Predicting the Likelihood of a Loan Application being approved

T. Hastings Reeves

•••

Data Science Intensive Capstone Project, May 2021 Cohort

Thanks to Springboard Mentor:
AJ Sanchez, PH.D., President of Exodus Software Services, INC.

Dream Housing Finance Company

- Current process of approval for loan applications:
  - Weigh specific criteria found in the application
  - Approve or deny loan request, individually.

Understanding the problem

Loan Application

- Key Factors
  - Loan ID
  - Gender
  - Married
  - Education
  - Income
  - Loan Amount
  - Loan Term
  - Property Area

Individual review

- Credit History, Applicant and Co-Applicant Income, and Loan Amount are key variables in helping loan officers weigh the criteria.

Approval or Denial

- Based on individual loan application.
- Case by case basis, from beginning of process to end.
- Time-consuming.

Project objective:

Utilize machine learning to expedite the beginning level decision making process for loan approval.

Data Acquisition & Wrangling

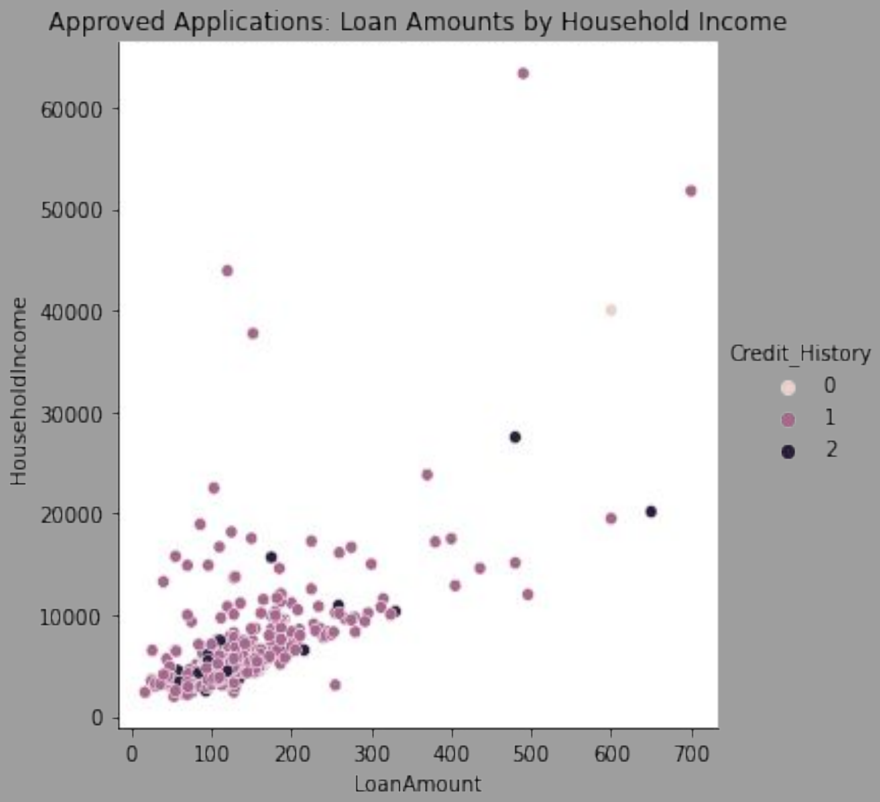
The data was supplied by Dream Housing Finance Company:

- 614 loan applications
- 13 categories of information
- Target: Loan Approval

Data Exploration:

Credit History:

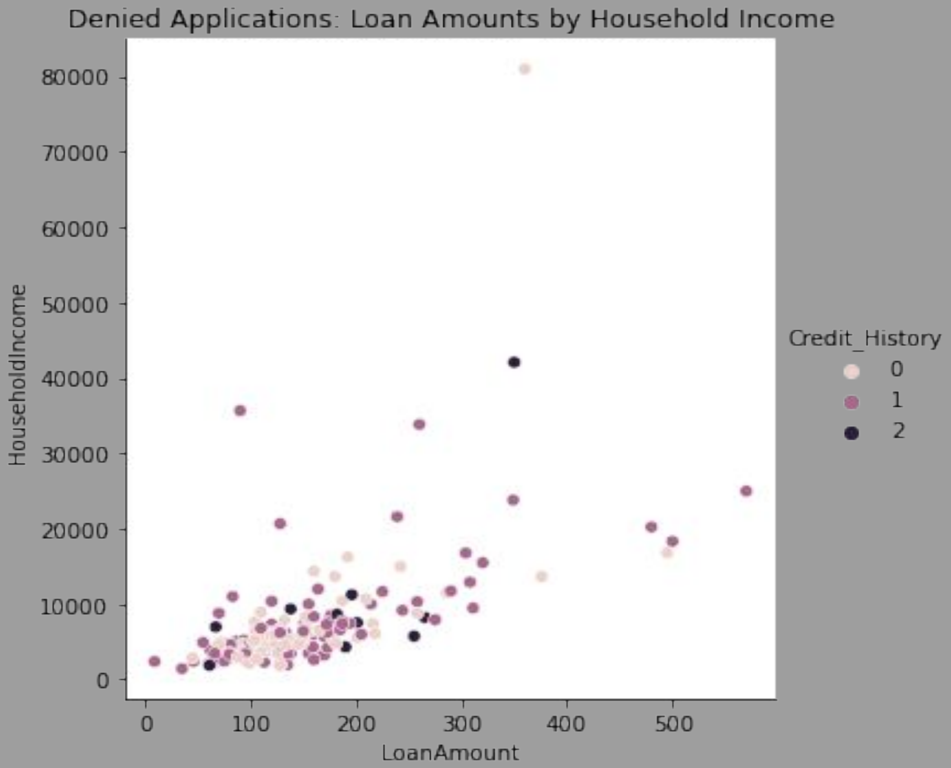
- Very few applications that were approved did not meet the Credit History Requirement.
- Some earners' income at $40,000/month did not meet the Credit History Requirements but was approved.



Data Exploration:

Credit History:

- The majority of the applications did not meet the Credit History Requirements were denied.
- One earners' income was $80,000/month, yet they did not meet the Credit History Requirement, and were denied.



Machine Learning

Baseline Model

- Type: Supervised Learning
- Binary Classification: 1 for Approved, 0 for Denied
- Imbalanced data: Nearly a 3:1 ratio from Class 1 to Class 0.
- Tools: Scikit Learn and imblearn
- Algorithm: Logistic Regression

Modeling Steps:

Pre-Processing:	• Train/Test split at 70/30, with stratify condition set to ‘y’. • pd.get_dummies for label encoding
Performance Metrics:	• Accuracy Report • Classification Report, AUC and ROC
Resampling:	• Random Undersampling, Random Oversampling, SMOTE • Baseline Model, Random Forest, Decision Tree
Best Models:	• Random Forest-SMOTE, Logistic Regression-SMOTE

Model Comparisons

MODEL:	Recall: Minority	Precision: Minority	F1: Minority	Recall: Majority	Precision: Majority	F1: Majority	AUC
Baseline Model: Logistic Regression(Training data)	0.42	0.87	0.56	0.97	0.79	0.87	0.744
Logistic Regression: Random Oversampling	0.55	0.27	0.36	0.31	0.6	0.41	0.738
Logistic Regression: Random Undersampling	0.6	0.49	0.54	0.71	0.8	0.75	0.741
**Logistic Regression: SMOTE	0.53	0.55	0.54	0.8	0.79	0.8	0.745
**Random Forest: Random Oversampling	0.53	0.65	0.58	0.87	0.8	0.83	0.744
Random Forest: Random Undersampling	0.62	0.52	0.57	0.74	0.81	0.77	0.738
*Random Forest: SMOTE	0.53	0.66	0.59	0.87	0.8	0.84	0.749
Decision Tree: Random Oversampling	0.55	0.55	0.55	0.8	0.8	0.8	0.674
Decision Tree: Random Undersampling	0.55	0.39	0.46	0.61	0.75	0.67	0.579
Decision Tree: SMOTE	0.55	0.52	0.54	0.77	0.79	0.78	0.662

- *Best Model
- **High Performing Models
- Baseline Model uses training data due to overfitting on Test set CR.

Loan Application Approval Or Denial

Findings

While we were not able to conclusively classify 1 or 0 we can set parameters to determine 4 sets of categories to group loans into:

1. **Definitely Approved**
2. **Tentatively Approved**
3. **Tentatively Denied**
4. **Definitely Denied**

Future Work:

1. Ensemble Methods utilizing Logistic Regression and Random Forest.
2. Clustering Algorithm to develop new insights.
3. Creating and finding evaluating the performance of multiple Random Undersampled models.