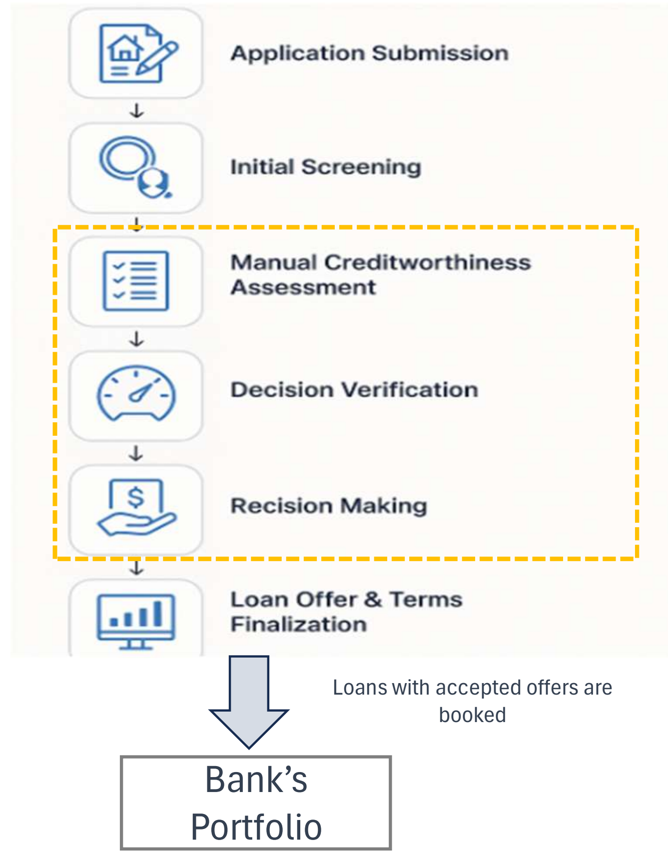
Capstone Project : **Loan Default Prediction** - Use Practical Data Science Techniques

Problem definition

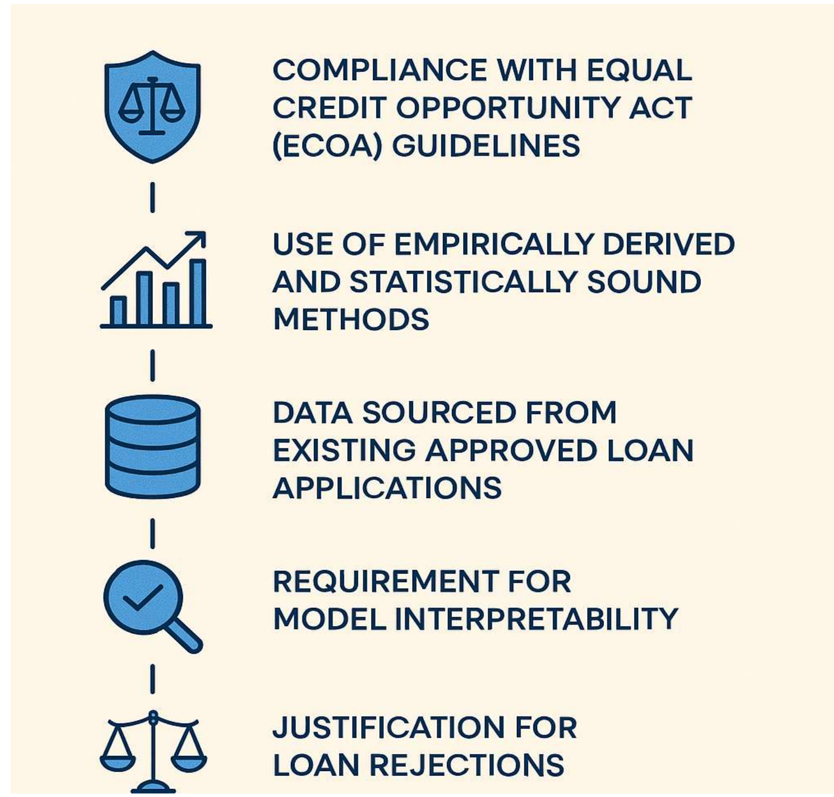
- Retail bank's profits come from interest income earned through lending products
  - Loans that do not repay and default lead to losses and impact profits
- Banks have judicious application process to check credit-worthiness of the applicants.
  - Manual checks and verification are effort intensive and prone to wrong judgment due to human error and bias



Loans that repay generate Income and Profits.
Loans that default lead to Loss

Objective

- To simplify the approval process for Home Equity Line of Credit by having predictive model
  - meets Regulatory guidelines
  - based on internal/existing loans and repayment behavior
  - based on sound model development techniques
  - Interpretable and explainability
    - *Able to explain adverse characteristics to rejected applications*

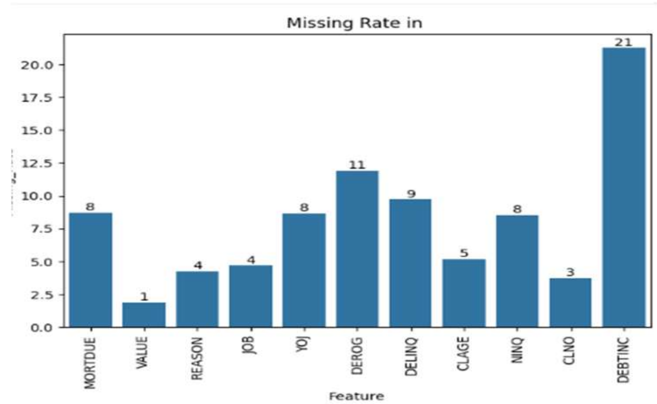


Solution Approach

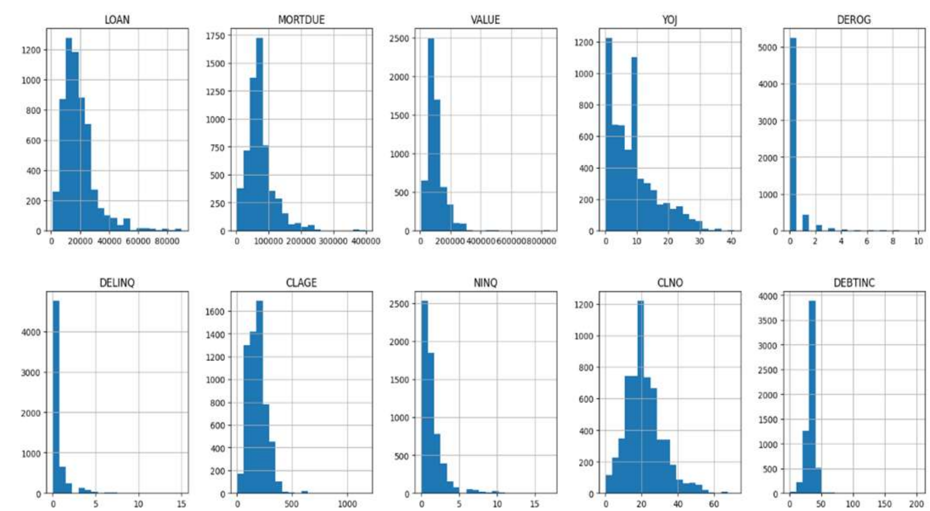
- Data Exploration
  - Data quality checks
  - relationship between different variables as well as dependent variable
- Evaluate Classification techniques for predictive modeling
  - Logistic Regression
  - Decision Trees
  - Random Forest
- Various model Performance metrics were evaluated
  - Accuracy
  - Misclassification - Precision and Recall
  - Receiver’s Operating Curve
- Model Interpretability and Transparency
  - Feature Importance
  - Shapley values

Data Quality

- Existing loan data with repayment behavior had 5,960 loans and 13 variables
- 11 variables has missing values, and 2 variable had 10+% missing rate



- Variable distribution show right skewed distribution across all numeric variables



Given the small data size and limit number of variables, missing value and outlier treatment was applied

Bivariate Analysis

- The bad rate 19.9% for loans

- The loans that default had
  - Relatively less work experience
  - Higher # of Delinquent Credit line
  - Less credit history
  - Higher # of recent inquiries
  - Higher Debt to income ratio

- Job types - Self-employed and Sales occupation had higher default rate

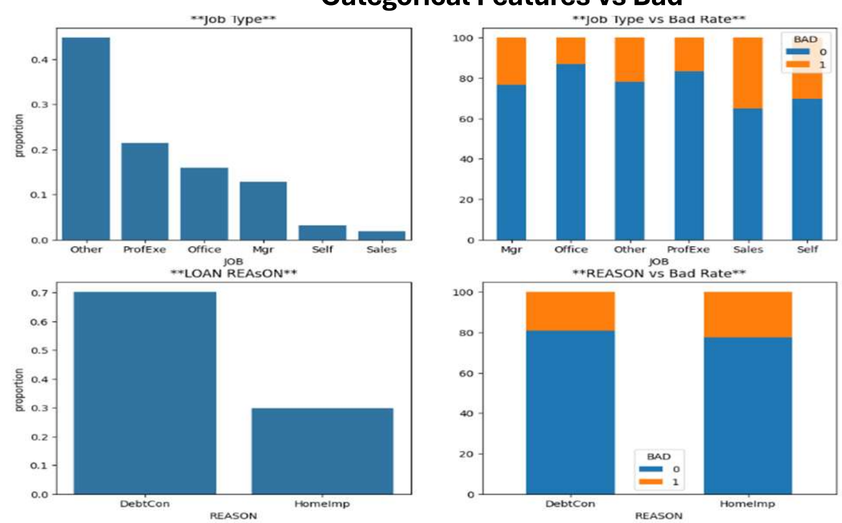
- There was no significant difference in default rate by loan purpose, existing number of Credit lines

- Loans that defaulted had lower existing mortgage balance, property values
  - Both these variables showed strong correlation

Comparison of Mean Values - Good vs.Bad Loan

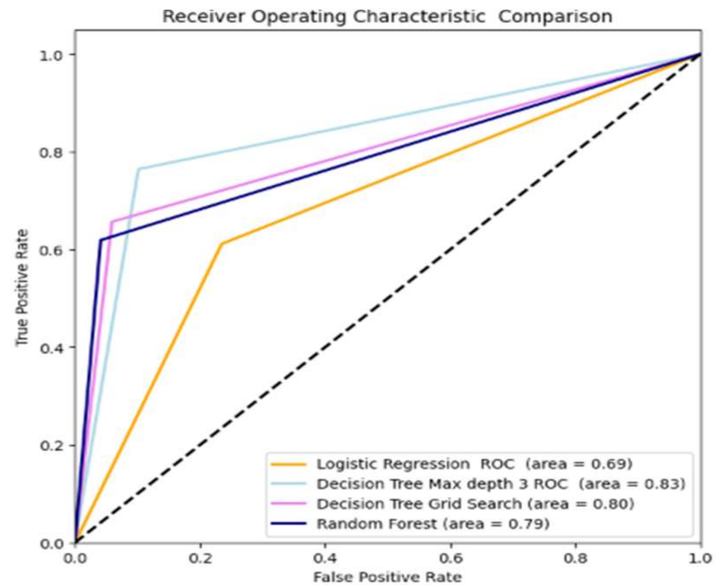
Dependent variable	Amount approved	Amt due on existing mortgage	Curr Property Value	Years at present Job	# of Major derog	# of Delq credit Lines	Age of Oldest TL	# of recent Inq	# of Existing Credit Lines	Debt to Income ratio
Good	19,028	74,829	102,596	9.2	0.1	0.2	187.0	1.0	21.3	33.3
Bad	16,922	69,460	98,173	8.0	0.7	1.2	150.2	1.8	21.2	39.4

Categorical Features vs Bad



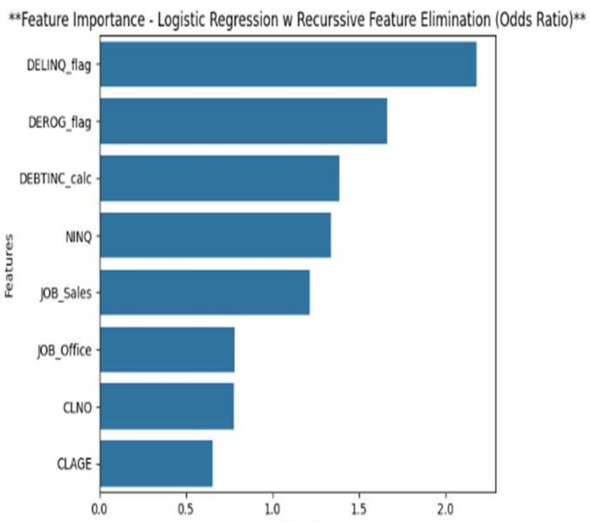
Proposed Model Solution Design

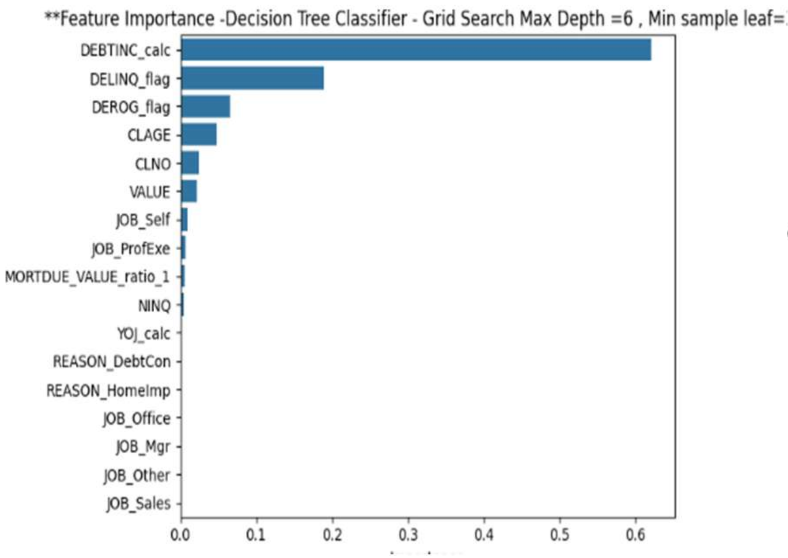
- Classification techniques evaluated
  - Logistic Regression
  - **Decision Trees**
  - Random Forest
- Model Performance Comparison
  - Accuracy
  - Precision and Recall
  - ROC
- Model Interpretability - Feature Importance

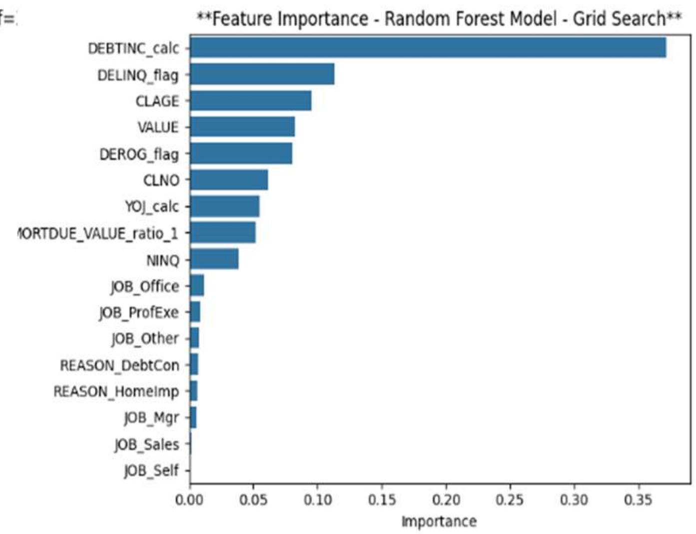


Model Name	Precision	Recall	Accuracy	F1_score
Logistic Regression w. RFE Train	65.4%	70.7%	75.6%	66.6%
Logistic Regression w RFE Test	65.0%	68.8%	73.1%	66.0%
Decision Tree(max depth 3) Train	77.6%	81.7%	86.4%	79.4%
Decision Tree (max depth 3) Test	80.7%	83.1%	86.9%	81.8%
DecisionTree with Grid Search - Train	83.1%	82.3%	89.5%	82.7%
DecisionTree with Grid Search - Test	83.3%	79.9%	87.8%	81.4%
RandomForest with Grid Search - Train	94.4%	89.1%	95.1%	91.5%
RandomForest with Grid Search - Test	85.5%	78.9%	88.3%	81.5%

Proposed Model Solution: Feature Importance



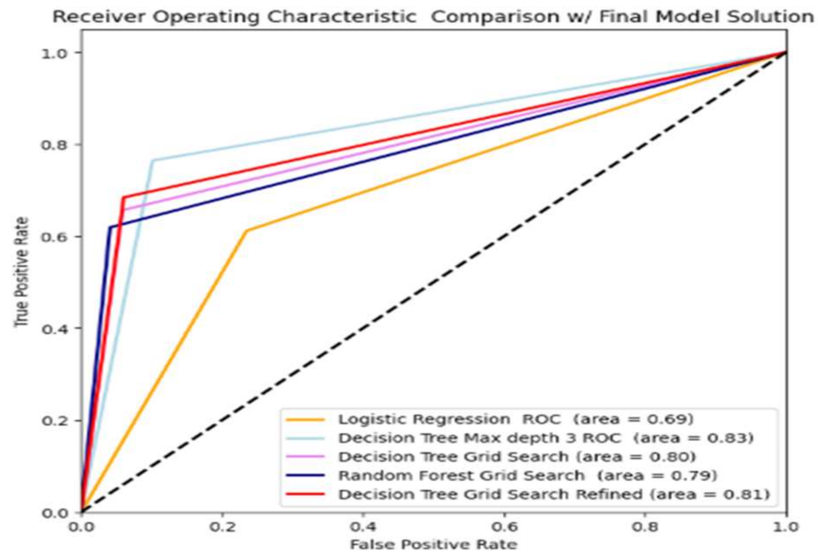




3 out of top 5 features are same across all the 3 classification techniques.

Final model Solution

- Decision Tree based model was selected
  - Easy to interpret
  - Feature selection – Transparency
  - Strong Performance
    - High Accuracy (~89%)
    - High Precision(~82%) and had better Recall(~82%) than Logistic regression
    - **Lower stability(~3 % drop in Recall and 2% drop in Accuracy)**

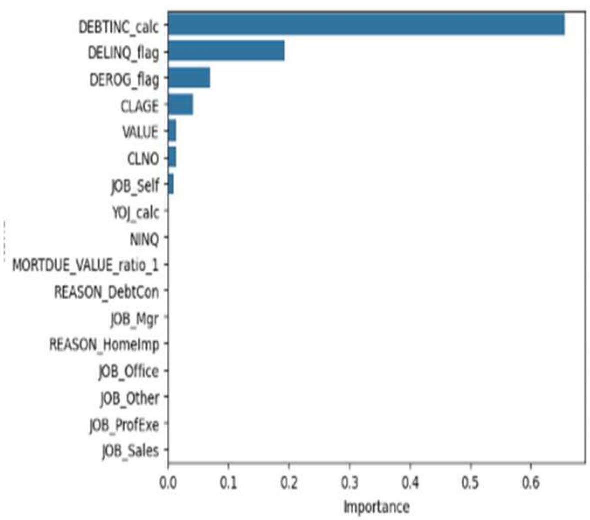


Decision Tree Model with additional hyperparameter tuning show stable performance and reduction in false positive rate.

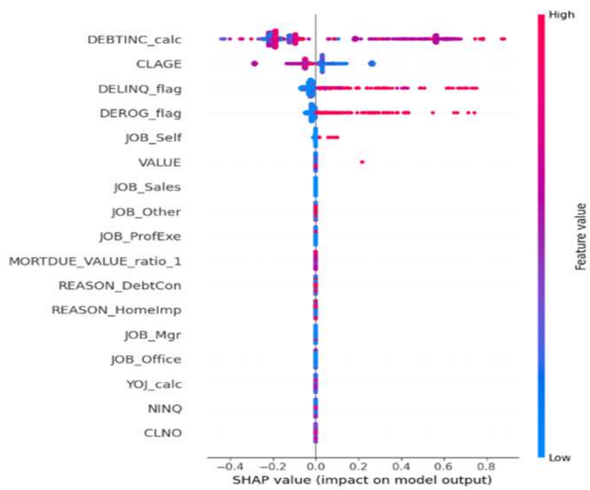
Model Name	Precision	Recall	Accuracy	F1_score
Logistic Regression w. RFE Train	65.4%	70.7%	75.6%	66.6%
Logistic Regression w RFE Test	65.0%	68.8%	73.1%	66.0%
Decision Tree(max depth 3) Train	77.6%	81.7%	86.4%	79.4%
Decision Tree (max depth 3) Test	80.7%	83.1%	86.9%	81.8%
DecisionTree with Grid Search - Train	83.1%	82.3%	89.5%	82.7%
DecisionTree with Grid Search - Test	83.3%	79.9%	87.8%	81.4%
RandomForest with Grid Search - Train	94.4%	89.1%	95.1%	91.5%
RandomForest with Grid Search - Test	85.5%	78.9%	88.3%	81.5%
Decision Tree GV Refined Train (Final Model)	**82.4%**	**82.6%**	**89.2%**	**82.5%**
Decision Tree GV Refined Test (Final Model)	83.8%	81.2%	88.3%	82.4%

Final model : Feature Importance and Interpretability

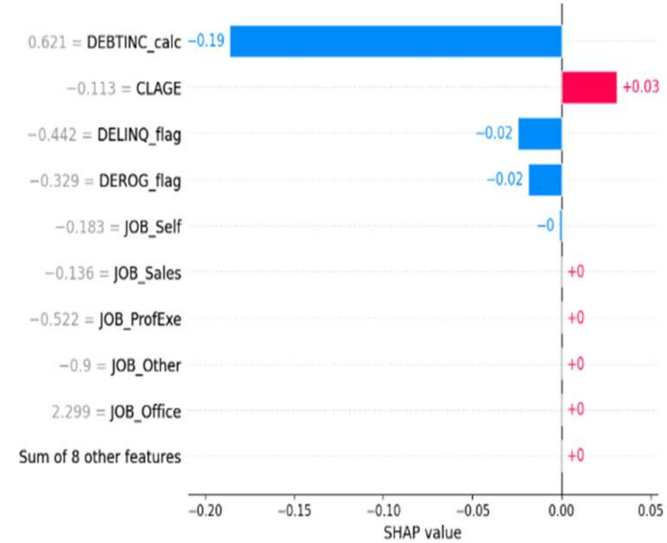
Feature Importance



Global Feature Impact



Localized Feature Impact on a loan



- Debt Income ratio, prior Delinquency and Derogatory behavior , length of credit history are top model features
- SHAP Value shows the impact each of features at overall level and loan level.
  - These calculation would need to be implemented in automated underwriting system to explain reason for rejection

Proposed Business Solution

- Business need to evaluate following areas as part of introducing the model in application decision process

Application Processing	Cost vs Benefit
Application Platform	<u>Benefit</u>: Reduce errors and have systemic controls on decision making , improved data gathering <u>Cost:</u> Enhancement/Upgrade to Platform to handle calculation of attributes and approval /decline decisions
Underwriting	<u>Benefit:</u> Reduce human errors or bias, Ability to focus on highrisk applications and accuracy of data <u>Cost</u>: Training underwriters to adopt and use the model effectively in application process
Credit and Data Science	<u>Benefit</u>: Control on lending decisions with use of enhance application platform, ability to increase application volume <u>Cost</u>: Introduce monitoring and identify the times frame to update the model and thresholds for approve /decline decisions
Regulatory	<u>Benefit</u>: Model inputs used for decision making can used to explain the adverse behavior and reason for decline <u>Cost</u>: Tracking the effectiveness of the model over the time.

Model Use and Implementation

- Define threshold or cut-off based on the new model to approve and decline customer.
- Introduce the new model as challenger to current process to evaluate its effectiveness in use
  - Make any adjustments before fully adopting the model for decision making.
- Need to train underwriting team so that model can be adopted in approval process
  - Address questions related to model-based declined.

Summary

- The predictive model based on Decision tree technique with parameters outperformed other techniques(Logistic regression and Random Forest)

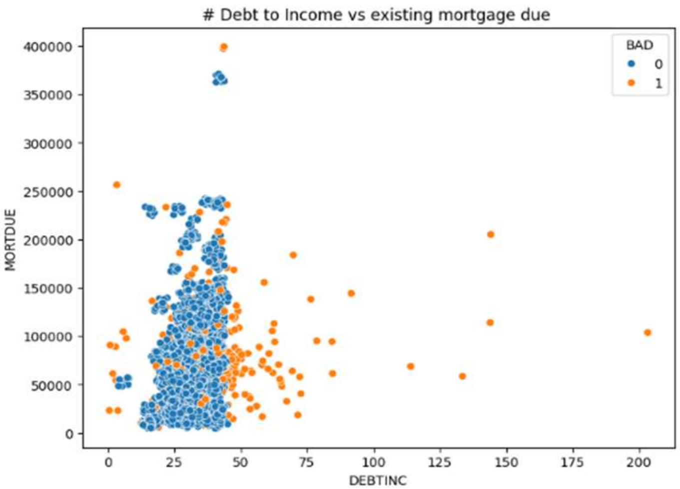
- Presence of Prior delinquency and Derogatory behavior along with High debt to income ratio and short Credit history were key predictors of default

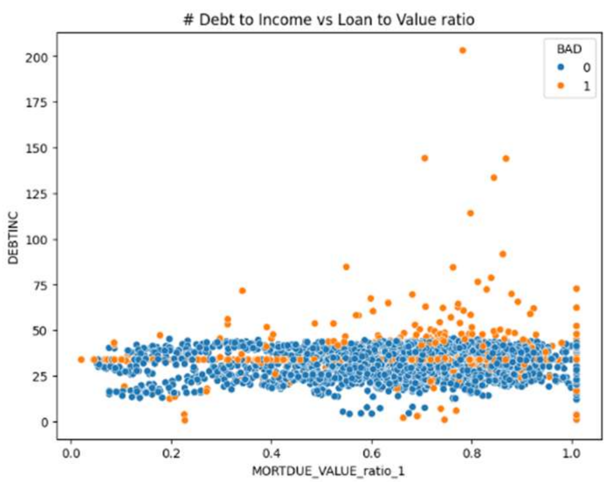
- Model can lead to improvement in application decisioning process
  - Requires that business enhances Origination platform, train and prepare staff to adopt the model in the business process for effectively using the model

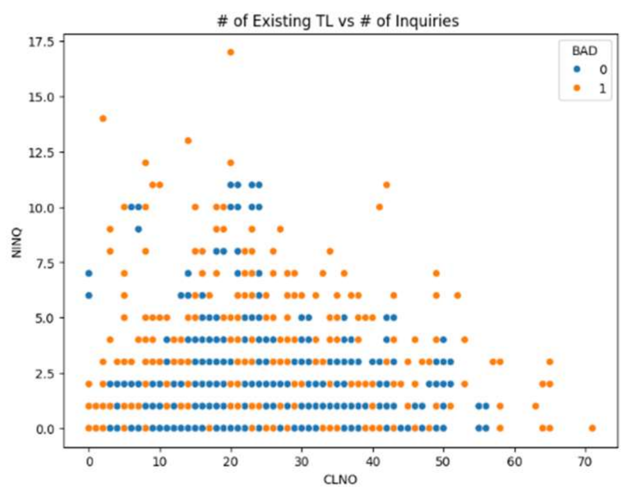
Risk and Challenges

- Improve the application data gathering to ensure availability of applicant characteristics to calculate and model inputs
  - If there are errors or missing data, model-based decisions will be inaccurate.
- Ensure that rules based on Decision tree model get correctly implemented and not get modified without stakeholder awareness
  - Incorrect implementation/changes lead to unintended consequences
- Timely communication of application decision and adverse behavior to ensure regulatory compliance
  - Miscommunication of reject reviews or reasons can lead to compliance issues
- The model is built on approved but will be used on all application
  - Bank need to consider evaluating effectiveness of model on all applications and in future look at developing model to include rejected applications

Multivariate analysis







- Higher concentration of defaults when Debt to income greater than 50% and mortgage amount is 500,000
- higher concentration of the defaults when years of experience is less than 10 years and age of oldest trade below 200 months.
- Defaults are observed when there are 4 or more inquiries.
- The combination of 5+ inquiries and 20+ # of existing trades also have higher defaults