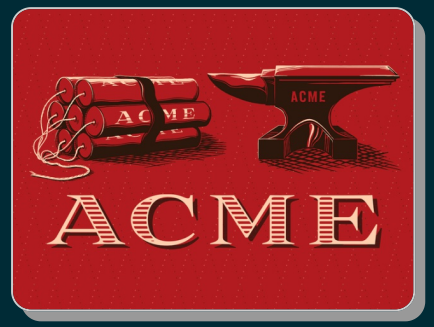


Machine Learning Model for Customer Loans at Acme Bank

Alfonso Tobar Arancibia

Data Scientist

2019-10-06

github.com/Rcubes/acme

tobar_with_R

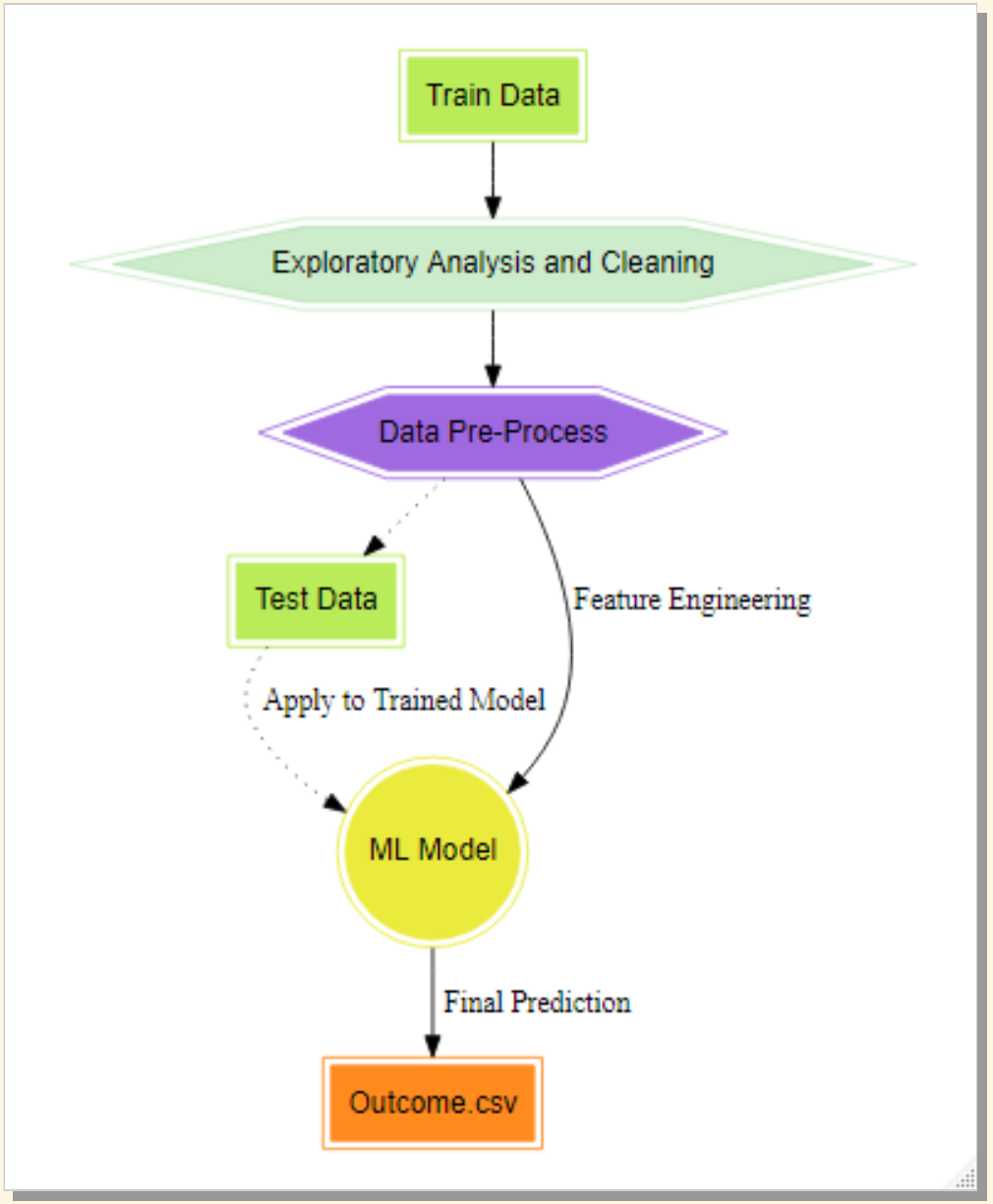
The Problem

The Problem

- **Acme Bank** has requested a Loan Evaluation Model using Machine Learning.
- The bank provided a Dataset containing Customer Information: Loan Amount, Customer Living State, Average Account Balance among others.
- The Models needs to Predict the **Loan Approval** by using the Data Provided.

The Data Flow

- Training Data goes to Exploratory Analysis and Cleaning.
- Data Pre-Process is applied to the Training and Testing Data.
- The Machine Learning Model is Trained.
- Test Data goes into the Trained Model to Predict Final Outcome.



Data Description

Feature	Type	Description
id	int	The unique ID assigned to every loan application.
Loan Amount	int	The loan amount applied for.(in $$)
Term	str	The number of months to repay the loan.
State	str	The state that the applicant is applying from.
Annual Income	int	The applicant's stated annual income. (in $$)
Income Verification Status	categorical	Has the bank verified the applicant's income information? Values are Verified, Partially Verified and Not Verified.
Average Account Balance	int	The applicant's average balance of all cash accounts over a period, e.g. bank and brokerage. (in $$)
Due Amount	int	Total late fees due from the applicant. (in $$)
Home Ownership	categorical	The home-ownership status provided by the applicant. Values are RENT, OWN, MORTGAGE, OTHER.
Loan Purpose	categorical	The purpose of the loan
Due Settlement	str	Has the applicant has opted for settlement against any loan in the past? Values are N (not opted) and Y (opted).
Installment Amount	int	The monthly payment owed by the borrower if the loan gets approved. (in $)
Payment Plan	str	Indicates if the borrower has suggested a loan repayment plan or not. Values are n (not suggested) and y (suggested).
Approve Loan	int	The approval status of the loan. Values are 0 and 1 where 0 indicates loan is not approved while 1 indicates the loan is approved.

Data Harmonization

Exploratory Analysis

This step aims to identify main issues within every Variable. The Main Issues found were:

- **Term**, **Income Verification Status** and **Due Settlement** have Missing Categorical Values.
- **Payment Plan** is Highly Imbalanced, having no values for *Suggested Loan Repayment* Option.
- **ID** is just a Data Identifier.
- **Loan Amount**, **Annual Income**, **Average Account Balance**, **Due Amount** and **Installment Amount** have missing Numerical Values.
- **Annual Income**, **Average Account Balance** and **Due Amount** are Right Skewed (High Concentration in values near to Zero).
- **Loan Amount** and **Installment Amount** have a strong correlation.
- **State** and **Payment Plan** are Independent (there is no relationship) to **Approve Loan** (Variable to Predict).

Data Pre-Process (Feature Engineering)

After Exploring the Data the Data needs to be treated properly before building the Machine Learning Model. The different Problems found in the Data will be treated as follows:

- **Id**, **Payment Plan** and **State** were dropped since they don´t contribute to the Model.
- Missing Categories in **Term** were Imputed using Mode (Most Frequent Value).
- **Income Verification Status**, **Due Settlement** missing Categories were Imputed using a Tree Based Model.
- All the Numerical Missing Values were Imputed using a K Nearest Neighbor Model. (Simulated values acording to Neighbors).
- **Installment Amount** and **Loan Amount** were transformed into a Ratio to avoid high Correlation.
- Skewed Variables were sanitized through a Box-Cox Transformation.

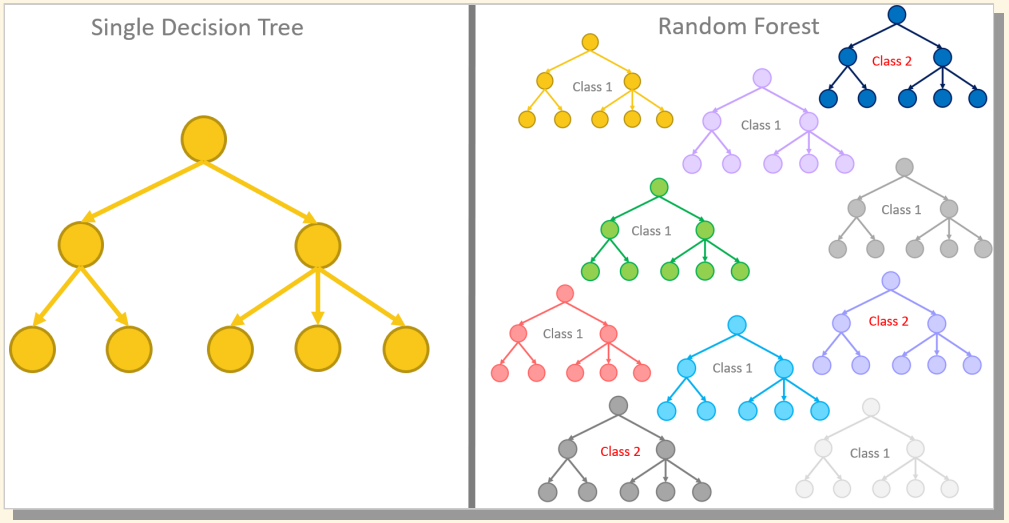
Machine Learning Model

Model Building

Model: Random Forest:

Data Split: 70/30

Main Parameter: Number of trees: 100



Validation Results

Confusion Matrix		
Prediction	**Truth**	
	Not Approved	**Approved**
Not Approved	245	7
Approved	2	313

Metric	Value
accuracy	0.9841270
sens	0.9781250
spec	0.9919028
bal_accuracy	0.9850139
precision	0.9936508
recall	0.9781250
AUC	0.9922444
Note: Values obtained in the Validation Set	

Honest Assessment

Once the Model has been tuned it is necessary to Validate the Results.

In this case a 10-Fold Cross Validation Technique was used.

Validation Metric: ROC AUC

This Metric corresponds to the Area under the ROC Curve.

Fold	Metric	Value
1	ROC AUC	0.9851368
2	ROC AUC	0.9995469
3	ROC AUC	0.9993316
4	ROC AUC	0.9912787
5	ROC AUC	0.9919655
6	ROC AUC	0.9892202
7	ROC AUC	0.9948245
8	ROC AUC	0.9929529
9	ROC AUC	0.9898073
10	ROC AUC	0.9955125
Note: ROC AUC for Metric for every Fold		

Conclusions

Conclusions

- A robust Random Forest model was built. The model was chosen because its easy Tuning and Power against Overfitting. These qualities are particularly useful since the model needs to be tested against unseen/unclassified data.
- To combat overfitting a Cross-Validation technique was applied. Solid Performance Metrics were obtained Using 10-Fold Cross Validation.
- The Prediction Results can be found in the *outcome.csv* file that can be found in this same Folder.

Model Final Variables	
Categorical	**Numerical**
Term	Annual Income
Income Verification Status	Average Account Balance
Due Settlement	Due Amount
Loan Purpose	Installment / Loan Amount Ratio
Home Ownership	
Approve Loan	
Note: Variables Obtained After Feature Engineering