Machine Learning Capstone Project

Introduction

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.

The Data

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.

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**: 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 rovolving 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 deliquent in last 2 years
- pub.rec: Number of derogatory pulic record borrower has (Bankruptcy, tax liens and judgements etc.)

Data Preparation

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.

Train and Predict

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. ExtraTreesClassifier is used to train the model and make prediction.

Conclusion

Our model has an accuracy of 70%. 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.