By: Ramsey Abdelrahman

Executive Summary

The Problem 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 981 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.

Research Question: Who can we approve for home Loan?

The Hypotheses

There were plenty of hypothesis generated from this dataset.

Service Level Hypothesis

1. **Marital Status:** Single applicants are more able to have approvals as they request less loan amounts.
2. **Education:** Applicants who are more educated are more able to get approval as they have higher incomes.
3. **Self-employment:** Applicants who are self-employed are more able to get approval as they get higher income.
4. **Co-application:** Applicants whose co-applicants are working and having income are more able to have loan approval than those who are sole-applicants.
5. **Property Area:** Applicants who are living in urban areas are more able to get approval as their salaries are higher.
6. **Credit History:** Applicants who have available credit history are more able to get approval than those who don't.
7. **The Loan Amount:** Applicants who have high income apply for higher loan amounts.

Demographic Level Hypothesis

1. **Dependents:** Applicants who live in urban areas have fewer dependents than who live in semi-urban or rural areas.
2. **Co-application:** Applicants who live in urban areas are more probable to have co-applicants than applicants in semi-urban or rural areas.
3. **Credit History:** Applicants who live in rural areas are less probable to have available credit history than the ones in urban and semi-urban areas.
4. **Marital Status:** Single applicants are more probably living in urban and semi-urban areas than rural areas.
5. **Dependents:** Singles applicants have fewer dependents than married applicants.
6. **Gender:** Males tend to be more the main applicants while the females are the co-applicants due to cultural reasons.
7. **Gender:** Males are more able to have approved loans as they have higher income.
8. **The Loan Amount:** Applicants who live in urban areas tend to apply for larger amounts than those living in semi-urban or rural.

Data Cleaning & Manipulation

To investigate the research question and validate the hypothesis, we used python on several stages to explore, examine, clean, and manipulate data to reach solid predictions. In binning stage, features have been divided within bins and categories to visualize distribution. For example: Gender_bins = [0,1,2,3], Gender_labels = {"Female", "Male", "Unknown"}. Re-Categorizing & Coding was a second stage to restructure the data. For example: bins = [0, 3800, 5516, 81000], labels = ['0-3800', '3801-5500', '5501-81000'], data['ApplicantIncome_Coded'] = pd.cut(data.ApplicantIncome, bins=[0, 3800, 5500, 81000], labels = labels, include_lowest=True), data['ApplicantIncome_Coded'] = coding(data['ApplicantIncome_Coded'], {'0-3800': 0, '3801-5500': 1, '5501-81000': 2}).

Machine Learning & Predicition

As the test set is the target to predict, 4 main classification models have be operated to choose best accuracy and performance: KNN, Logit, Random Forest, & Perceptron. SVM & MLP have been also explored additionally. The two datasets (train & test) were marked initially for the training & testing stages (e.g. train['Trainsource']='train'). The two datasets were concatenated into one large dataset for pre-processing and feature engineering. The two datasets, then, were separated based on the marks created at the begining. Four objects were created based on the required features (predictors) to predict the target: X_train, X_test, y_train, y_test. As the target is (y_test), each model was run twice. The first time is to predict the y_test and use it as y_true in the second time for performance measuring and calculating accuracies during cross-validtion. Gridsearch and hyperparameter were also ran eventually for performance improvement.

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 Random Forest:

- Score: 0.94 Best Accuracy: 0.85

Comparison of the ML Models Results					
ML Algorithm	**Accuracy %**	**Score**	**Approved**	**Not Approved**	**Total**
KNN	**79%**	**0.81**	**320**	**47**	**367**
Logit	**83%**	**0.83**	**308**	**59**	**367**
Perceptron	**81%**	**0.78**	**212**	**155**	**367**
Random Forest (RF)	**85%**	**0.94**	**274**	**93**	**367**