Loan Default Prediction

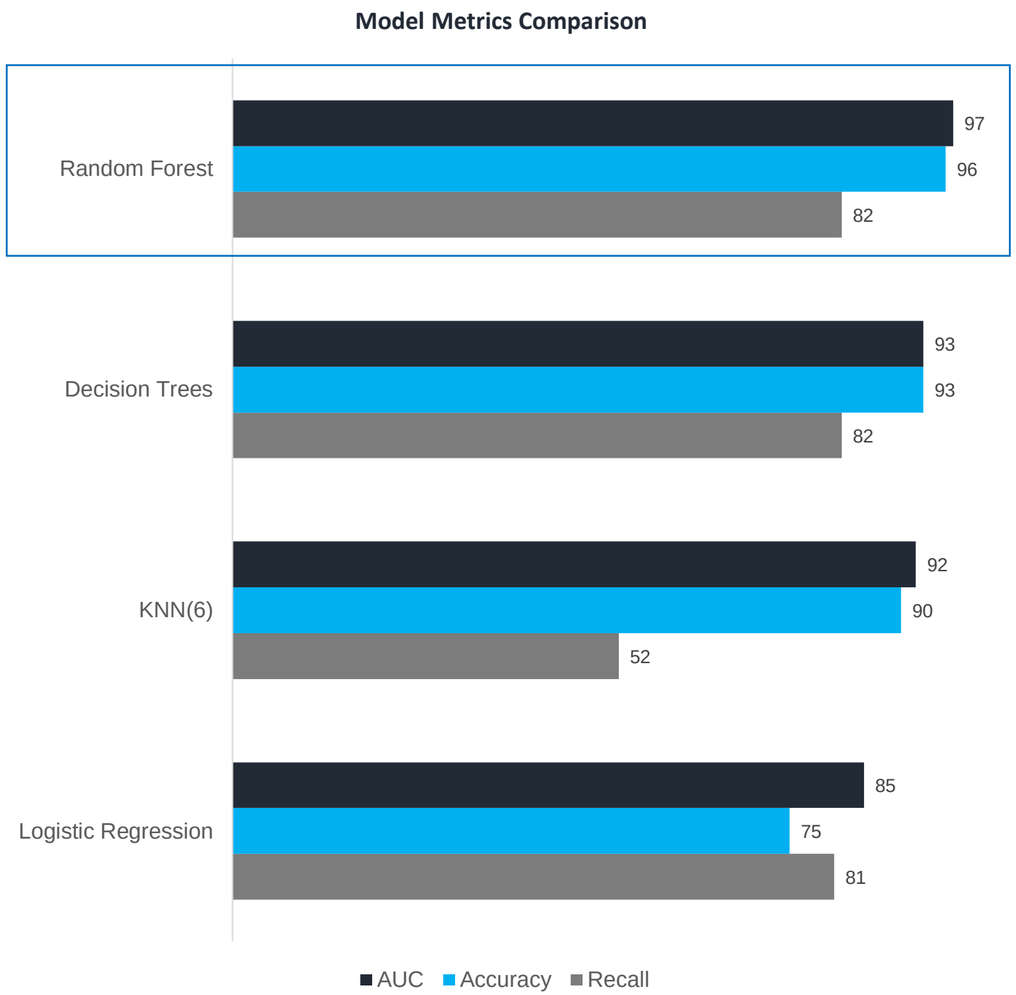
Capstone Project Report – Milestone 2 | Author: Shouvik Nandy

Contents

Refined Insights

Data Preparation	1. There were many observation which had more than 9 out of 12 variables missing, those observations were dropped from the dataset. 2. There were no duplicate records in the data. 3. All missing values were imputed. 4. Since the scale of the numerical independent variable were different, the variables was standardized to have unit standard deviation. This will provide common scale for measuring the impact of variable with other. 5. The categorical parameters were encoded using one-hot encoding. The categorical variables were not ordinal hence we using this method. 6. The class imbalance problems was manged by oversampling the minority class in the training data set using SMOTE technique.
Model Development	1. Logistic Regression model performed better on the scaled dataset. 2. Oversampling of minority class for the training dataset also increased the performance to logistic regression. 3. KNN model didn't perform well on this dataset. 4. Decision Tree performed better than Logistic Regression model. 5. Ensemble technique RandomForest was used to generalize the tree based model. The ensemble model performed the better on this dataset.
Model Parameter Tuning	1. Cross validation methods used to tune model parameters. 2. RandomSearch technique utilized to find initial model parameter values. 3. These values were then tuned further by using Grid Search technique.
Model Evaluation Metrics	Recall, Accuracy & AUC were used to evaluate model performance.

Model Comparison



Score on Validation Data in %

Model	Scaled Data	SMOTE	Hyper Parameter Tuned	Recall Defaulted Loans	Overall Accuracy	AUC
Logistic Regression	No	No	No	14	81	-
Logistic Regression	Yes	No	No	47	86	-
Logistic Regression	Yes	Yes	No	59	86	-
Logistic Regression	Yes	Yes	Yes	81	75	85
KNN(10)	Yes	Yes	No	49	88	-
KNN(6)	Yes	Yes	Yes	52	90	92
Decision Trees	Yes	Yes	No	84	95	-
Decision Trees	Yes	Yes	Yes	82	93	93
Random Forest	Yes	Yes	No	84	96	-
Random Forest	**Yes**	**Yes**	**Yes**	**82**	**96**	**97**

Proposal Final Solution

Based on the model evaluation, Random Forest is the best model for this dataset

Recall & Accuracy

Recall Score (82%) for Random Forest model along with accuracy(96%) is better when compared to other models.

AUC

Random Forest model has very high AUC score (97%), that means on the validation data the model is better in separating the target variable when compared with other models.

Explainable

The tree is lost in Random Forest as its an ensemble technique, we will use SHAP values based on Game theory to explain the model prediction results.

Key Recommendations

Threshold Analysis & Dimensionality Reduction

We can evaluate updating the threshold to evaluate improvement in Recall score, as its is important that we don't approve a loan which has high probability of being defaulted.

We can further evaluate dimensionality reduction technique to improve model performance at the cost of exploitability.

Feature Importance

Features important for classification of the loan application has be identified and addressed for both scenarios:

- Global (overall for the model)
- Local (for specific observation)

Implementation Considerations

There are lot of missing data in this dataset. The model performance can be increased if the missing data are being captured for the new loans.

Status Updates

Solution Design

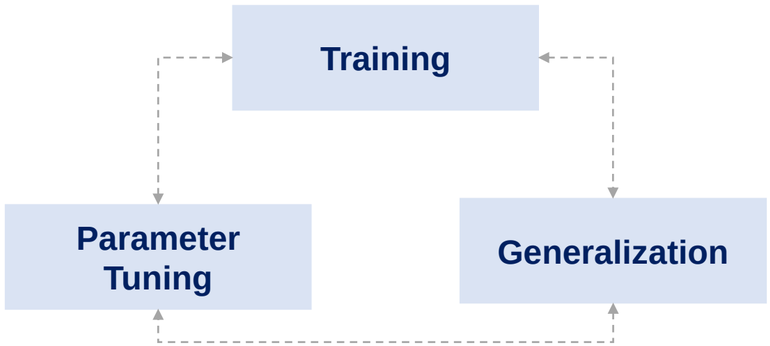
- *Handle missing values*
- *Encode Categorical variables*
- *Scaling of numerical variables*
- *Split data in train , validate and test sets*
- *Use SMOTE for oversampling of imbalance class*

- *Models: Logistic, Decision Tree, ADABoost & Random Forest*
- *Feature Selection*
- *Tune hyperparameters using Random and Grid methods*
- *Generalization & Regularization for overfitting using cross validation methods*
- *Utilized model performance metrics (Recall, Accuracy & AUC) for optimization*

- *Predict target class using the best model*

- *Measure Model performance*
- *Identify and report import features of the model*
- *Utilized SHAP for tree based models*

EDA ❯ **Data Preparation** ❯ **Model Development** ❯ **Prediction** ❯ **Model Evaluation**



- *Completed*
- *In-Progress*
- *Not Started*

Appendix

Notebook	Description	Format
D:\Learning\Data nce\MIT-DS\8. Ca	Notebook with Data processing & Models related to Milestone 2 deliverables.	IPYNB
Capstone_Loans_Default_Prediction_Shouvik_V2.html	Notebook with Data processing & Models related to Milestone 2 deliverables.	HTML