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 predicts 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.

Metrics

I will be using accuracy score to measure model performance.

Analysis

Data Exploration

I will be using data from Lenging 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.

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**: 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

Using ExtraTreesClassifier in sklearn, the features are ranked by importance as shown in figure below
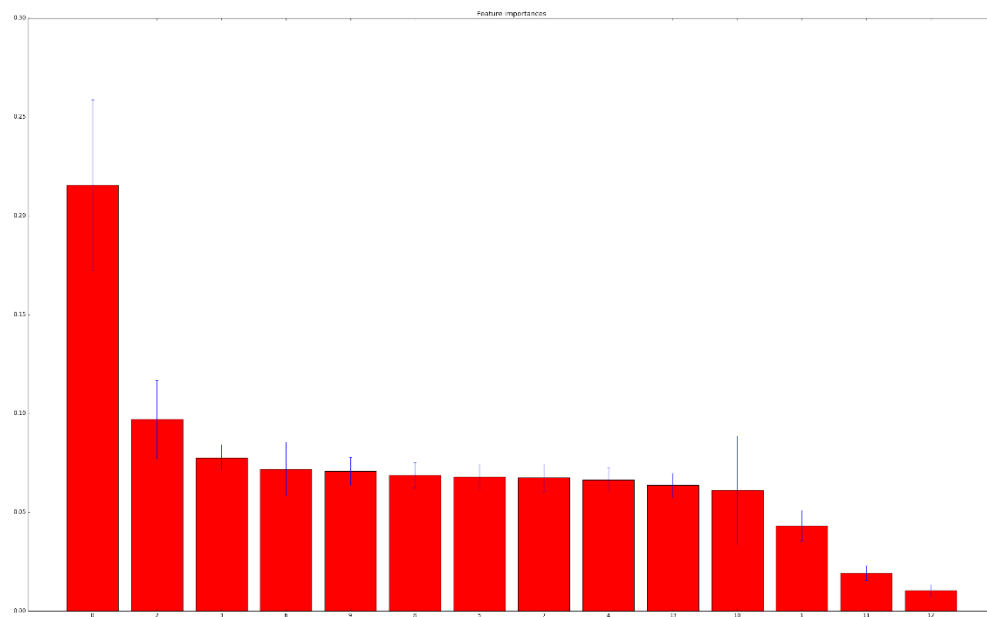
It seems the most important features in the order of importance are:

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.

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. 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 accuracy score.

Benchmark

I will consider 70% accuracy as acceptable criteria to predict if a loan will default.

Methodology

Data Preprocessing

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.

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

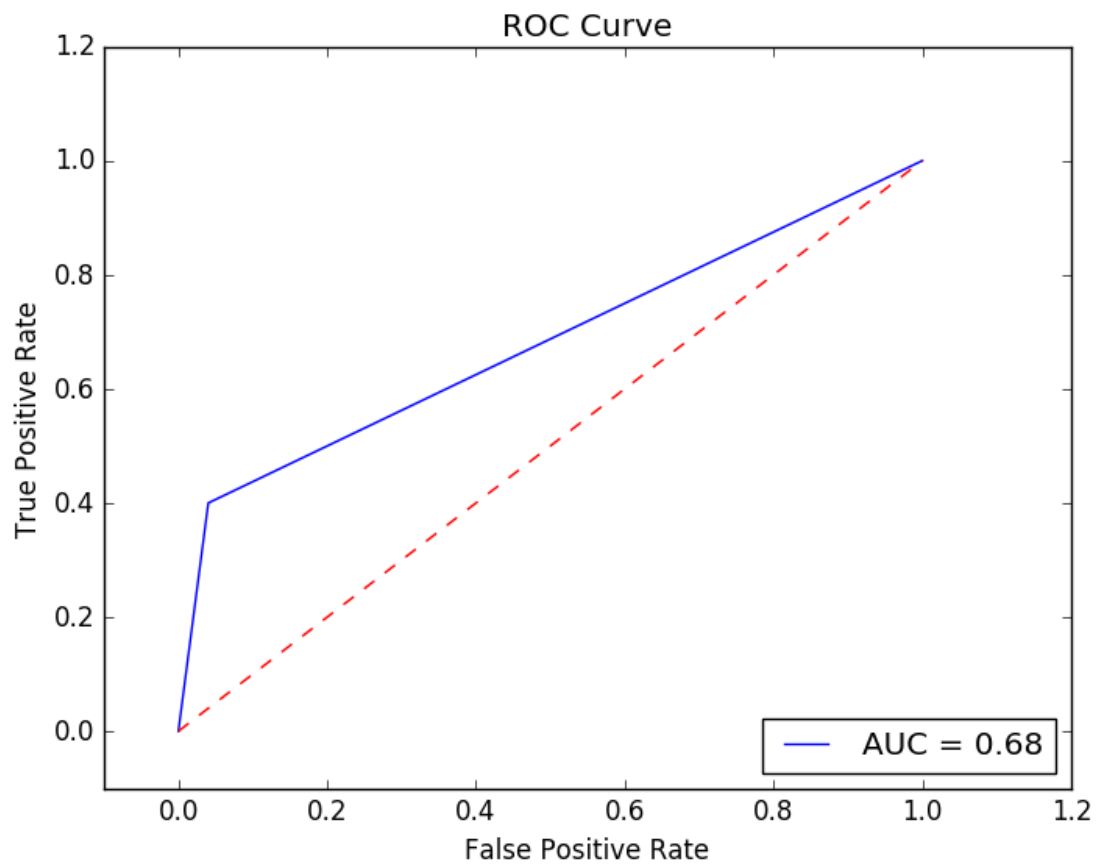| Method	Training accuracy	Test accuracy	Roc auc
LogisticRegression	0.7542	0.746	0.7238
SVM	1	0.68	0.5
ExtraTreesClassifier	1	0.78	0.737

Refinement

ExtraTreesClassifier is used as the final model as it gives the best accuracy.

Results

Model Evaluation and Validation

The picture below shows the ROC curve plotted using the ExtraTreesClassifier

Justification

The final model has accuracy score of 78%, which is more than the required 70% accuracy.

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.