C3879C Capstone Project

Loan Prediction

Adelene Ng

Introduction

- ABC Finance is a finance company dealing with home loans
- They wish to automate their loan process – currently it is manual
- Wish to have a intelligent online system – customer furnishes details like occupation, income
- System will be able to validate if he/she is eligible for the loan in real time

Problem Statement

- Manual loan approval process is tedious and time consuming
  - A lot of Paper work
  - Repetitive workflows
- Error prone
- Reduce the risks around the loans

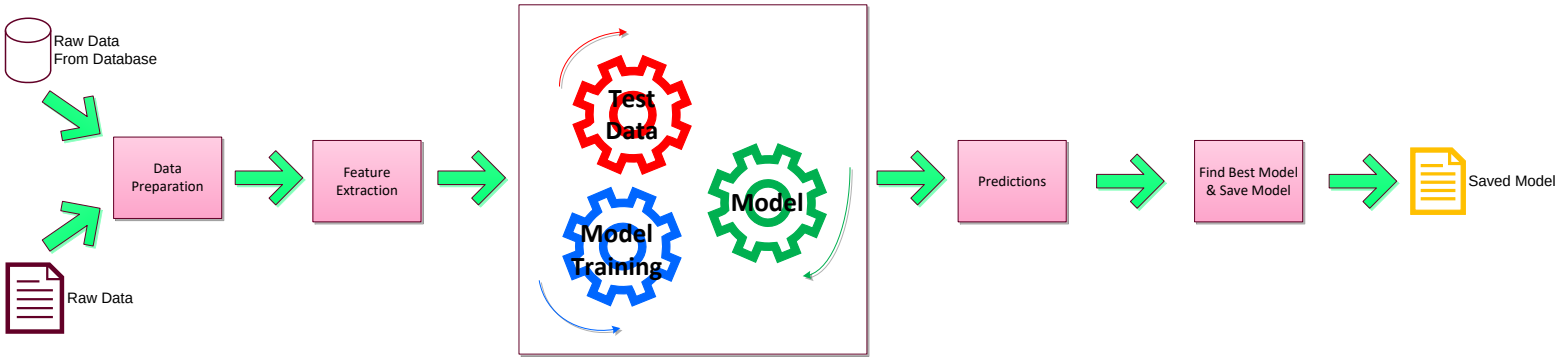
Data Set

Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	Credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural

Feature Engineering

New Features	Description
EMI (Equated monthly instalments)	This is the monthly amount to be repaid by the applicant. This is calculated by taking the ratio of loan amount with respect to loan amount term.
Balance Income	This is the income left after the EMI has been paid
Total Income	This is the sum of Applicant Income and Co-applicant Income

System/Software Design

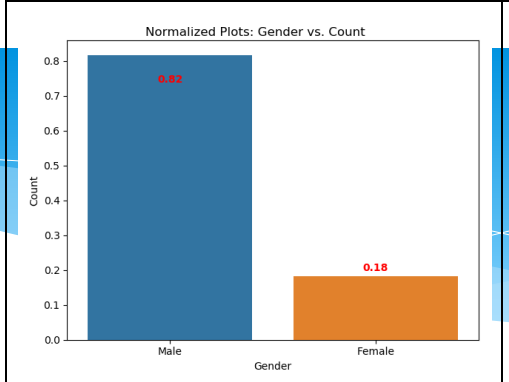


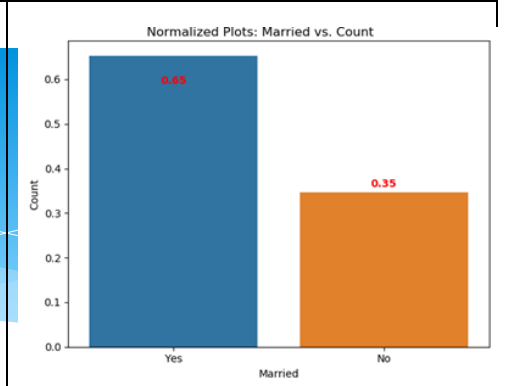
Data Analysis

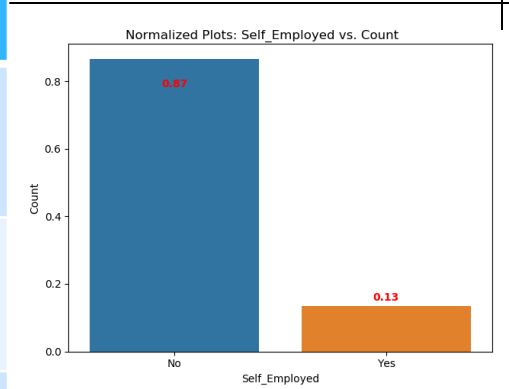
- Categorical data
  - Also called a nominal variable is one that has two or more categories, but there is no intrinsic ordering to the categories.
- Ordinal data
  - variables have natural, ordered categories and the distances between the categories is not known.
- Numerical data

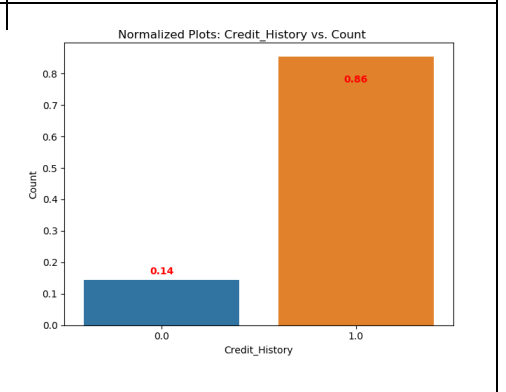
Categorical Data Analysis

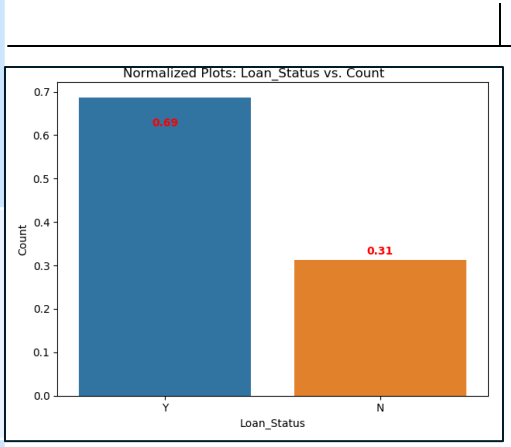
Features	Normalised
Gender	80% of the borrowers are men, 20% are women
Marital Status	65% of the borrowers are married, 35% are single
Self Employed	85% of the borrowers are self-employed, 10% are employed
Credit History	15% of the borrowers have a bad credit history, 85% have a good credit history
Loan Status	69% of the loans were approved, 32% are rejected



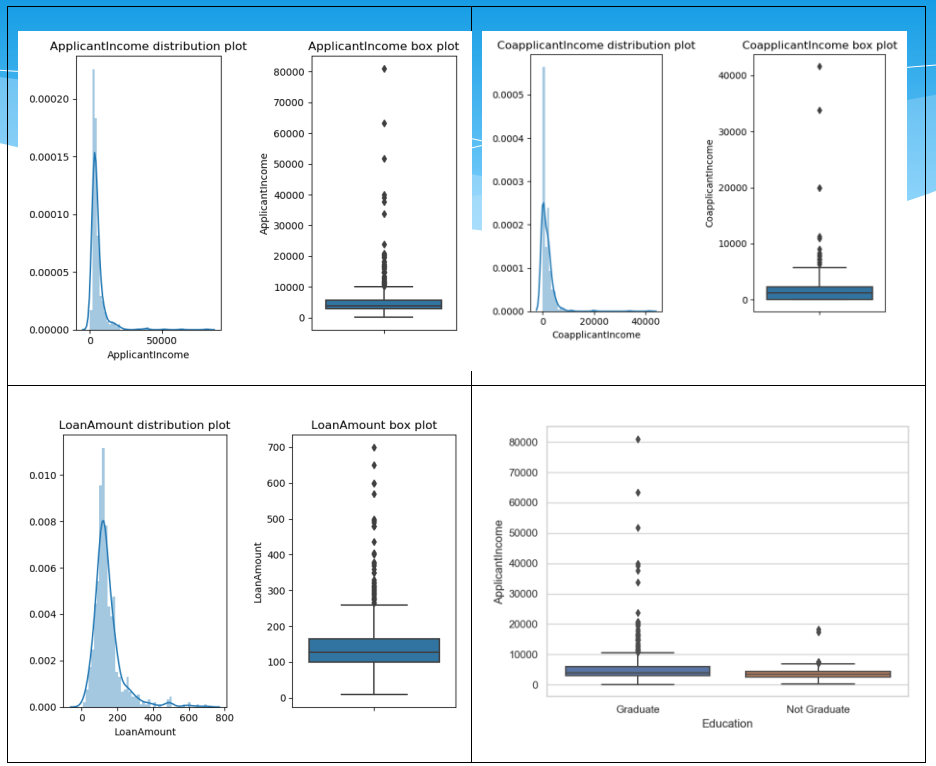






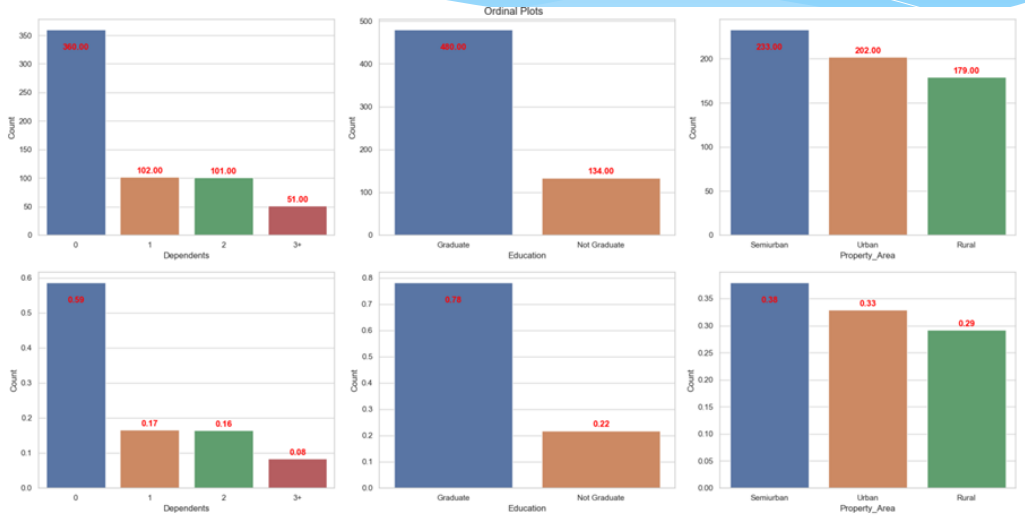


Numerical Data Analysis



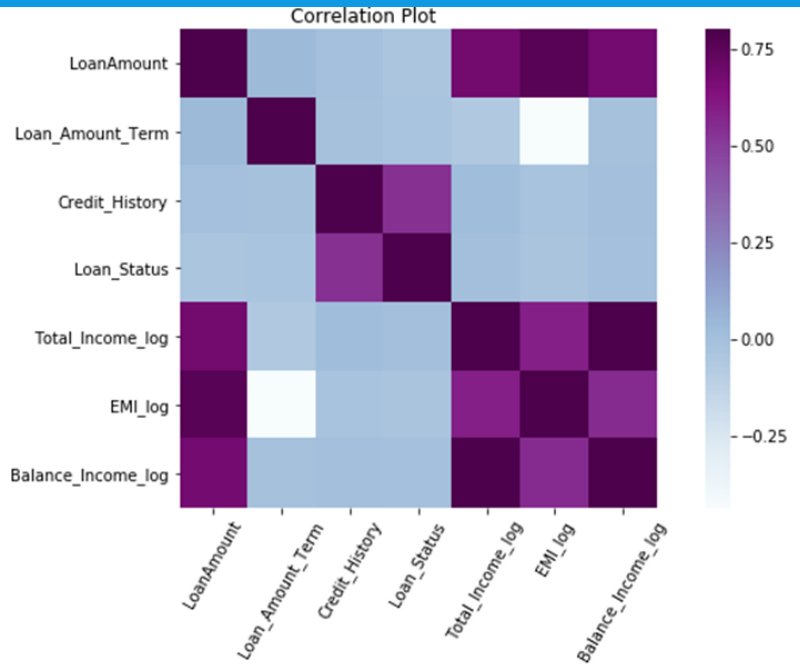
	Distribution Plot	Box Plot
Applicant Income	Not normally distributed	Presence of outliers
Co-Applicant Income	Not normally distributed	Presence of outliers
Loan Amount	Fairly normally distributed	Presence of outliers
Education		Presence of outliers. High number of graduates with very high incomes

Ordinal Data Analysis



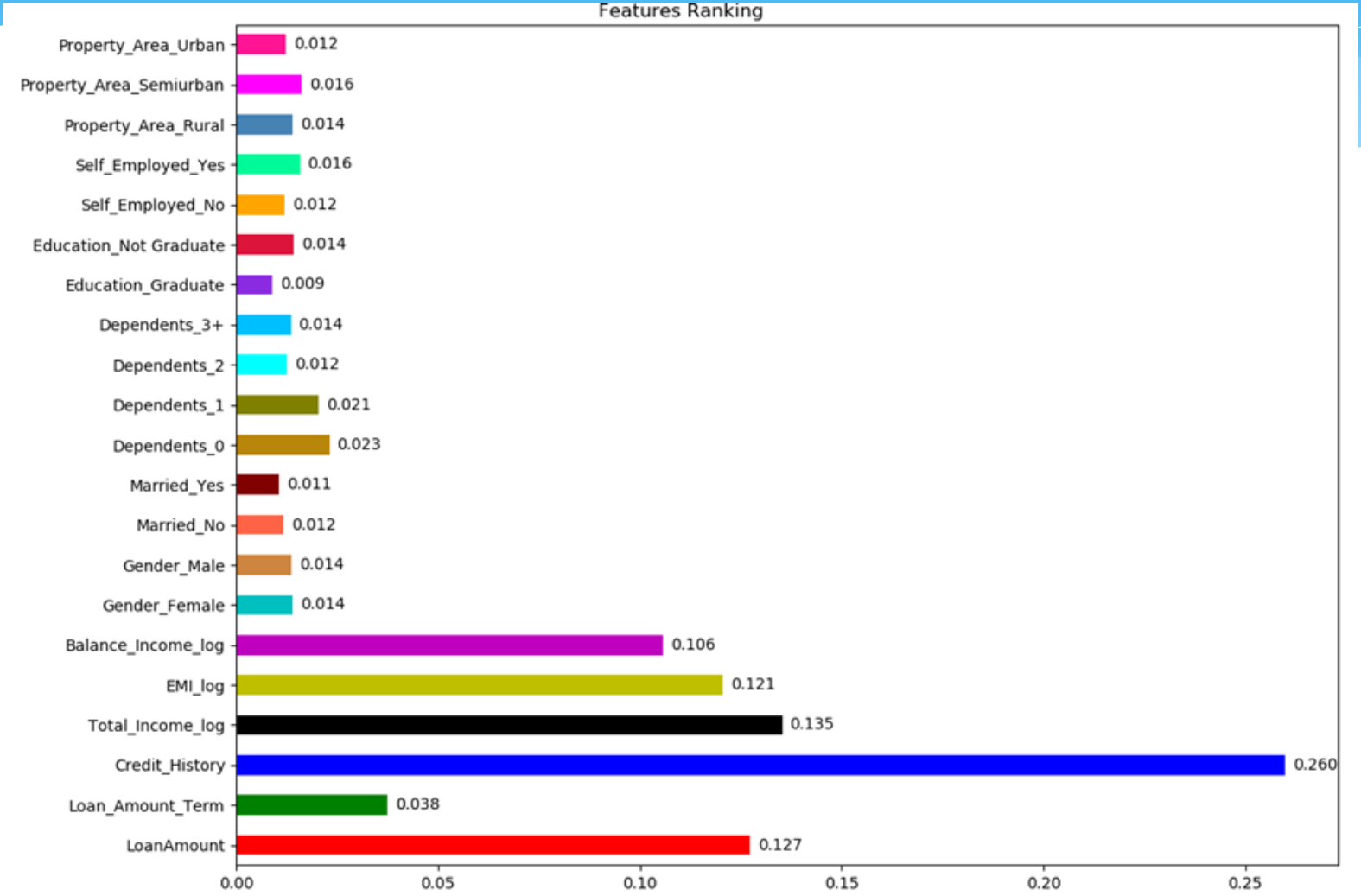
	Normalized	Un-normalized
Dependents	59% of applicants have no dependents	360 applicants have no dependents
Education	78% of applicants are graduates	480 applicants are graduates
Property	38% of applicants live in semi-urban areas	233 applicants live in semi-urban areas

Data Correlation



Features/Variables	Most Correlated Variables
Loan Amount	EMI, Total Income, Balance Income
Loan Amount Term	No strong correlation with any of the variables
Credit History	Loan Status
Loan Status	Credit History
Total Income	EMI, Balance Income
EMI	Total Income, Balance Income
Balance Income	EMI, Total Income

Feature Ranking



Front End (1)

Loan Demo

Applicant Information

First name: John Last name: Tan

Gender Type

Gender: (•) Male ( ) Female

Marital Status Type

Marital Status: ( ) Single (•) Married

Dependents Information

Number of dependents: 0

Education Type

Education: ( ) Graduate (•) Non Graduate

Employment Type

Self Employed: ( ) Yes (•) No

Property Area Type

Property Area Urban

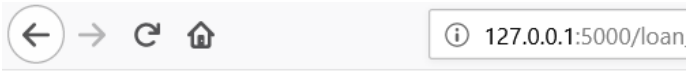
Income Information

Applicant Income (monthly): 2165 CoApplicant Income (monthly): 3422

Loan Terms

Loan Amount (in thousands): 152 Loan Term (in months): 360 Credit History (0 or 1): 1

Submit



Hello John Tan Congratulations, your Loan is APPROVED

Front End (2)

Loan Demo

Applicant Information

First name: Mary Last name: Cheok

Gender Type

Gender: ○ Male ◉ Female

Marital Status Type

Marital Status: ○ Single ◉ Married

Dependents Information

Number of dependents: 2

Education Type

Education: ◉ Graduate ○ Non Graduate

Employment Type

Self Employed: ○ Yes ◉ No

Property Area Type

Property Area Urban

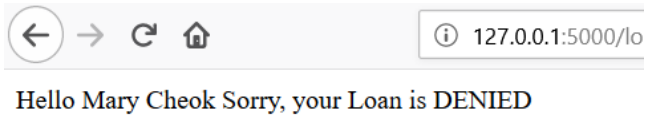
Income Information

Applicant Income (monthly): 2165 CoApplicant Income (monthly): 3422

Loan Terms

Loan Amount (in thousands): 152 Loan Term (in months): 360 Credit History (0 or 1): 0

Submit



Results

Accuracy of the different ML algorithms (using default parameters)

Algorithm Used	Accuracy
Logistic Regression	**80.79%**
Decision Tree	68.74%
Random Forest	78.18%
XGBoost	80.11%
Bagging	76.87%
Ada Boosting	80.61%
Voting Ensemble	78.21%

Results

Accuracy of the different ML algorithms (Grid Searching)

Algorithm Used	Parameters	Accuracy
Logistic Regression	'C': 1.0, 'dual': False, 'max_iter': 100	80.61%
Random Forest	**'criterion': 'entropy', 'max_depth': 5, 'max_features': 'sqrt', 'min_samples_leaf': 8, 'min_samples_split': 3, 'n_estimators': 10**	**81.11%**
XGBoosting 1	'max_depth': 3, 'min_child_weight': 3	76.06%
XGBoosting 2	'learning_rate': 0.01, 'subsample': 0.8	80.13%
XGBoosting 3	'max_depth': 3, 'min_child_weight': 5	76.55%
XGBoosting 4	'learning_rate': 0.01, 'n_estimators': 250	79.80%

Future Work

- a service hosted on the private cloud within the financial institution (because of data protection laws and client confidentiality) & the integration with financial services production back end data systems
- authorization and authentication should be enabled
- larger data sets should be used for training the model
- apart from balance income, EMI and total income, other features such as interest rate, debt-to-income ratio of the borrower (amount of debt divided by annual income), the number of days the borrower has had a credit line, the borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments) can be investigated as well.
- how to retrain and redeploy the model as new data comes in
- using a queueing framework to handle long running jobs in order not to tie up server resources