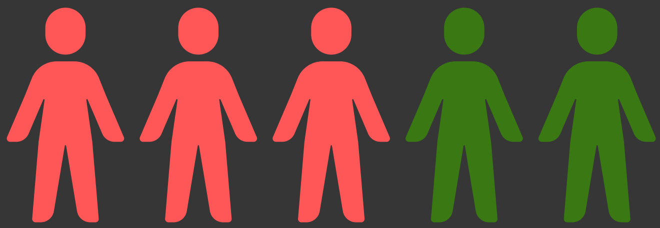
JANUARY 2022

$ERA

Customer Default Prediction Model Report

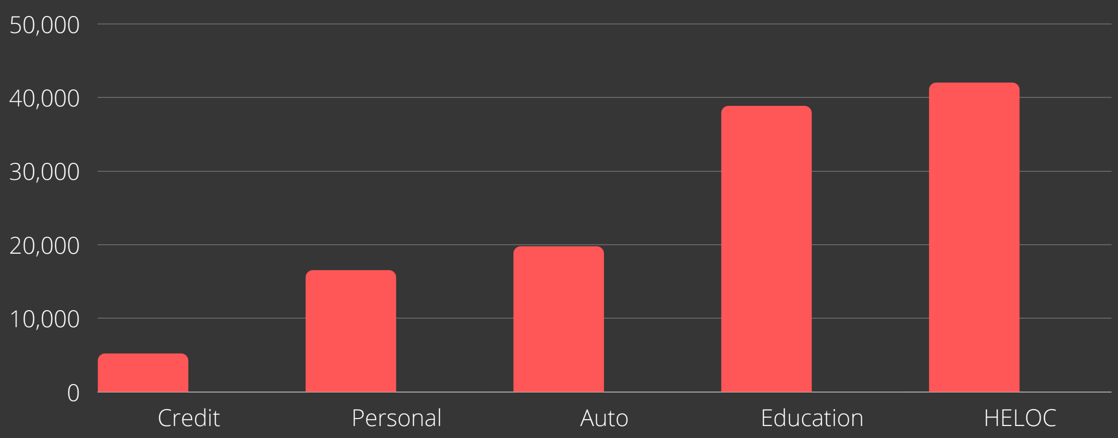
Prepared by Justin Jimenez

Background



According to a recent report from Experian, the average individual debt in America is $92,727. With that amount of debt, it is no surprise that approximately 62% of middle-class Americans report that they struggle with paying down consumer debt. All over America people are drowning in debt for their cars, their homes, their credit cards, and their degrees:

Average Debt ($) by Category



***Data reported from Experian**

$ERA Objective

$ERA's objective is to provide users with the tools and services that they need to pay down their debt quicker and achieve financial freedom. $ERA is creating a new automated financial service called Augmented Debt Reduction, which utilizes Lines of Credit to pay down high-interest debt quickly. The company is aiming to release the product in Q2 of 2022, but they need a process for evaluating the ideal customer for this new product. $ERA's mission is to help people achieve financial freedom, so a process to identify potential customers that are likely to succeed with ADR is essential to that mission.

Problem Scope Definition

$ERA needs a classification model that will be able to classify potential new users into "likely to default" and "not likely to default". The model must be interpretable to adhere to compliance regulations and audits, it cannot utilize features that can only be obtained from a hard credit check, and it must minimize the number of incorrectly classified bad borrowers. The model must be ready for deployment by Q2 2022. The key stakeholders are the CEO and CTO of $ERA.

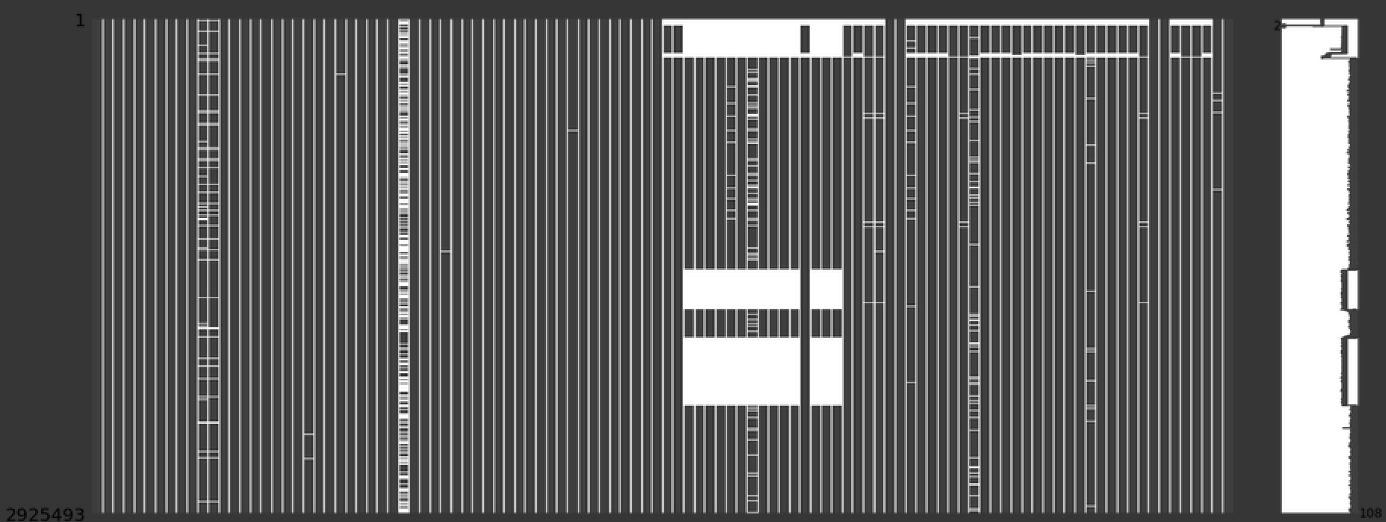
1. CLASSIFY USERS LIKELY TO DEFAULT
2. EASILY INTERPRETED
3. NO FEATURES FROM CREDIT CHECKS
4. MAXIMIZE IDENTIFICATION OF BAD BORROWERS
5. READY FOR DEPLOYMENT Q2 2022

Data Wrangling

$ERA does not currently have access to LOC data, so a substitute dataset was used. LendingClub services Peer-to-Peer (P2P) loans, which have similar interest rates and are used for similar purposes as that of the LOCs that $ERA plans to provide, so this project utilized P2P data in place of LOC data.

The P2P dataset from LendingClub contained 2.9 million loans from 2007 to Q4 2018 and had 142 features. The data was sourced from Kaggle. Some of these loans were paid off or charged off, and some were current when the dataset was last updated.

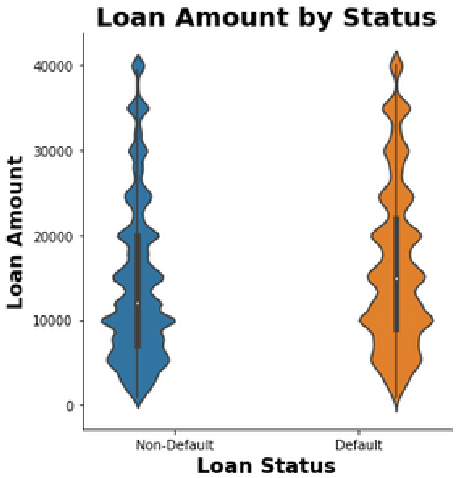
Many features were missing values for large amounts of observations. The features missing more than 60% of values were dropped. The missingness matrix of the remaining data suggests that about 29.6% of observations were missing several features at random (the large white blocks in the matrix).

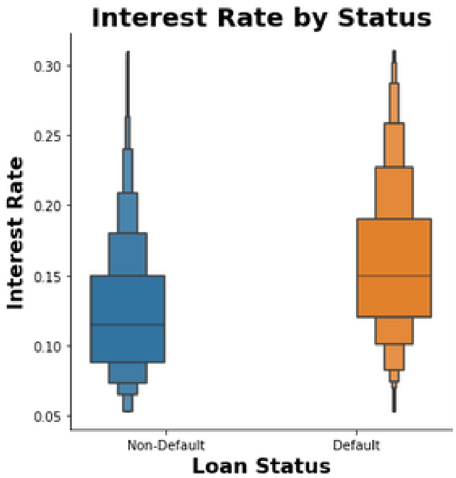


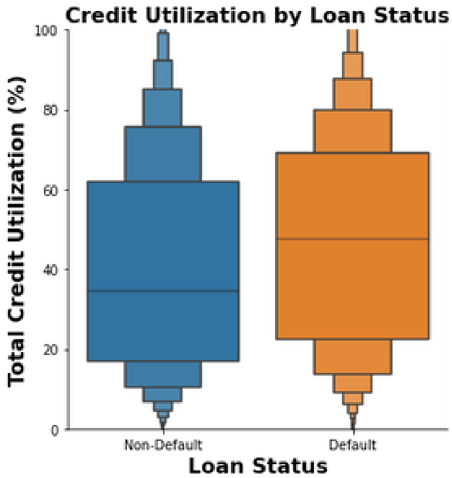
From this culled set, observations that were missing less than 4% in the remaining features were also dropped along with the data that was missing at random. Outliers were then filtered out with IQR filtering, and the dataset was checked for duplicated observations before imputing the last missing values with the median of each feature. Lastly, the data types were changed to reflect the nature of the features (discrete, continuous). The dataset then had 102 features and 1.7 million loans.

Exploratory Data Analysis 1

The EDA notebook contains extensive visualization of each feature by loan status and loan purpose. The difference in statistics for each feature was depicted with contingency tables and tested with either a chi-squared test or t-test for statistical significance. All but 2 features were found to have statistically significant differences between loan statuses.

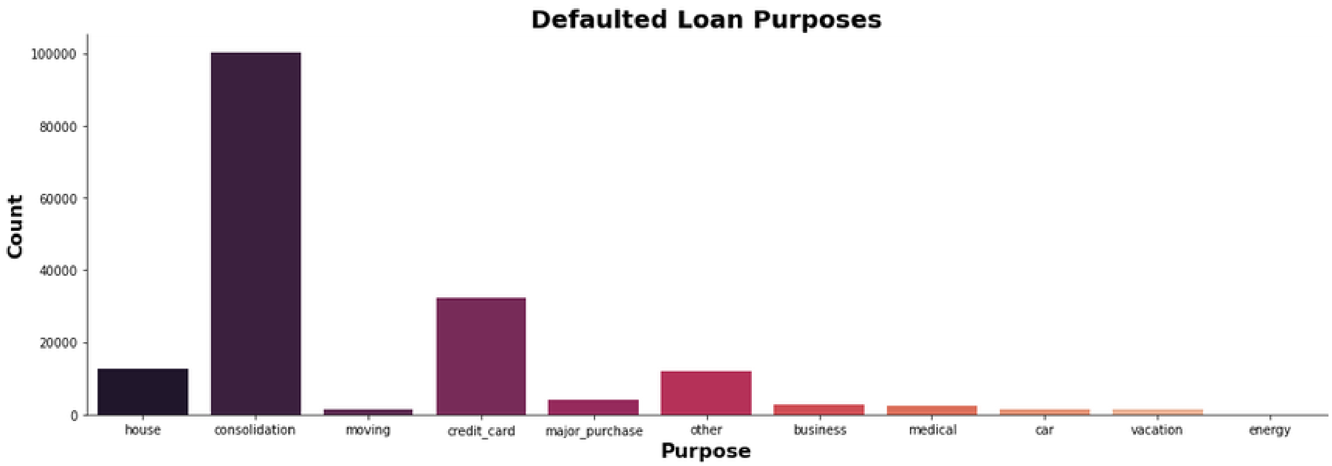




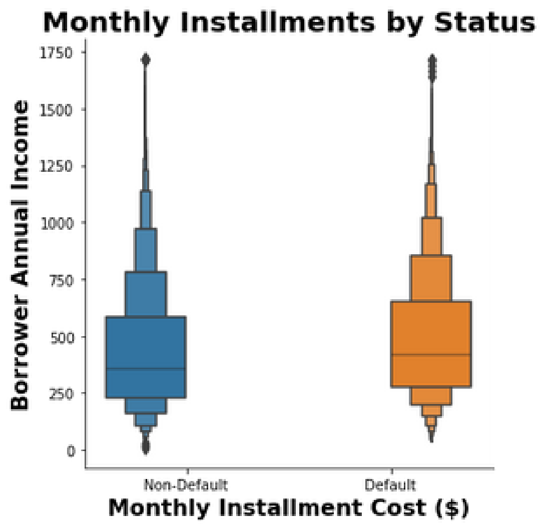


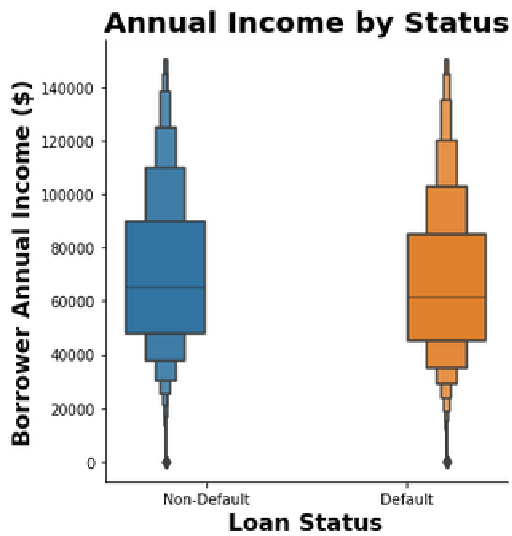
There were many trends in the characteristics of defaulted borrowers. Loan amount (ranging from $1,000 to $40,000), interest rate (ranging from 5% to 30%), and total utilization rate were all found to be greater among defaulted borrowers.

Defaulted borrowers also tended to take loans out for debt consolidation or credit card payments rather than other purposes like businesses or vacations.



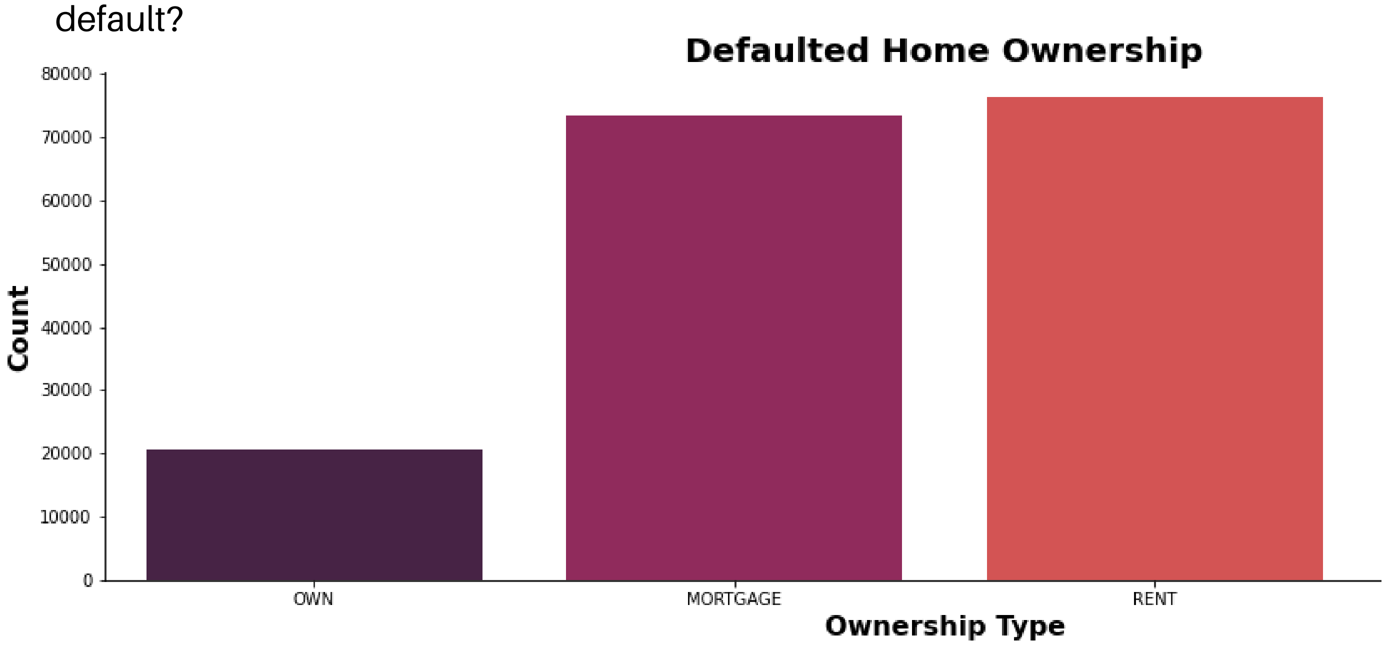
Exploratory Data Analysis 2





Defaulted borrowers also tended to pay higher monthly installment costs and earn less annually ($6,000 less on average). Most non-defaulted borrowers rent or mortgage their home, whereas defaulted borrowers rent in greater proportion.

Other interesting trends among defaulted borrowers include lower credit limits, younger credit, more recent activity, more bankruptcies, and more chargeoffs. Although some of these characteristics are obvious, it begs an important question: How much do these features contribute to an individual's liklihood of default?

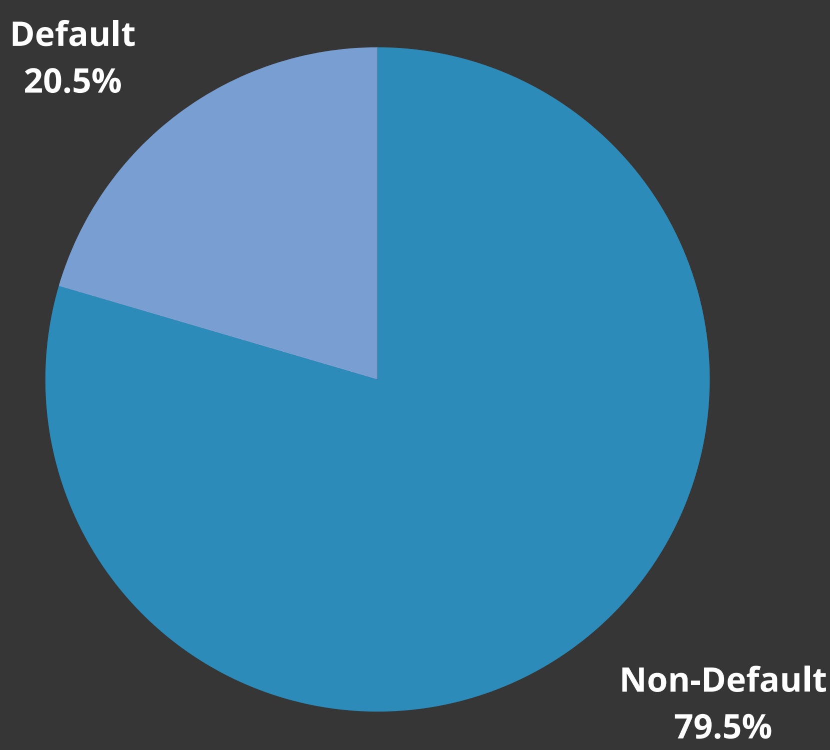


Preprocessing

In EDA, the loans that were not either paid off or charged off were dropped from the dataset along with some features that were irrelevant or too difficult to clean (like employment title). Going into the preprocessing phase, the dataset was at 71 features and 836,000 loans.

Because the model is intended to make predictions on data that users provide on a form, many features that can only be obtained from credit checks were dropped, and surrogate features for properties like DTI and utilization rate were engineered from features that a user could realistically provide. The resultant dataset had 23 features.

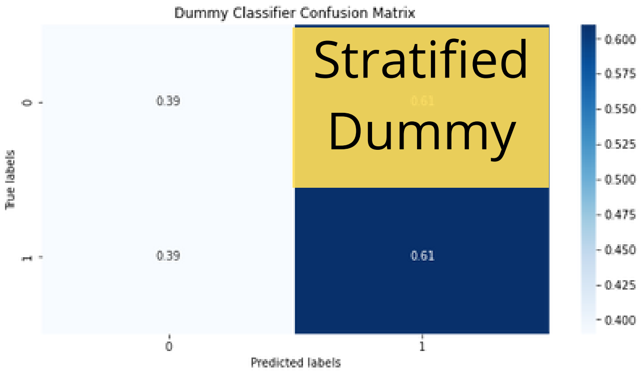
These features were inspected for multicollinearity, one hot encoded, and split into 80% train and 20% test sets.

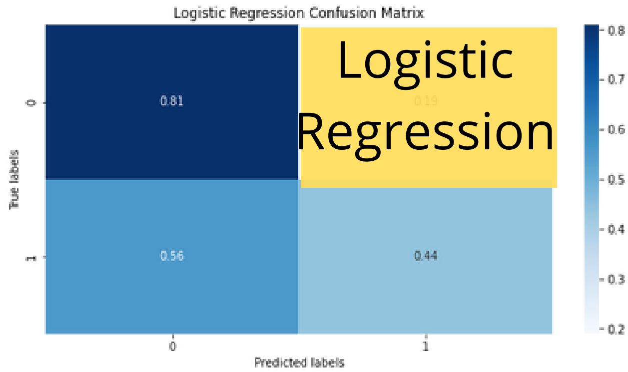


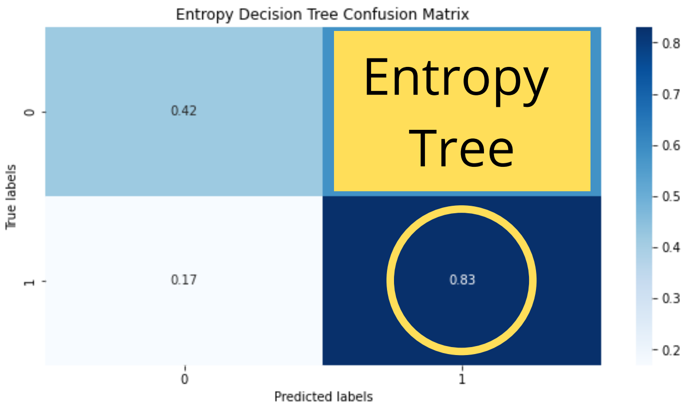
The dataset was significantly imbalanced between classes (1:4) and simple Logistic Regression models scored 0.01 recall on the test set.

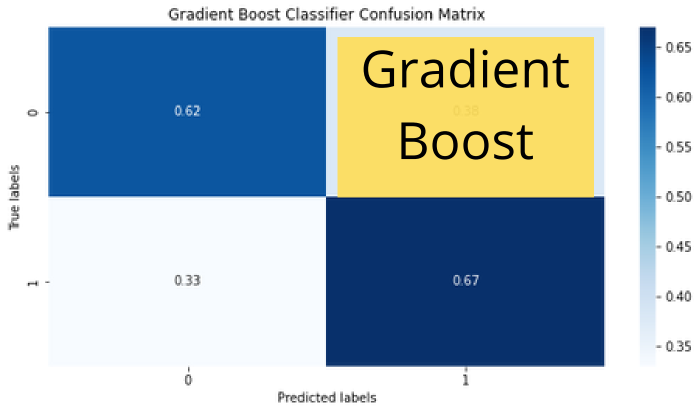
Therefore, the data was resampled using a SMOTE Edited Nearest Neighbor sampler. Lastly, a standardized scaler was fit to the train set and transformed both the train and test sets.

Model Training







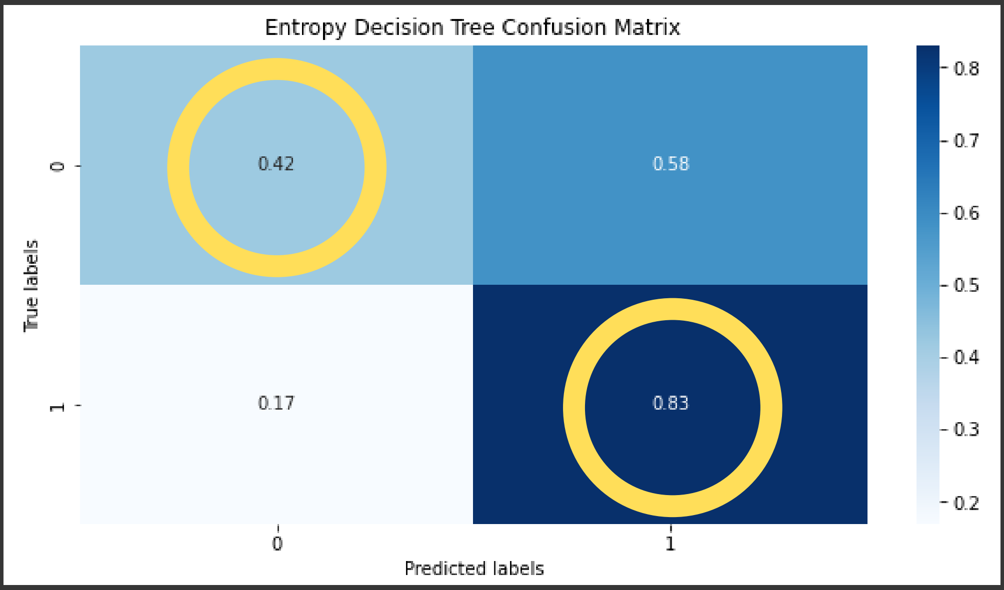


I decided to develop four models: Logistic Regression, Decision Tree (Gini Impurity and Entropy), and Gradient Boosting Classifier. I selected these models with the constraint of model interpretability in mind. A stratified dummy classifier was used as a baseline comparison.

All of the models were scored on recall (with consideration for specificity and precision) because of the scope requirement of maximizing correctly classified bad borrowers. A small increase in recall can amount to millions of dollars, but it should not come at a significant expense to the specificity of non-default predictions. The models were optimized with fivefold randomized search cross validation.

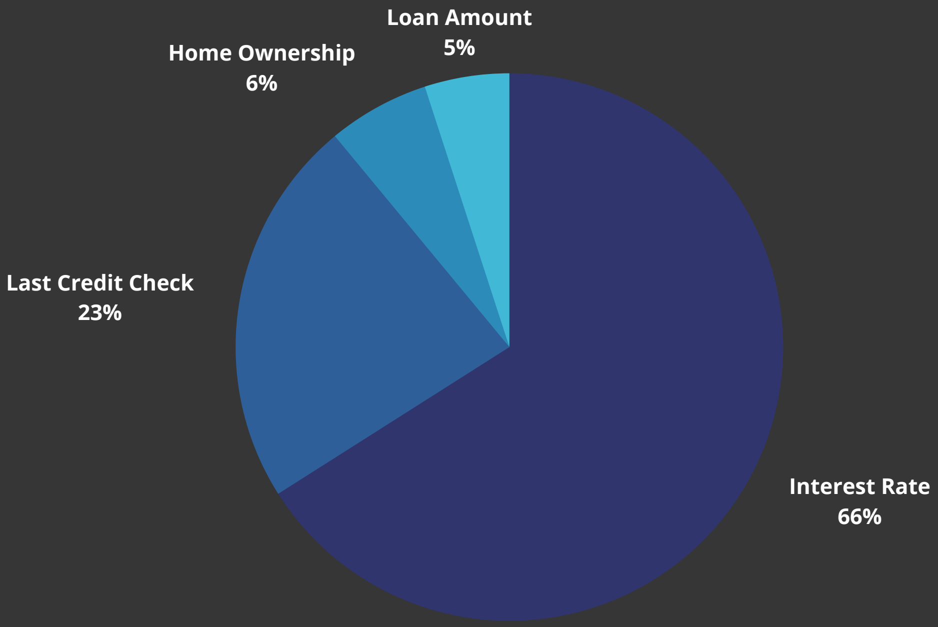
All four of the models showed a significant discrepancy between training and testing performance. For example, the entropy decision tree scored a 91% in recall on the training set, and an 83% on the testing set.

Model Evaluation



The entropy decision tree achieved the greatest recall score (83%), about 22% higher than the dummy classifier. However, this did come at some expense to specificity (42%) and precision (27%). The other models performed similarly in these two metrics, suggesting that it is characteristic of the nature of this problem.

Although the gradient boost classifier strikes a nice balance between recall (67%) and specificity (62%), I selected the entropy decision tree for its simplicity.



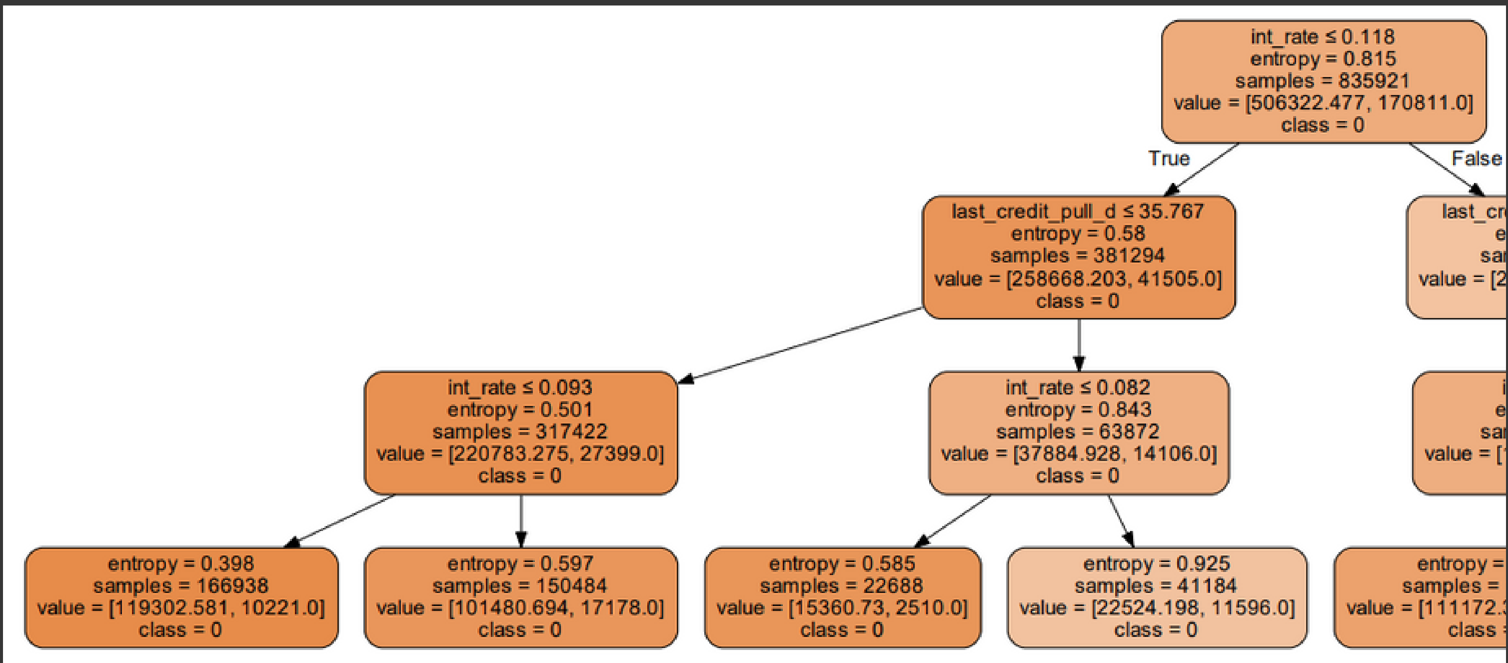
The optimized entropy decision tree only considers four features, depicted in the pie chart (left).

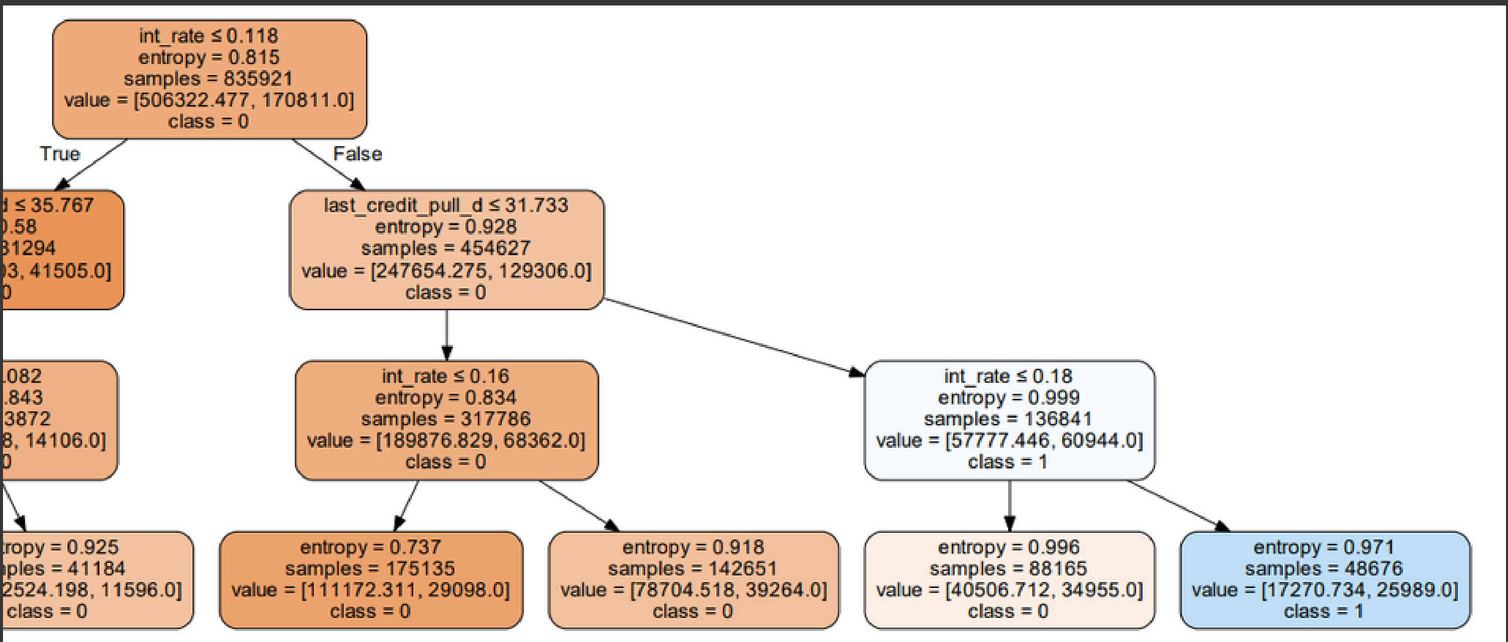
Hyperparameters:

- Max Depth of 3 Branches
- 0.76:1 Class Weight (Non-Default:Default)
- Min Samples Split = 46
- Min Samples Leaf = 2

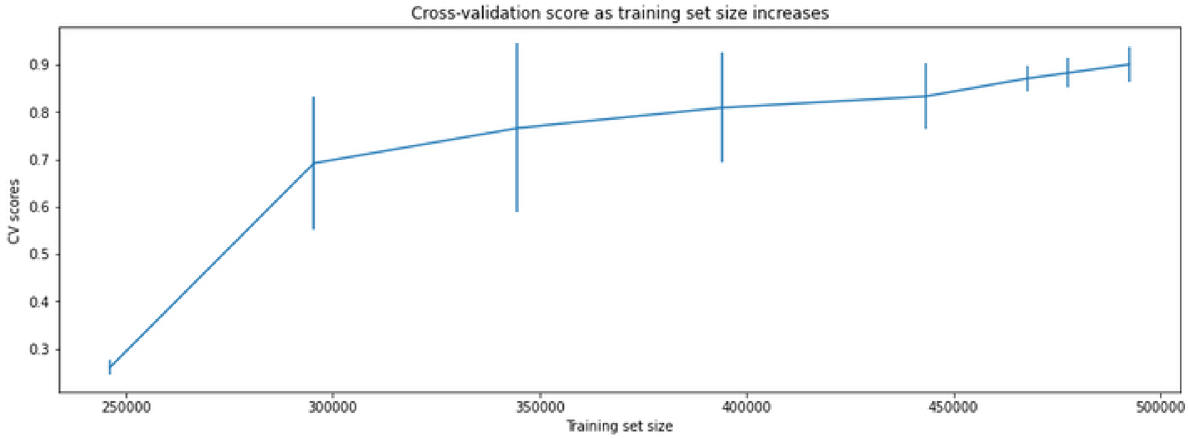
Model Visualization

The features depicted on the previous page are the features that were considered when the entropy decision tree was trained on 80% of the data. After refitting to the entire set of completed loans, the feature importance was updated to 63% for interest rate and 37% for credit pull date. Loan amount and home ownership were no longer used as predictors. The following visualization represents an entropy decision tree with the optimized hyperparameters and trained on the full dataset.





Data Quantity Assessment



The plot above shows the incremental increase in training recall scores as dataset size increases. It suggests that we may see marginal improvements to training recall with more data. However, due to the discrepancy between training and testing scores, it is difficult to tell how substantial the improvement to testing scores will be with additional training data.

Improvements

My first recommendation would be to attempt hyperparameter optimization with a Bayesian optimizer rather than randomized search. Grid search CV is too computationally expensive for a dataset of this size, and randomized CV produced inconsistent results. Bayesian optimization may be a better option.

The results might also improve with feature engineering such as Deep Feature Synthesis. One of the most important features that was dropped before training was the total payment received. Obviously, leaving this feature in would result in leakage since borrowers who paid more than the loan amount are likely to be fully paid off and vice-versa. It is possible to filter out borrowers who paid more than the loan amount but doing so would overfit the model to defaulted loans and create poor performance for non-default predictions. In practice, collecting data on percentage of debt paid off may be a critical feature for improvements.

Implementation

This model is intended for deployment and integration with a web app to enhance the lead acquisition experience. In practice, the web app should request user contact information and other information that can be parsed into the necessary features for prediction. The model would be integrated into the API and the data would be stored in a database. After the user completes the form, they will be able to view their estimated savings (from the calculator) and they will receive an approval or rejection based on the prediction results.

A separate log and dashboard would be useful for monitor model runtime performance. The model could be updated with a batch learning process when the dashboard indicates a drop in performance below a certain threshold.

Lastly, this model and the associated dataset could serve as the foundation for other tools and models related to predicting the interest rate an applicant will receive.

IMPLEMENTATION PLAN PROPOSAL

5-Step Implementation Plan