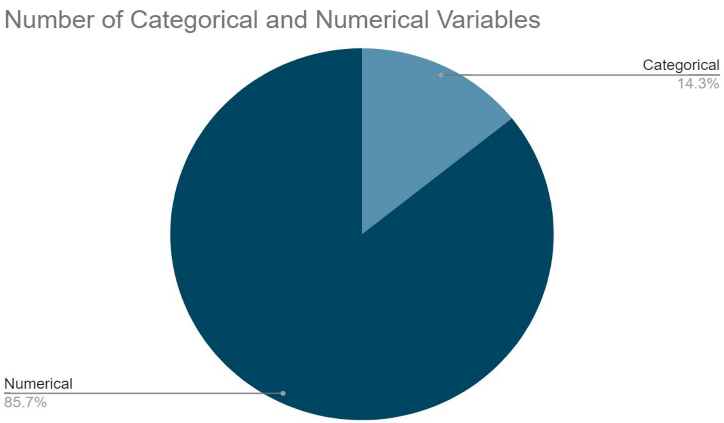
Final model selection ← Model Evaluation ← Model Building

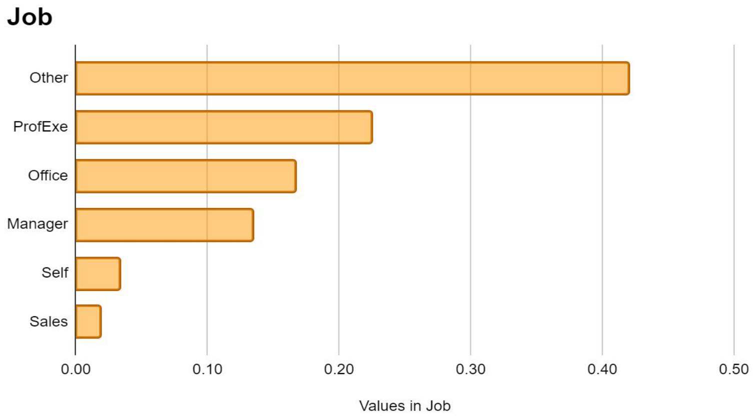
Objective

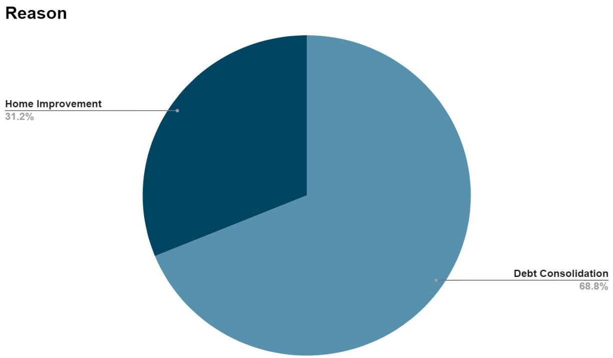
- To develop a robust and accurate classification model to predict which clients are likely to default on their home equity loans based on the Home Equity dataset (MEQ)
- To leverage the 12 input variables to build a time and cost efficient model to analyse the loan default risk for every individual
- To minimize the risk of loan default by identifying high default risk individuals

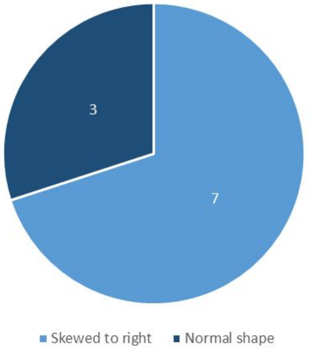
EDA Summary

- There are a total of 12 variables
- Most are numerical variables with only 2 categorical ones, ‘***REASON***’ and ‘***JOB***’
- **6** Job categories and most of the loans bought for debt consolidation
- **7** numerical columns were right skewed and rest **3** were of normal shape



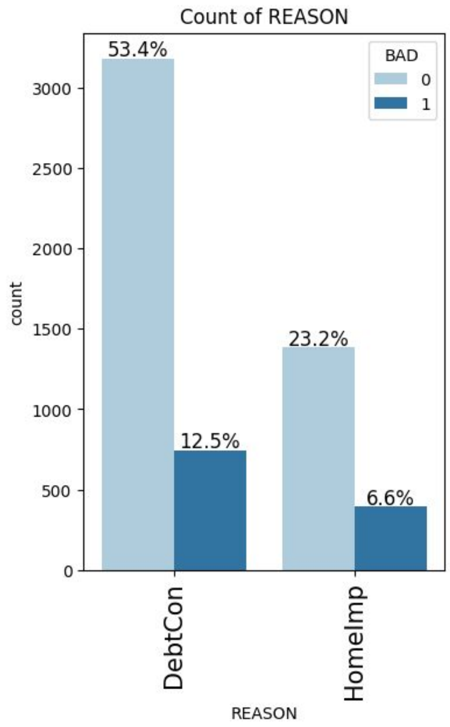


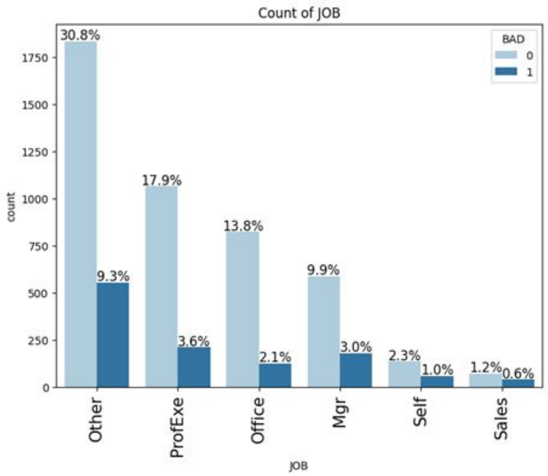




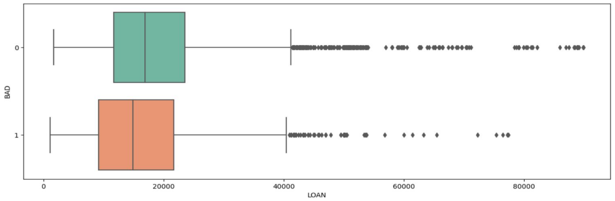
From the 65.9% of debt consolidators, 12.5% of this chunk have defaulted on their loans while 6.6% out of the total 23.2% home improvement applicants have defaulted

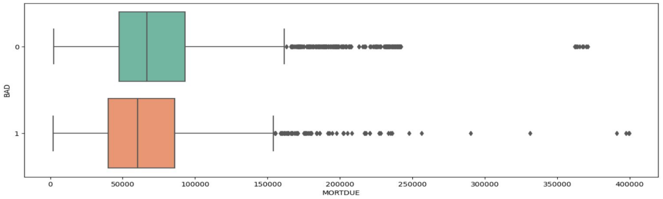
Loan amount , Amount of mortgage due, The Value of the property and The Debt to Income ratio have no effect on the chances of defaulting

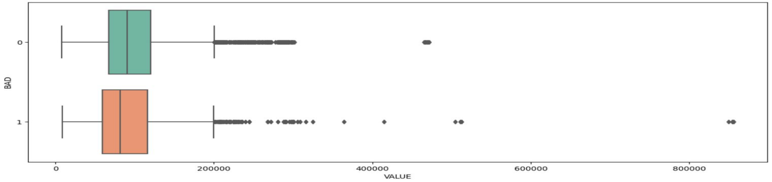


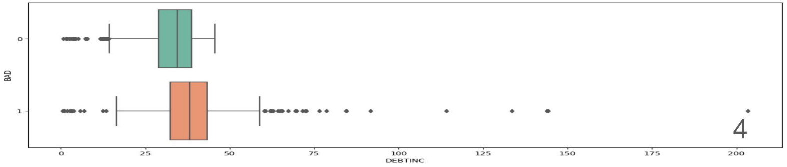


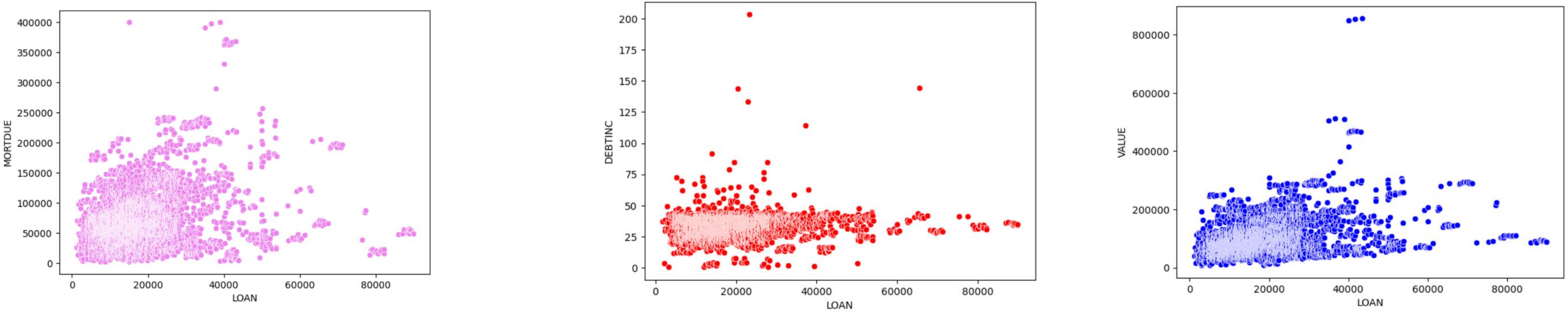
We further divide the job bars to check the percentage of defaulters in each job class



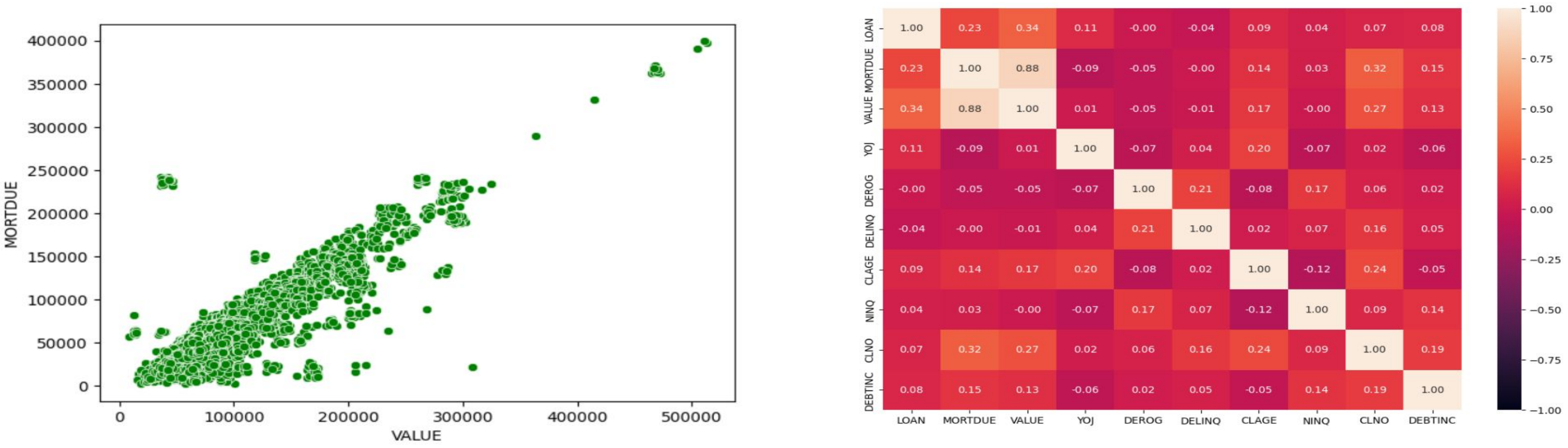








There is no correlation between the Loan with respect to Mortgage Due, Debt to Income Ratio and Value of Property



	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
LOAN	1.00	0.23	0.34	0.11	-0.00	-0.04	0.09	0.04	0.07	0.08
MORTDUE	0.23	1.00	0.88	-0.09	-0.05	-0.00	0.14	0.03	0.32	0.15
VALUE	0.34	0.88	1.00	0.01	-0.05	-0.01	0.17	-0.00	0.27	0.13
YOJ	0.11	-0.09	0.01	1.00	-0.07	0.04	0.20	-0.07	0.02	-0.06
DEROG	-0.00	-0.05	-0.05	-0.07	1.00	0.21	-0.08	0.17	0.06	0.02
DELINQ	-0.04	-0.00	-0.01	0.04	0.21	1.00	0.02	0.07	0.16	0.05
CLAGE	0.09	0.14	0.17	0.20	-0.08	0.02	1.00	-0.12	0.24	-0.05
NINQ	0.04	0.03	-0.00	-0.07	0.17	0.07	-0.12	1.00	0.09	0.14
CLNO	0.07	0.32	0.27	0.02	0.06	0.16	0.24	0.09	1.00	0.19
DEBTINC	0.08	0.15	0.13	-0.06	0.02	0.05	-0.05	0.14	0.19	1.00

There is a linear correlation between the Current Value of a Property and the Mortgage Due, a high positive correlation

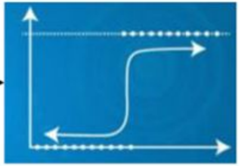
Data Preprocessing Summary

Missing value treatment	KNN Imputer (K=3)
Outliers treatment	99th whisker
Data transformation	Standard Scaler

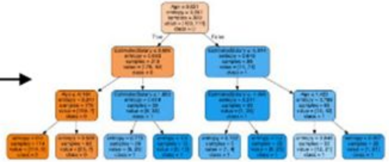
No changes in data distribution after pre-processing

MODEL SELECTION

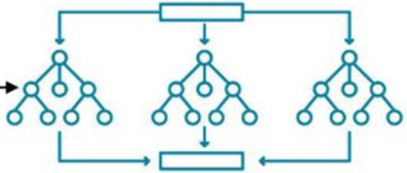
Logistic regression



Decision Trees



Random Forest



Model Performance

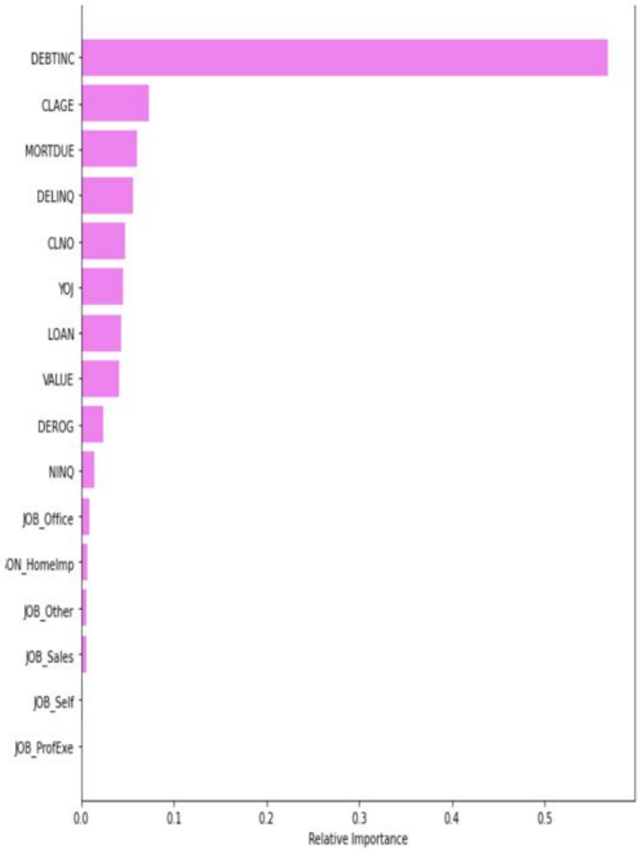
	Baseline model				*Tuned model*			
	AC	**RC**	**PR**	**F1**	**AC**	**RC**	**PR**	**F1**
Logistic Regression	**0.81**	**0.23**	**0.62**	**0.34**	**0.71**	**0.60**	**0.36**	**0.45**
Decision Tree	**0.86**	**0.50**	**0.69**	**0.58**	**0.77**	**0.78**	**0.45**	**0.57**
Random Forest	**0.89**	**0.61**	**0.83**	**0.70**	**0.89**	**0.69**	**0.77**	**0.73**

Model Performance Summary

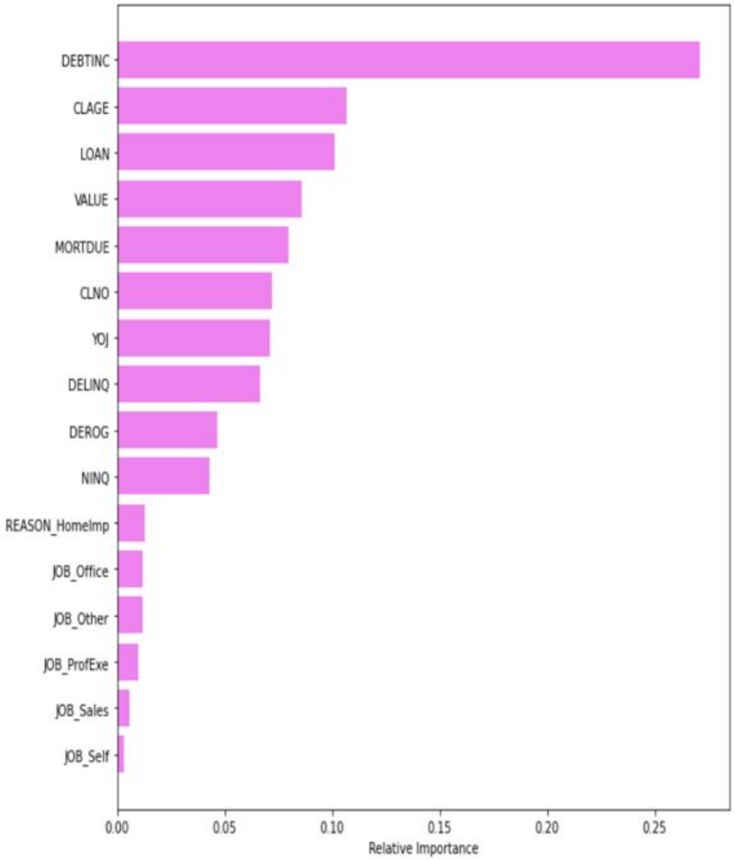
Model	Accuracy	Recall	Precision	F1	Pros	Cons
Logistic Regression	0.81	0.23	0.62	0.34	Accuracy is High, Easy to interpret.	Recall is very low among all other models.
Tuned Logistic Regression	0.71	0.60	0.36	0.45	Recall is increased that of baseline model	Precision is lowest among all the models.
Decision Tree	0.86	0.50	0.69	0.58	Accuracy is very high. Easily interpretable.	Precision value is low. Accuracy is also lower than random forest.
Tuned Decision Tree	0.77	0.78	0.45	0.57	Recall is highest among the models.	Accuracy is lower than the Untuned decision tree.
Random Forest	0.89	0.61	0.83	0.70	Accuracy is highest among all other models.	Non interpretable.
Tuned Random Forest	0.89	0.69	0.77	0.73	Highest accuracy, good recall score	Non Interpretable.

Important Features

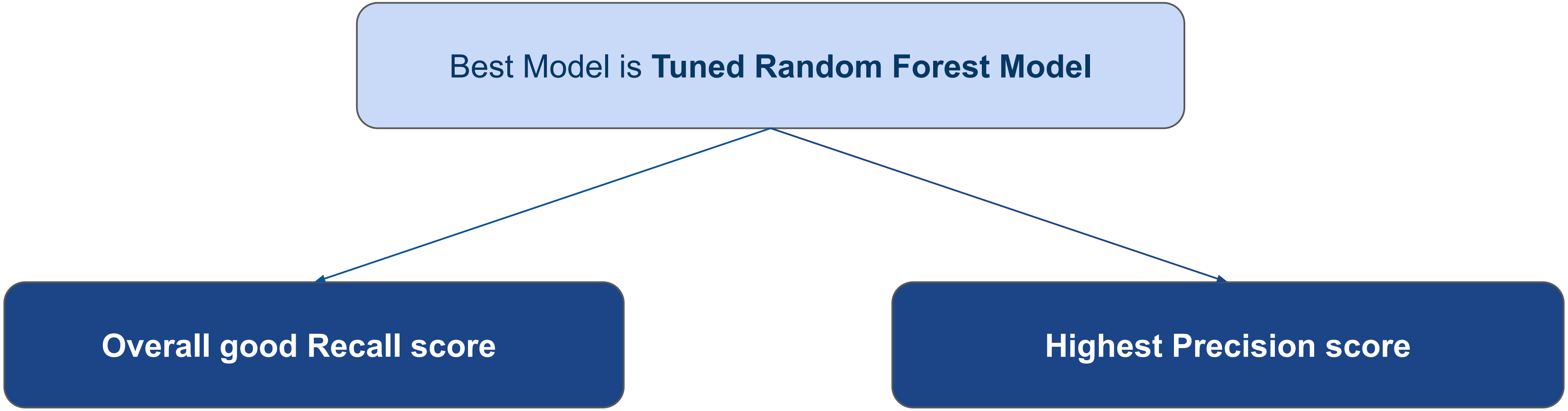
Decision Tree



Random Forest



Final Model Selection



Best Model is **Tuned Random Forest Model**

Overall good Recall score

Highest Precision score

- Tuned Decision tree is giving the highest recall however the precision of this model is very low at 45%
- On overall basis, tuned Random forest is giving a better prediction with a good recall and highest precision
- The f1-score of the model is also highest for Random forest

Summary

- Random forest model can predict the loan defaulters almost 70% of the time .
- The model is also giving the highest overall precision.
- The most important feature that are considered for the prediction are DEBTINC, CLAGE, LOAN, VALUE and MORTDUE.
- The Debt/income ratio is the most important feature, but also the one with the most missing data(21.5%) which is similar to the proportion of the defaulted customers(20%).

Business Recommendations

Debt to income ratio is a very powerful tool in predicting defaulters.

- The bank can use debt to income ratio as a initial indicator when evaluating a loan.
- Those with a higher debt to income ratio can also be made aware of the potential difficulties of paying off a loan when already in a larger portion of debt to income.
- Need alternate business solution for applicants without the info of debt/income ratio

Those who have a higher value of their current property and are asking for a ***larger loan amount*** are generally more likely not to default. This makes sense as those who are wealthier are more financially stable.

- It is important to not develop bias towards those who are less wealthy and potentially deny them the opportunity to grow into a better living. Therefore rather than a hard cut off banks can use this metric as a means to be more critical or feel more secure in providing loans.

EDA showed that certain ***JOBS*** have a higher portion of loans compared to others, this despite the fact that each had relatively equal default percentages.

- If the bank is giving out loans preferably to certain jobs, this should be mitigated as there is no indication that any job is less likely to default. Rather such actions would deny the bank potential profits from those jobs that are given less access.
- If the banks is giving out loans equally and certain jobs are simply applying more, then that presents an opportunity. For those jobs that are applying less, it could be beneficial to research why that is and if they could be converted into potential customers.