Short Term Credit Default

Ben Miller, William Brickowski, Spencer Hanson, and Max Moghadam

Project Goals

- Can we develop a way to predict credit default?
- What attributes make it likely that a loan will default?
- What geographical regions are more prone to default?

Data preparation work

- Data cleaning
    - Useless attributes
    - Omitted Data
- Merging datasets
    - Merging on the basis of loan sequence number
- Encoding data for particular classification techniques
    - SciKit Learn LabelEncoder to convert to numeric values

Tools used

- iPython Jupyter Notebook
- Python
- SciKit Learn
- Pandas
- Numpy
- MongoDB

Logit Regression

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                default   No. Observations:               208432
Model:                          Logit   Df Residuals:                   208424
Method:                           MLE   Df Model:                            7
Date:                Wed, 25 Apr 2018   Pseudo R-squ.:                  0.1056
Time:                        18:42:10   Log-Likelihood:                -25741.
converged:                       True   LL-Null:                       -28780.
                                        LLR p-value:                     0.000
=============================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------
betaNot                      -0.3665      0.265     -1.385      0.166      -0.885       0.152
1. creditScore               -0.0119      0.000    -47.408      0.000      -0.012      -0.011
10. DebtToIncome              0.0254      0.001     22.927      0.000       0.023       0.028
12. OriginalLoanToValue       0.0167      0.001     17.640      0.000       0.015       0.019
13. OrigninalInterestRate     0.5564      0.022     25.269      0.000       0.513       0.600
TX                           -0.4200      0.059     -7.174      0.000      -0.535      -0.305
FL                            0.3946      0.050      7.908      0.000       0.297       0.492
CA                            0.0827      0.039      2.116      0.034       0.006       0.159
=============================================================================================
```

Gaussian Naive Bayes

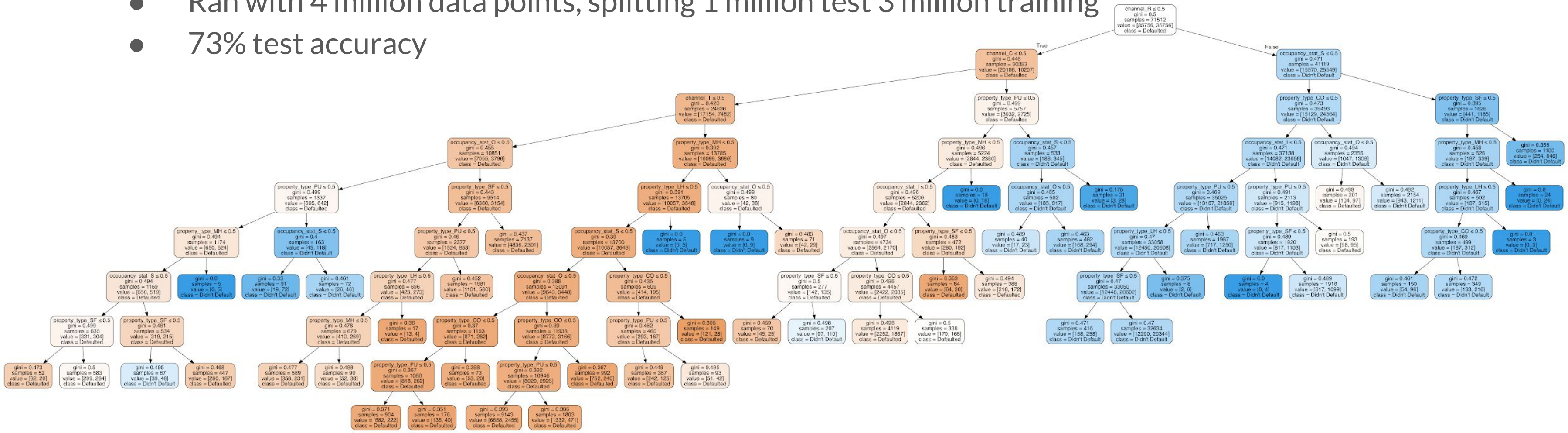
- Read in all of the data from a CSV into a pandas dataframe
- Removed all instances containing null values
- Reduced the dataframe to the important attributes: credit_score, debt_to_income_ratio, original_loan_to_value, interest_rate, default
- Used SciKit Learn LabelEncoder to convert all data types into numeric values
- Normalized the data using the formula: (observation-mean)/std

$$x_i = \frac{x_i - mean(x)}{\sigma(x)}$$

- Normalized the data using the formula: (observation-mean)/std
- Split the data into: feature_train, target_train, feature_test, target_test
- Implemented the machine learning algorithm using SkiKit Learn's GaussianNB Module
- Accuracy score: 98.18216%

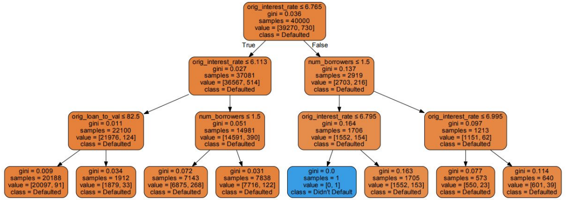
Decision Tree

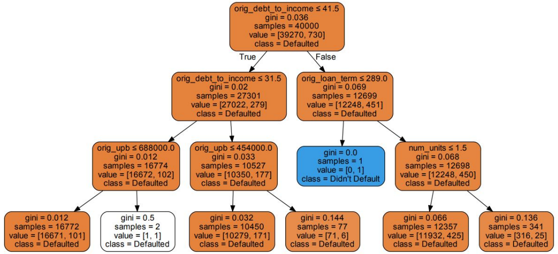
- Read in data from MongoDB, converted into numpy arrays
- Cleaned the data using a random insert strategy
- Ran with 4 million data points, splitting 1 million test 3 million training
- 73% test accuracy

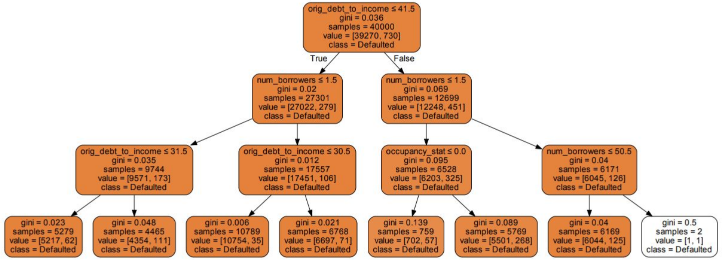


Random Forest

- Generate many random decision trees
  - Each has k random nodes, k << total nodes
  - Use bagging to easily create train and test
- Select the more accurate trees
- Combine them into a large random forest
- Majority vote
- Accuracy: 98.2%
- Recall, or True positive rate: 1.25%
- False Positive Rate: <.001%

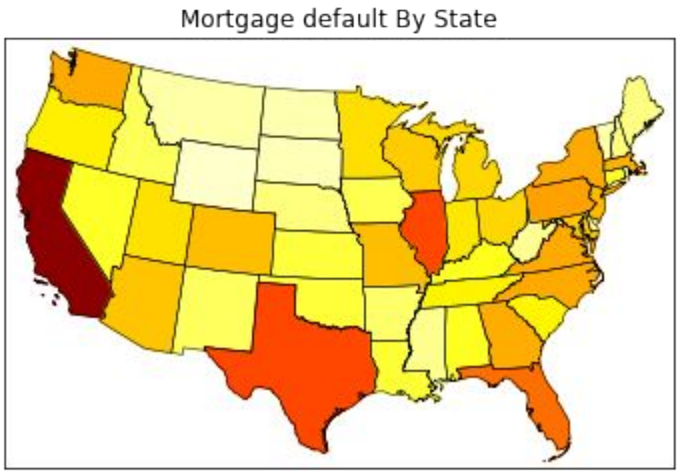






Knowledge gained

- Found methods to predict loan default
- Visualized various states with different defaults
- Made sense of which attributes contributed more to whether or not a default occurred



How that knowledge can be applied.

- These methods can be used by credit rating agencies in the financial industry such as Moody's Investors Services and Standard & Poors's in order to more accurately assess credit risk
- Such methods can also be used by economist to analyze the types of loans that are being given out in order to predict if another housing bubble is to come (characterized by predicting a number of defaults above a particular threshold)