DS4400 Project: Loan Default Prediction

Matthew Martin

Background

- Over $44 billion of loans issued as of December 2018 since 2011
- Lenders face risk of having loan default
- Charged off loans
- Mitigate risk of loss of money if can determine if borrower will not make payments

Dataset

- From Lending Club
- Loan data from 2007-2018
- 145 features
  - Categorical
  - Numerical
- ~ 1303607 data points
  - 1041952 paid off loans
  - 261655 defaulted loans
- Training/Test Split
  - 80/20
  - Before October 2016 in training

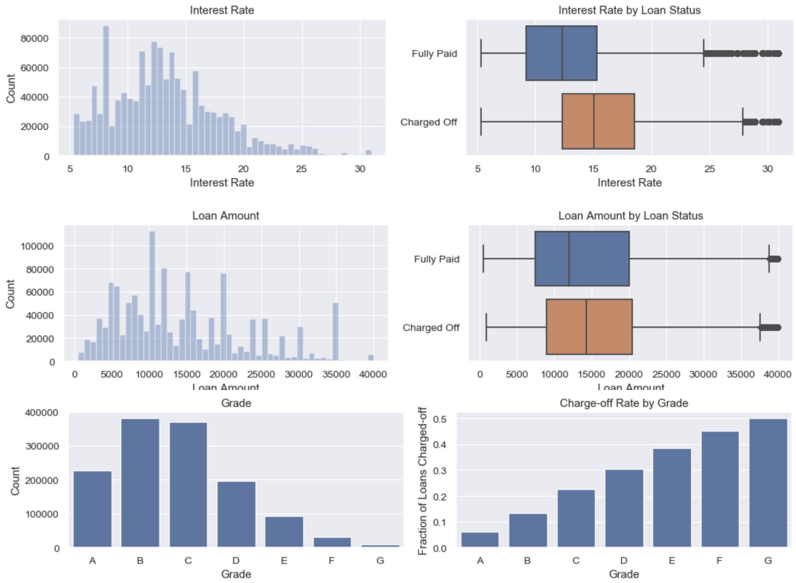
Data Preprocessing

- Deleted features with more than 30% of data missing
- Grouped variables together
  - i.e. State address -> Region
- Dropped features that only apply to one class
  - Based upon feature descriptions
- Converted dates to datetime objects in order to split test and train data

loan_amnt	funded_amnt	int_rate	installment	annual_inc	issue_d	loan_status	dti	delinq_2yrs	earliest_cr_line	...
30000	30000	22.35	1151.16	100000.0	2018-12-01	0	30.46	0.0	2012	...
40000	40000	16.14	975.71	45000.0	2018-12-01	0	50.53	0.0	2009	...
20000	20000	7.56	622.68	100000.0	2018-12-01	0	18.92	0.0	1999	...
4500	4500	11.31	147.99	38500.0	2018-12-01	0	4.64	0.0	2003	...
8425	8425	27.27	345.18	450000.0	2018-12-01	0	12.37	0.0	1997	...
20000	20000	17.97	507.55	57000.0	2018-12-01	0	22.18	0.0	1995	...

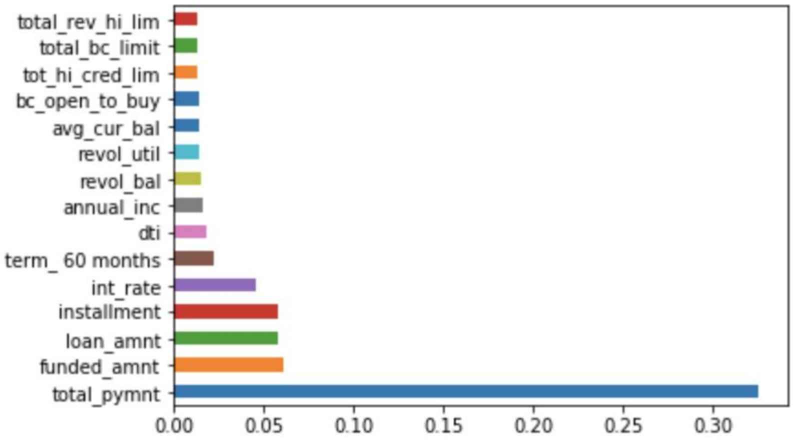
Feature Exploration

- Example features
  - Loan amount
  - Annual income
  - Debt to income ratio
  - Interest rate
- Trends
  - Loans defaulted at higher interest rates
  - Loans defaulted with borrowers of lower income than paid loans
  - Lower grade loans more likely to default



Feature Exploration

- Feature Importance
- Some multicollinearity
- Loan Status correlations
  - Strongest positive
    - Interest rate
    - 60 month term
  - Strongest negative
    - Total payment
    - Grade B loans
    - Open to buy on bank cards



acc_open_past_24mths	0.100731
grade_D	0.108647
grade_E	0.127284
term_ 60 months	0.175899
int_rate	0.258412
total_pymnt	-0.318482
total_pymnt_inv	-0.318038
grade_B	-0.106295
bc_open_to_buy	-0.077103
avg_cur_bal	-0.071585
tot_hi_cred_lim	-0.070042

Model Results

Model	Train Accuracy	Test Accuracy
Logistic Regression (0.25)	92.9%	98.5%
Logistic Regression (0.50)	94.4%	98.2%
Logistic Regression (0.75)	92.8%	96.7%
Logistic Regression (0.90)	90.2%	94.7%
LDA	89.6%	94.8%
SVM	93.9%	97.6%
Random Forest (10)	99.5%	91.8%
Random Forest (50)	99.9%	97.8%
Random Forest (100)	1.0%	98.3%
Random Forest (250)	1.0%	98.6%
Adaboost (10)	94.5%	98.8%
Adaboost (50)	95.4%	99.2%
Adaboost (100)	95.8%	99%
Neural Network	95.4%	99%

Model	Area Under Curve
Logistic Regression (0.5)	98.3%
LDA	95.1%
Linear SVM	97.6%
Random Forest (250)	98.6%
Adaboost (50)	99.2%
Neural Network	99%

Challenges

- Dealing with large amount of features
- Gaining domain knowledge
  - Needed to get rid of features that solely predicted one class
- Data preprocessing
- Label imbalance
  - Needed to find more data
  - Undersample

Conclusion

- Hard to classify if borrower will default on loan solely upon features before loan is approved
  - Spread between feature values associated between class labels is small
- Feature associations
  - Lead to paid
    - Higher total payment
    - Higher credit limit
    - Higher annual income
  - Lead to default
    - Higher funded amount
    - Higher debt to income ratio
    - Higher interest rate
    - Higher past due delinquencies
- May need to look into if important features are impacting prediction too much

Future Work

- Do deeper analysis on borrowers who defaulted
  - See what trend of pre acceptance variables may lead to profit even if loan defaults
  - Try to see if there is an ideal recommendation for loan features per borrower applicant based on his or her credentials
- Regression problem
  - Try to predict the interest rate for a specific loan and borrower