Loan Default Rate Prediction for Investor

Business Model and Business Goal

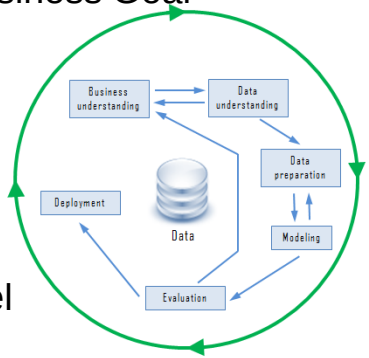
Data Preparation

Data Understanding

Feature Engineering

Machine Learning Model

Model Evaluation and Recommendation



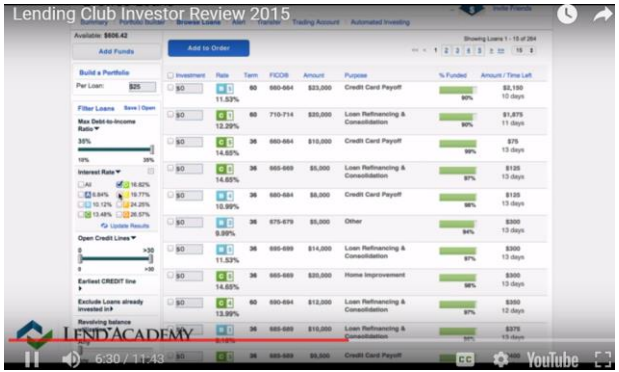
week1
1.Data Exploration
2.Come Up Potencial Problem and Possible work directions

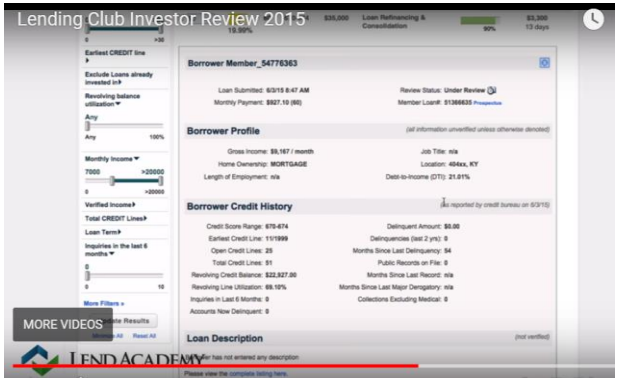
week2
1.Understand business model
2.Data Preparation and Cleaning
3.NLP

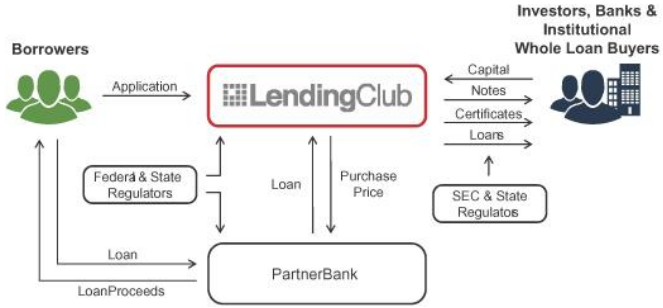
week3
1.Feature Engineering
2.Modeling

week4
1.Machine Learning Models
2.Wrap up

Business Model and Business Goal







Lending Club is a peer to peer loan platform. In this project, we will predict will a loan default or not based on lending club' historical data so that we can provide insight to lending club investors how to choose a profitable loan

Data Preparation

Combine the dataset from 2007 to 2017

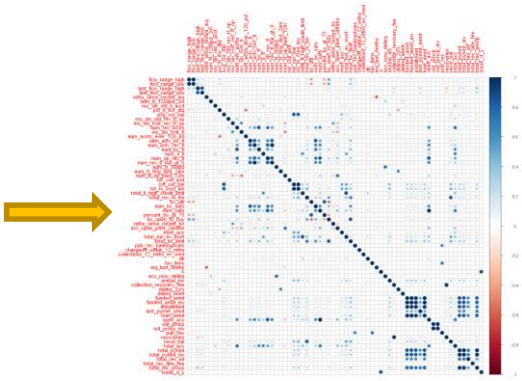
```
loan = pd.concat([loan0711, loan1213, loan14, loan15, loan16, loan17])
```

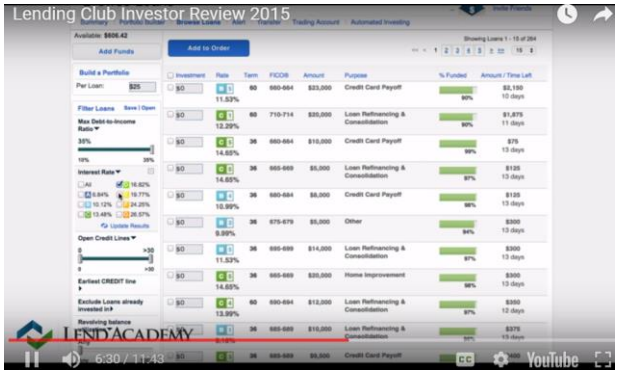
```
loan.shape
```

```
(1543253, 151)
```

```
loan.loan_status.value_counts()
```

```
Fully Paid                                         671952
Current                                            659808
Charged Off                                        176257
Late (31-120 days)                                  18839
In Grace Period                                      9878
Late (16-30 days)                                    3697
Does not meet the credit policy. Status:Fully Paid   1988
Does not meet the credit policy. Status:Charged Off   761
Default                                                52
Name: loan_status, dtype: int64
```

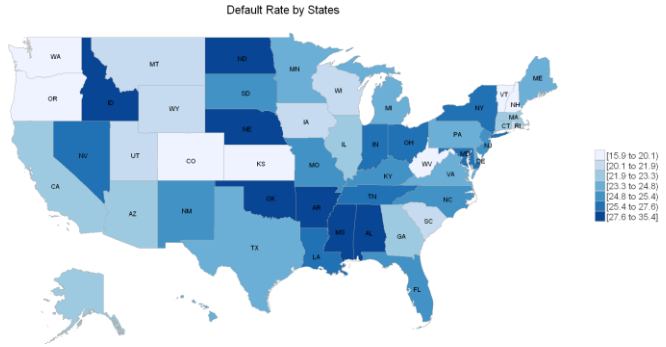


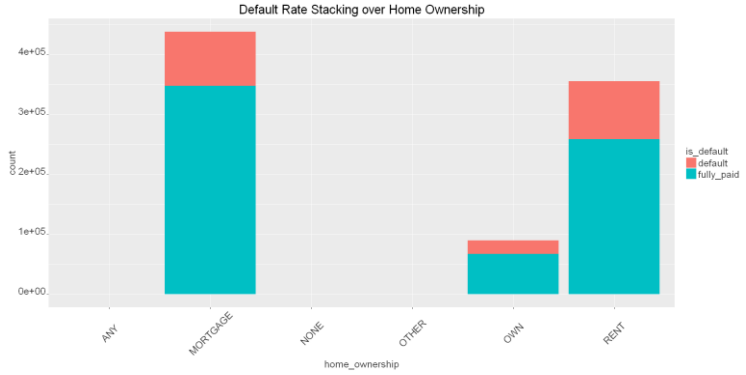


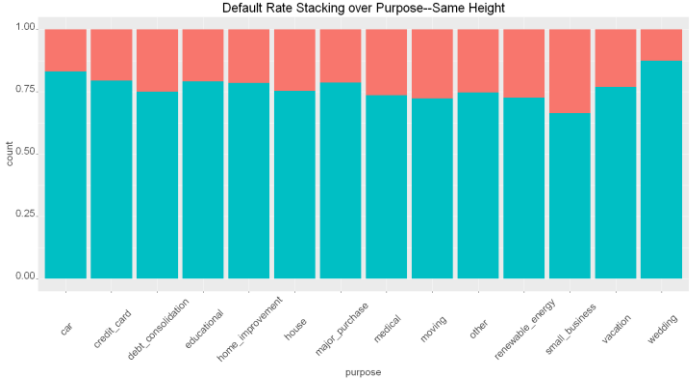
Is_fullypaid
1:Fully Paid
0:Charged Off

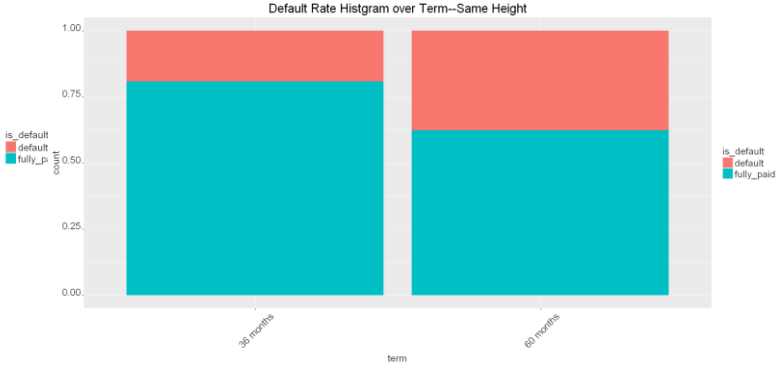
Business Goal → Predictors
Investor known features and some replacement features

Data Understanding









Feature Engineering

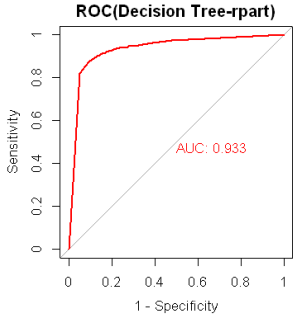
1. Missing Value: Delete columns with too mang missing value and impute other missing value with median
2. Collapsing: addr_state, emp_length, Purpose, scale numeric data into range 0 to 1
3. 21features

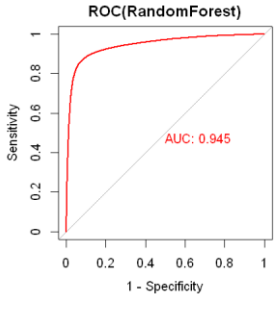
Target: 'fully_paid'
Predictor:

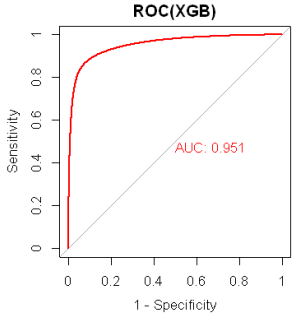
- Loan Payment: 'int_rate','grade', 'term', 'purpose_1', 'installment',
- Borrower Features: 'home_ownership_1','emp_length_1', 'dti', 'last_fico_range_low', 'addr_state_1' ,'annual_inc'
- Borrowers' credit history: 'revol_bal','revol_util','pct_tl_nvr_dlq','delinq_2yrs','delinq_amnt','acc_now_delinq','chargeoff_within_12_mths','collections_12_mths_ex_med','mths_since_recent_inq','mo_sin_old_rev_tl_op','inq_last_6mths'

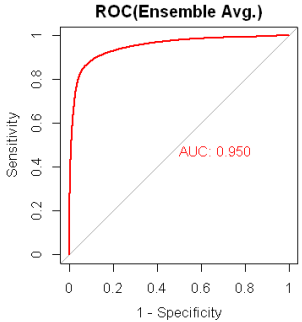
Machine Learning Modeling

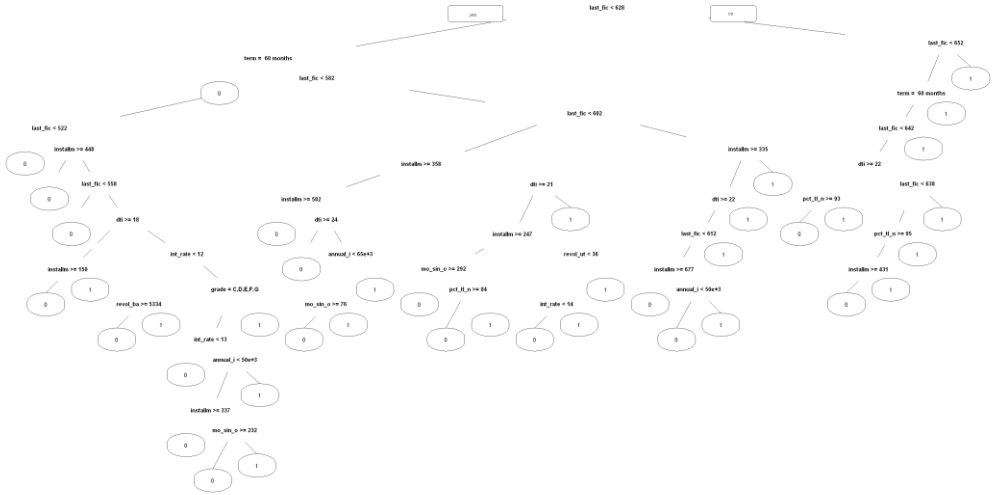
1. Glmnet logistic regression with regularization
2. Decision Tree
3. Random Forest-Tune parameter using Grid Search(Downsampling)
4. XGBoost











Wrap Up

1. Provide investment insights to investor and other stackholders
2. Need more payment features to uncover hidden information. Tried to put payment features, but failed
3. Test the model on small amount data and then implement the model to large amount data
4. Balance the time between data exploration and data modeling.
5. Cannot put text variable such as description into the model due to too much missing value.
6. Prefer decision tree: interpretability