Machine Learning Capstone Project

Definition

Project Overview

Many investment banks trade loan backed securities. If a loan defaults, it leads to devaluation of the securitized product. That is why banks use risk models to identify loans at risk and predict loans that might default in near future. Risk models are also used to decide on approving a loan request by a borrower.

Problem Statement

The goal is to find a model to predict loans that are likely to default and identify attributes that contribute to bad loans, so that it can be used to deny future loan applications.

Since we are trying to predict if a loan will default or not as opposed to predicting the amount of profit or loss that will be incurred, this is a classification problem.

The loan data used here are for consumer loans. So most loans are expected to be short term loans and missing monthly installment has heavy penalty.

Metrics

Looking at the dataset, we can see that out of 5000 loans there are 1533 loans that have defaulted and 3467 loans that were paid. With such an imbalanced dataset, accuracy score might not be a good choice. So we need to look at F1 Score as the metrics to measure model performance.

Analysis

Data Exploration

I will be using data from Lending Club (https://www.lendingclub.com/info/download-data.action). Lending Club is the world's largest online marketplace connecting borrowers and investors.

The file LoansImputed.csv in project folder contains complete loan data for all loans issued through the time period stated.

Variables in Data Set

Dependent Variable

- **not.fully.paid**: A binary variable. 1 means borrower defaulted and 0 means monthly payments are made on time

Independent Variables

- **credit.policy**: 1 if borrower meets credit underwriting criteria and 0 otherwise
- **purpose**: The reason for the loan
- **int.rate**: Annual interest rate for the loan (14% is stored as 0.14)
- **installment**: Monthly payment to be made for the loan
- **log.annual.inc**: Natural log of self-reported annual income of the borrower
- **dti**: Debt to Income ratio of the borrower
- **fico**: FICO credit score of the borrower
- **days.with.cr.line**: Number of days borrower has had credit line
- **revol.bal**: The borrower's revolving balance (Principal loan amount still remaining)
- **revol.util**: Amount of credit line utilized by borrower as percentage of total available credit
- **inq.last.6mths**: Borrowers credit inquiry in last 6 months
- **delinq.2yrs**: Number of times borrower was delinquent in last 2 years
- **pub.rec**: Number of derogatory public record borrower has (Bankruptcy, tax liens and judgements etc.)

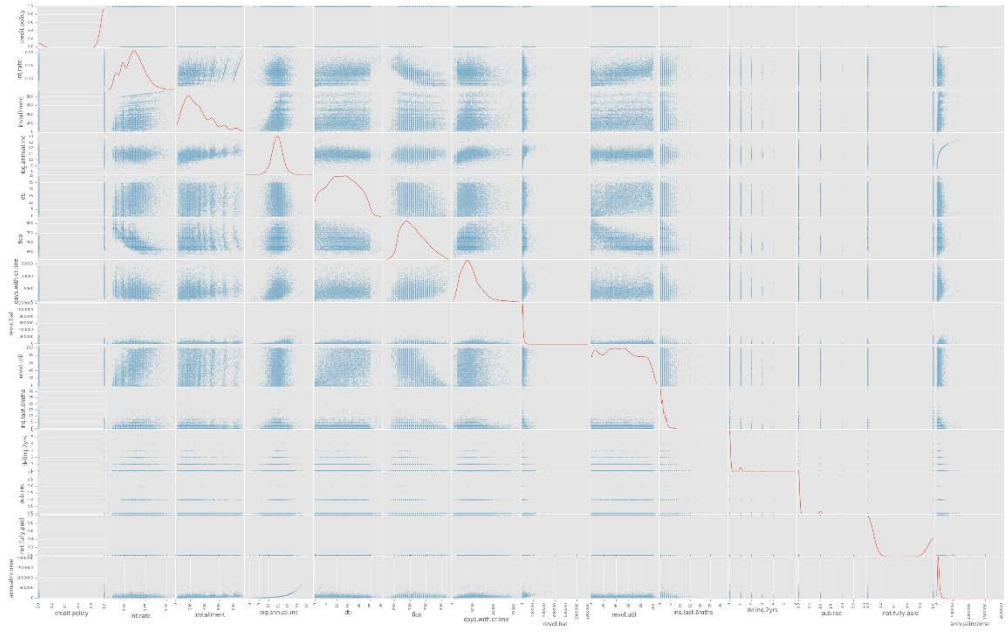
The following observations can be made from looking at statistical summary of the data

- 100% of the accounts that are not default, had met credit underwriting criteria and out of the accounts that are defaulted 66% had met credit underwriting criteria. This means not meeting credit underwriting criteria should be used as primary reason to deny loan application.
- Mean interest rate of loans that defaulted is 13% and for loans that are current is 11%. That means loans that default tend to have higher interest rates.
- Mean installment of loans that defaulted is 342 and for current loans it is 293. This means loans that default tend to have higher installment payments.
- Mean debt to income ratio of defaulted loans is 13.2 and for current loans it is 11.9. This means loans that default tend to have higher debt to income ratio

- Mean revolving balance of default loans is 21066.3 and for current loans it is 13576. This means loans that default tend to have higher revolving balance.
- Mean credit utilization of default loans is 52.25 and for current loans it is 43.8. This means loans that default tend to have higher credit utilization
- Mean credit inquiry of default loans is 2.33 and for current loans it is 0.998. This means loans that default tend to have higher credit inquiry
- Income level, fico score, purpose, days with credit, public record and delinquent in past 2 years don't seem to have a significant correlation with loans being default.

Exploratory Visualization

The plot below shows correlation matrix of all the features in the dataset



From the plot it can be seen that there is strong correlation between interest rate, amount of credit line utilized and debt to income ratio and there is inverse correlation between fico score and interest rate.

Algorithms and Techniques

This is a classification problem as we are trying to predict whether a loan will default or not. The dependent variable not.fully.paid is binary variable that can only be 0 or 1.

The algorithm needs to be fast as loans applied online need approval decision immediately. There is not a lot of data or time needed to train the model. So neural network would not be a good fit.

I will be using Logistic Regression, Support Vector Machine and Extra Trees Classifier to build the model and make prediction and pick the best model based on F1 score.

Logistic Regression is easy to interpret and can work with a lot of features. Since we have 14 features, logistic regression is a good choice.

Support Vector Machine with right kernel function can help avoid overfitting and it also works well with large number of features.

Extra Trees Classifier is an ensemble algorithm that is comparably fast, works with large number of features and can give better accuracy than Logistic Regression or Support Vector Machine.

Benchmark

Credit utilization seems to be one of the primary feature that determines if a loan is likely to default. We could assume that everyone who has 70% credit utilization is going to default. Looking at the dataset 51.14% of the loans that defaulted had credit utilization of 70% or more. We can use this as benchmark to evaluate model performance.

Methodology

Data Preprocessing

There were no missing values in the dataset. Looking at the data it seems purpose is a set of categorical string values consisting of 'debt_consolidation', 'all_other' , 'credit_card', 'small_business', 'home_improvement' , 'educational', 'major_purchase'. This is converted to numerical factor values from 0 to 6. OneHotEncoding is then applied on purpose. Feature scaling is then performed to normalize the data.

Implementation

First the dependent variable is separated from independent variable to create labels and features as separate data frames. The data set is split into training and testing set. Logistic Regression, Support Vector Machine and Extra Trees Classifier algorithms are tried to see which one gives the best result. The table below shows the results

Method	Training accuracy	Test accuracy	F1 Score
LogisticRegression	0.8	0.787	0.53
SVM	0.81	0.79	0.517

ExtraTreesClassifier	1	0.78	0.67

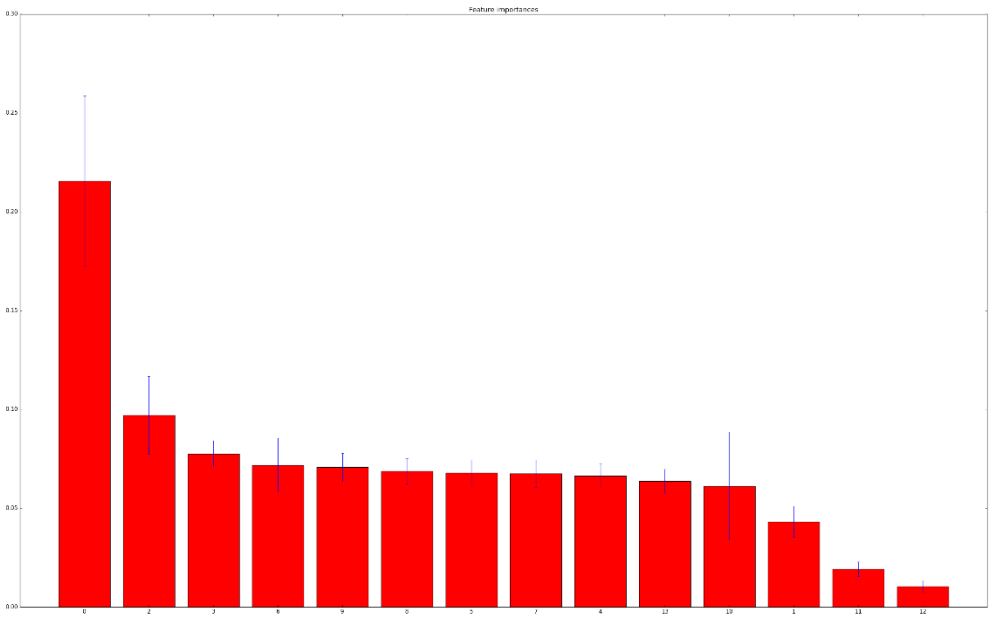
Refinement

ExtraTreesClassifier is used as the final model as it gives the best F1 SCore. The models run very fast and return results within seconds. No further performance improvement is needed. The models work well with default parameter settings.

Results

Model Evaluation and Validation

Using ExtraTreesClassifier in sklearn, the features are ranked by importance as shown in figure below



It seems the most important features in the order of importance are:

1. credit.policy
2. int.rate
3. installment
4. fico

5. revol.util
6. revol.bal
7. dti
8. days.with.cr.line
9. log.annual.inc
10. annualincome
11. inq.last.6mths
12. purpose
13. delinq.2yrs
14. pub.rec

This is in line with what we found from data exploration.

Justification

The final model has accuracy score of 78% and F1 score 0.67, which is more than the rough estimate of 51.14%

Conclusion

This model can be used to predict loans that will be at risk. This can be used to decide whether to approve a loan or not and proactive action can be taken on existing loans that are likely to default.