Home Loan Approval Prediction

Cherin Hamadi
Final Project

Project Statement

Metro Financial Services as fictitious company which is in the area of providing home loans wants to lower its risk of default payments. Financial organization wants to know, which customers should be given the loans for buying the home. The home loans data set has the data of 990 loans with their loan status (approved vs. not). Certain features have been defined and could have an impact of the loan status. The aim is to build up a predictive model to find out which features of the applicants could affect the loan status and in which sense.

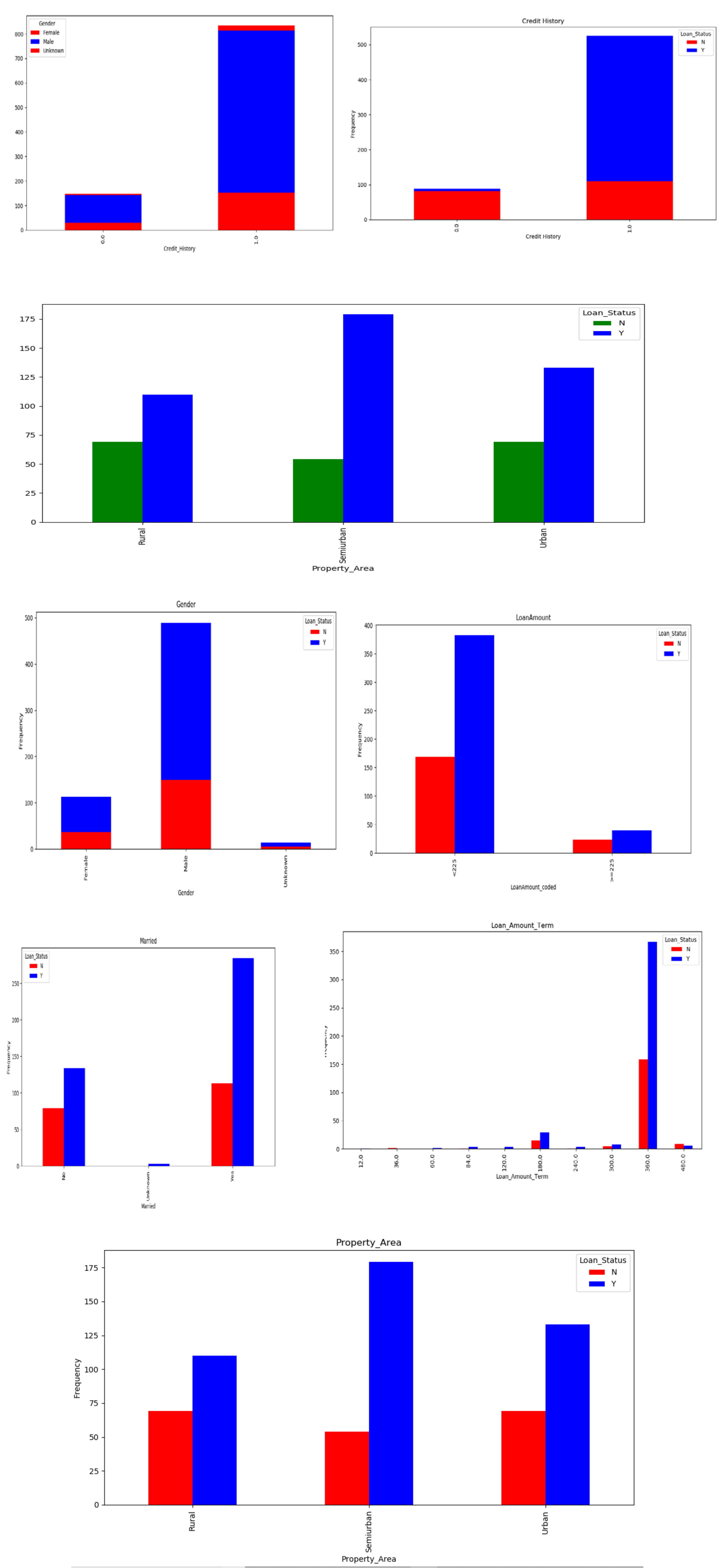
The Hypotheses

Marital Status: Single applicants are more able to have approvals as they request less loan amounts.
Dependents: Applicants who live in urban areas have fewer dependents than who live in semi-urban or rural areas.
Credit History: Applicants who have credit history are more likely to get loan approval
Education: Applicants who are more educated are more able to get approval as they have higher incomes.
Self-employment: Applicants who are self-employed are more able to get approval as they get higher income.
Co-application: Applicants whose co-applicants are working and having income are more able to have loan approval than those who are sole-applicants.
Gender: Males are more able to have approved loans as they have higher income.
The Loan Amount: Applicants who live in urban areas tend to apply for larger amounts than those living in semi-urban or rural.
Property Area: Applicants who are living in urban areas are more able to get approval as their salaries are higher.
Credit History: Applicants who have available credit history are more able to get approval than those who don't.
The Loan Amount: Applicants who have high income apply for higher loan amounts.

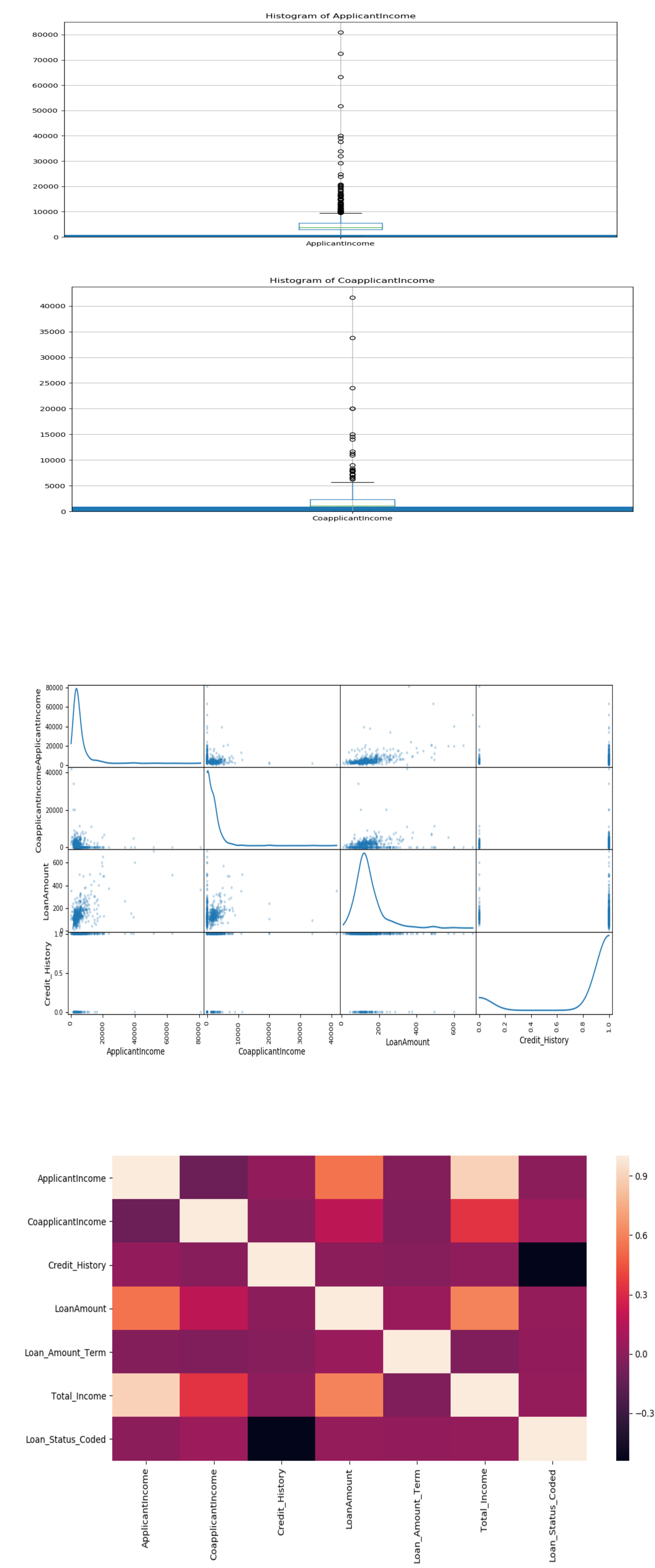
Data Cleaning & Manipulation

The mean, median, and mode were used primarily to replace missing values of some features. In some cases, missing values were replaced by "Unknown" as it found to have some impact on results. For example,
data['Gender']=data['Gender'].fillna('Unknown')
data['Married'] =data['Married'].fillna('Unknown')
data['Dependents'] =data['Dependents'].fillna('0')

Data Visualization



Statistical Insights



Machine Learning & Predication

KNN, Logit, Random Forest, & Perceptron have been used mainly to measure the correlation of the features. The two datasets (train & test) were marked initially for the training & testing stages (e.g. train['source']='train'). The two datasets were concatenated into one large dataset for pre-processing and feature engineering. Two valid ways were possible to code the features: Dummies & Encoder.

Key Findings

The loan amount requested affects positively on the loan approval based on the coefficient & level of significance.

Marriage & Credit history affect also positively on loan approvals and have significance.

Based on coding, large amount of request loan is more probable to be approved (also positive & significant).

People who are living in urban and semi-urban areas are more to be approved. (Significant & positive)

Unavailable credit history affects negatively on loan approval (significant & negative).

The best model performance was for Perceptron:
Score: 0.94
Best Accuracy: 0.82

Comparison of the ML Models Results

ML Algorithm	Accuracy %	Score	Approved	Not Approved	Total
KNN	80%	0.88	315	52	367
Logit	95%	0.90	313	54	367
Perceptron	94%	0.92	202	156	367
Random Forest (RF)	90%	0.91	278	97	367

METRO COLLEGE OF TECHNOLOGY