Project 4: Logistic Regression
By: Mohak Kant, Don Pathirage, Dipendra Yadav

3. a) Validation of implementation using synthetic data.

Generated synthetic data from Gaussian distribution having two classes.

Class C1: mean = 1, variance = 0.11 (Multivariate Normal)   Total number of data points = 1200
Class C2: mean = 1.6, variance = 0.12 (Multivariate Normal)   Training: Testing = 1000:200

Number of Epochs: 1000, learning_rate = 0.01, Batch Size =100, Momentum = 0.99
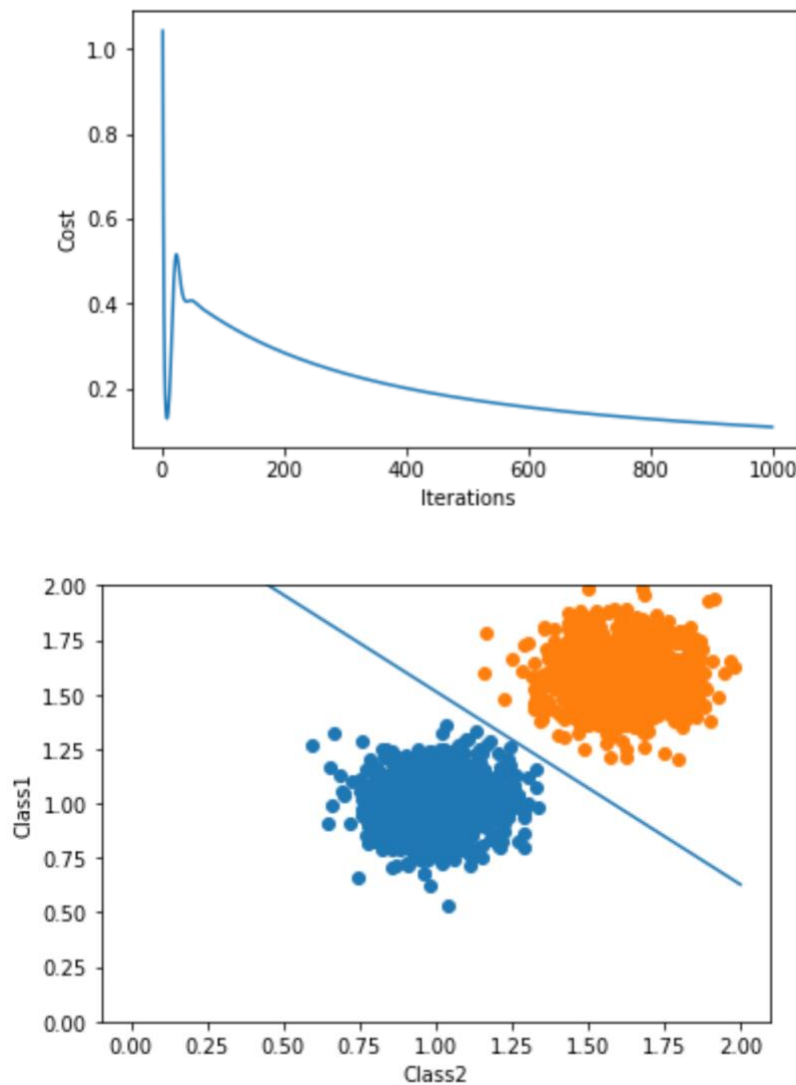

*Figure 1 Logistic Regression on Synthetic Data, Plot of Cost Function vs Iteration and Decision Boundary*

3. b) Description of Selected Dataset

Dataset: Loan Prediction data provided by Analytics Vidhya

Determining loan eligibility can be a lengthy, time-consuming process. We would like to generate a model to predict loan eligibility using applicant features such as gender, marital status, education, number of dependents, income, loan amount, credit history, etc. The selected dataset contains 614 such loan eligibility applications.

Univariate Analysis:

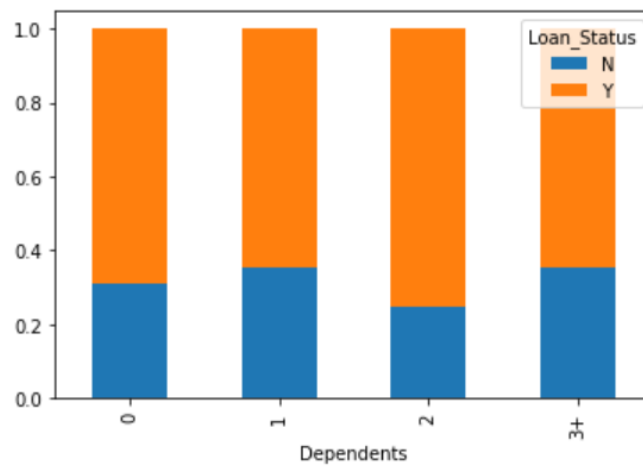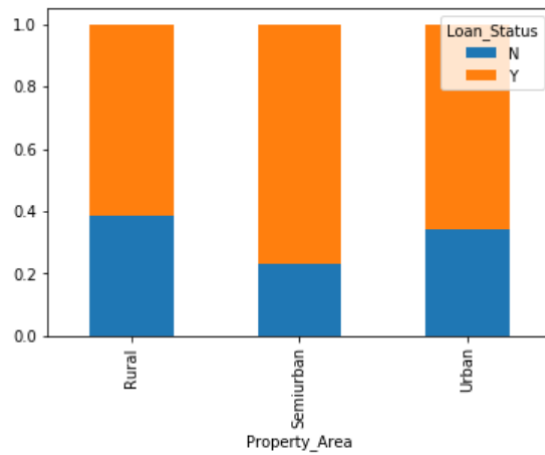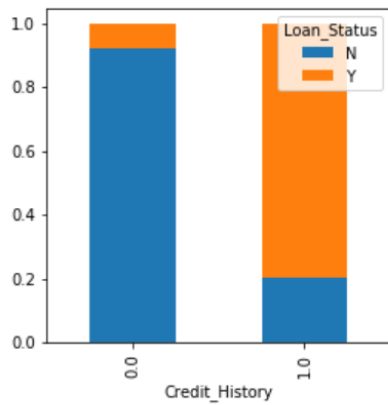- Loans of 422 out of 614 people were approved, an approval rate of ~69%.
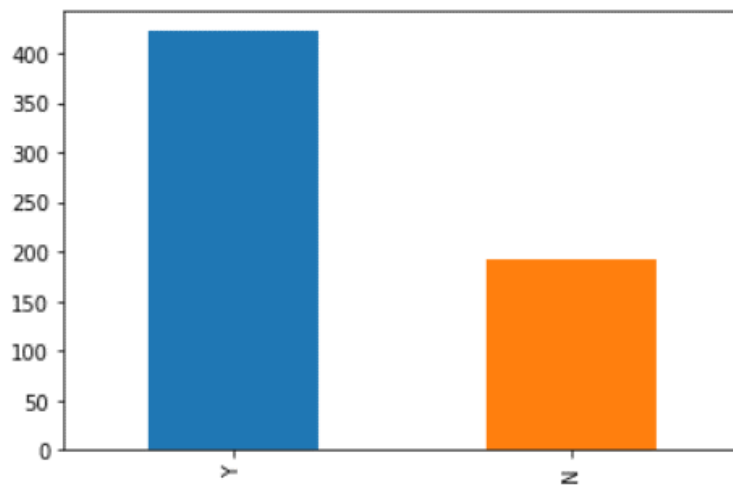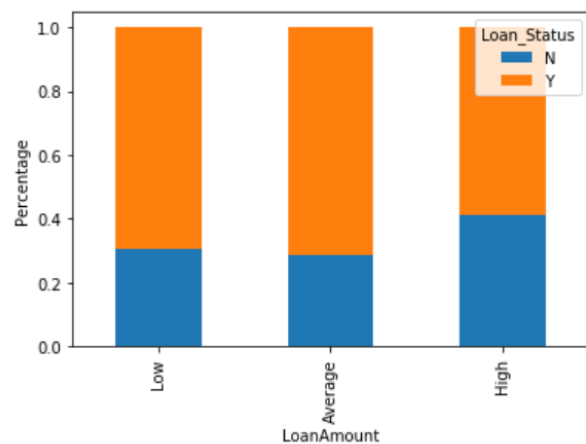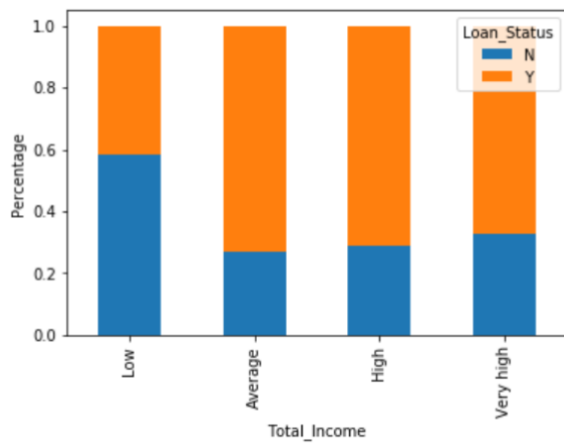

Bivariate Analysis:
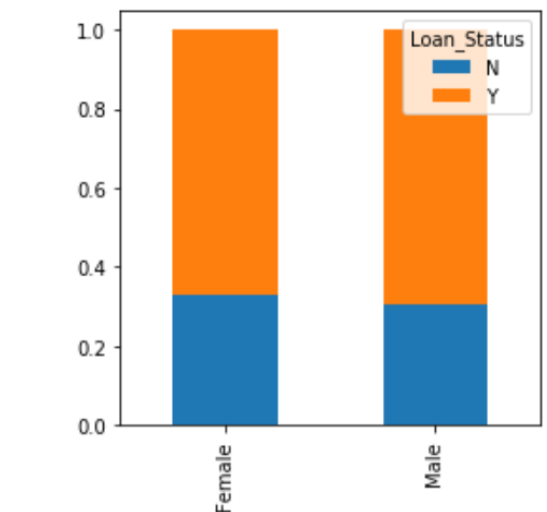
Inferences:
- Gender does not appear to affect approval rate.
- Higher approval rate for married applicants.

Inferences:
- Applicants with credit history have a higher approval rate.
- Approval rate in semi-urban areas is greater than that of urban and rural areas.
- Distribution of loan status is similar for 1 or 3+ dependents.

Inferences:

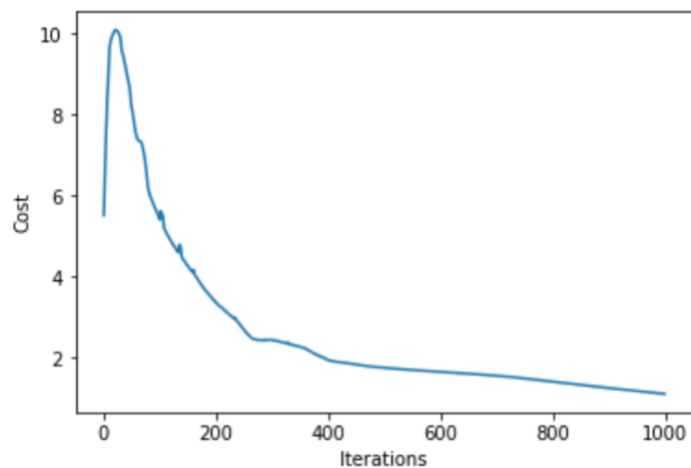- Loan approval for applicants having low total income is far less than that of applicants with average, high or very high incomes.
- Approval is higher for low and average loan amounts than high loan amounts.

Correlation Heat Map:
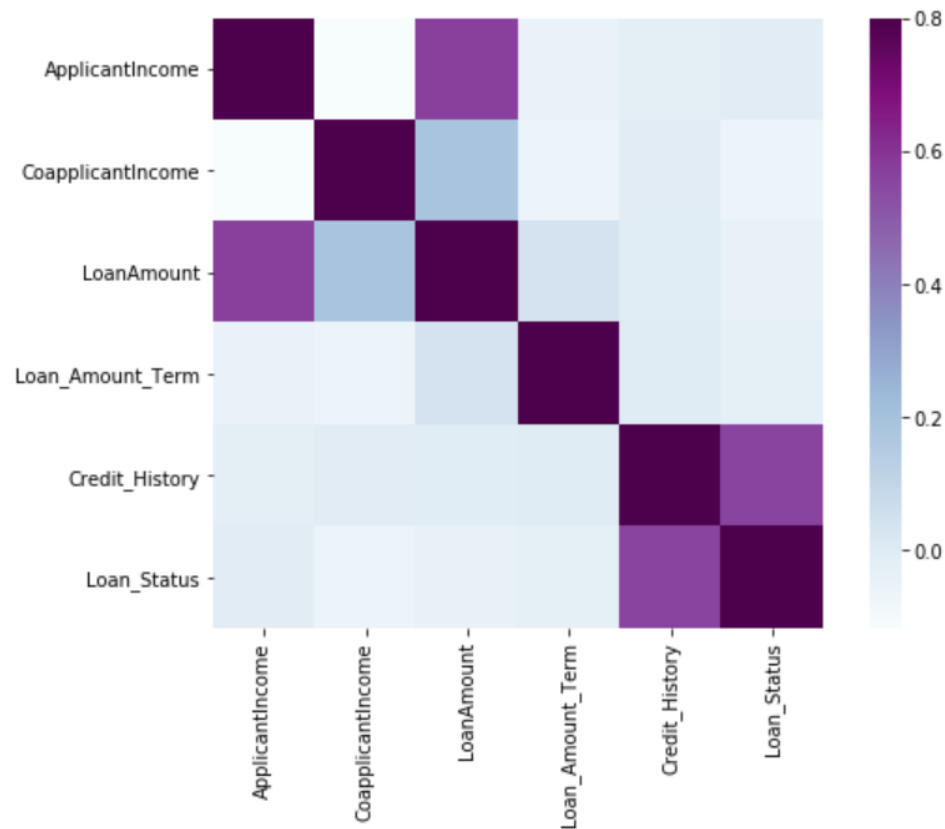
3. c) & d) Performance on Selected Dataset

Number of Epochs: 1000, learning_rate = 0.01, Mini Batch = 100, Momentum = 0.99, Polynomial Map of Degree = 2

Our Accuracy:

Training-Testing Split Ratio	Training Accuracy	Testing Accuracy
90-10	83.69%	82.73%

Our Training and Testing Time:

Training-Testing Split Ratio	Training Time	Testing Time
90-10	68.01 seconds	0.03 seconds

Benchmark Accuracy:

Training-Testing Split Ratio	Training Accuracy	Testing Accuracy
90-10	98.62%	100.00%

Benchmark Training and Testing Time:

Training-Testing Split Ratio	Training Time	Testing Time
90-10	0.023 seconds	0.001 seconds

Works Cited

[1] Analytics Vidhya, "Loan Prediction," Analytics Vidhya, [Online]. Available:
https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/.