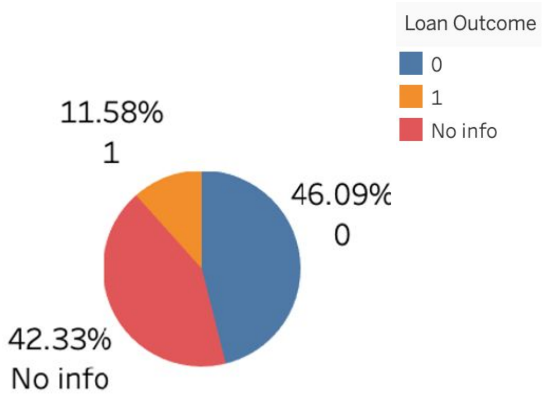
Good Loan vs. Bad Loan: Can we predict loan outcome?

Executive summary

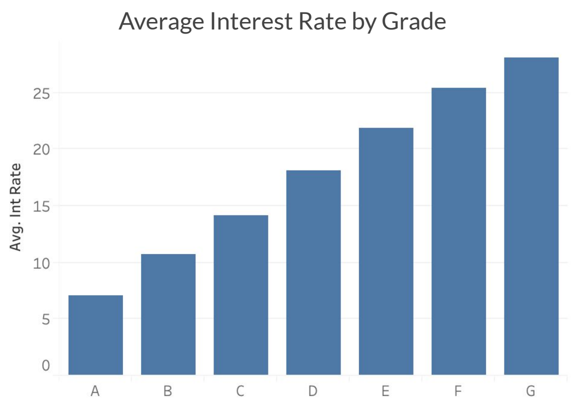
1. Problem:
    a. How could different variables influence the outcome of loans?
    b. Can we build a model to predict outcome?
2. Exploratory Data Analysis
    a. Focusing on good loans & bad loans (paid off or not)
3. Split train & test dataset
    a. 70% train, 30% test
4. Feature Selection
    a. loan_amnt, int_rate, grade, purpose, emp_length, home_ownership, annual_inc, term, region
5. Model Fitting
6. Diagnostic Test

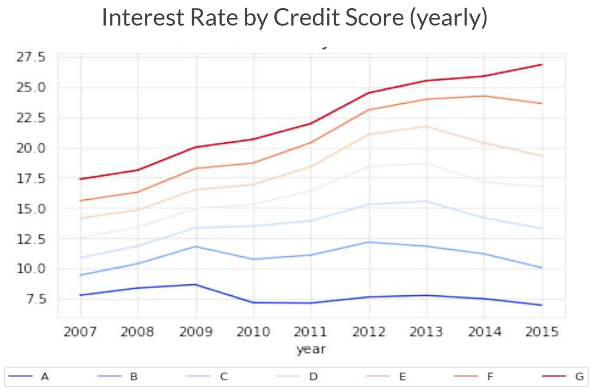


Distribution of loan outcome

Data

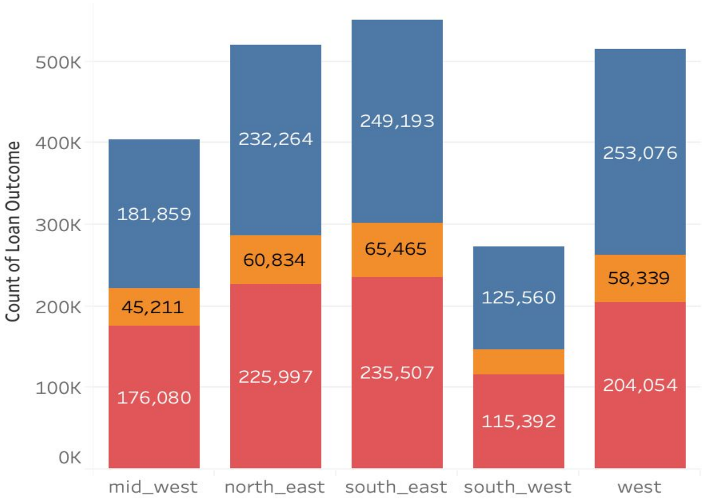
- Kaggle: [Lending Club Loan Data]
- 2.26M rows, 145 variables
- 21 empty columns, only 33 variables do not contain any missing values
- There exists multicollinearity among variables
  - eg. Grade and Interest Rate (lower grade, higher interest rate)
- Convert Loan_Status to Loan_Outcome for future use
  - Fully Paid => 0 (good loan)
  - Charged Off & Default => 1 (bad loan)
  - Others => No Info
- Convert State value to Regional information
  - mid_west, north_east, south_east, south_west & west



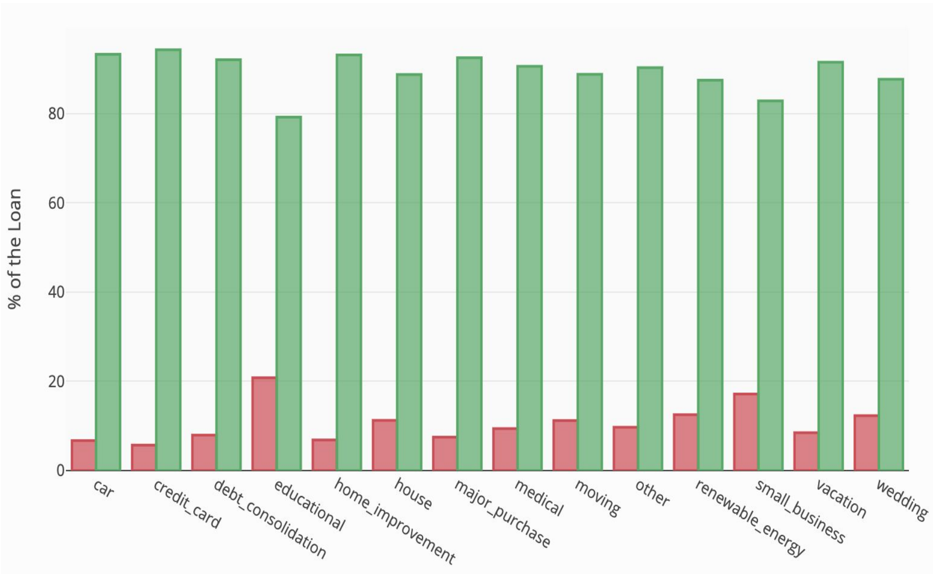


Variables related to Loan Outcome

Loan Outcome by Reigon



Loan Outcome by Purpose



Logistic Regression Model

- Multicollinearity between int_rate and grade (see VIF)
- Model Selection (Under AIC criteria)
- Outlier Test on Pearson Residuals
- Coefficient Interpretation:

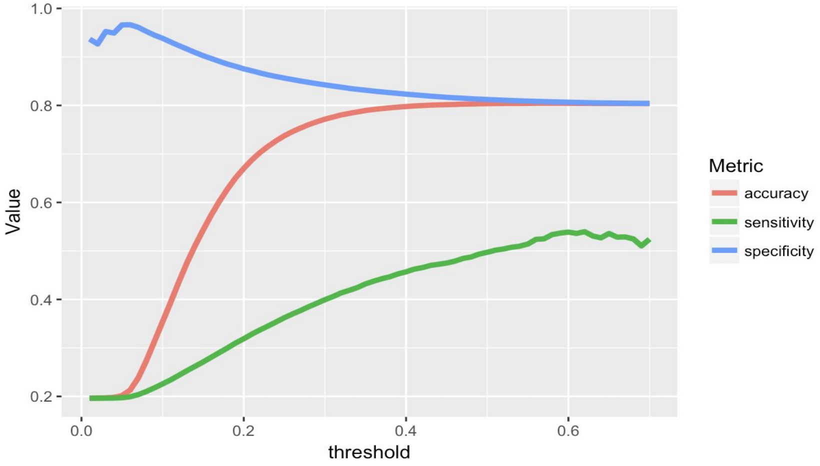
loan_amount 1.0000105; int_rate = 1.1125410; annual_inc: 0.9999974

	GVIF	Df	GVIF^(1/(2*Df))
loan_amnt	1.622483	1	1.273767
int_rate	9.320019	1	3.052871
grade	9.569355	6	1.207092
emp_length	1.058270	10	1.002836
home_ownership	1.198763	3	1.030676
annual_inc	1.352625	1	1.163024
term	1.434132	1	1.197553
purpose	1.168219	13	1.005998
addr_state	1.053736	4	1.006564

	GVIF	Df	GVIF^(1/(2*Df))
loan_amnt	1.618395	1	1.272162
int_rate	1.273189	1	1.128357
emp_length	1.057413	10	1.002795
home_ownership	1.198010	3	1.030568
annual_inc	1.352473	1	1.162959
term	1.425591	1	1.193981
purpose	1.161258	13	1.005767
addr_state	1.054163	4	1.006615

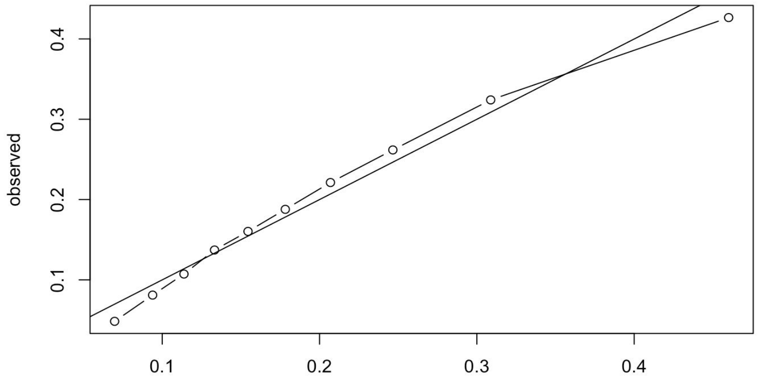
(Intercept)	-3.486e+00	3.323e-02	-104.900	< 2e-16	***
loan_amnt	1.154e-05	3.934e-07	29.321	< 2e-16	***
int_rate	1.060e-01	6.348e-04	167.004	< 2e-16	***
emp_length1 year	-6.434e-03	1.375e-02	-0.468	0.639761	
emp_length10+ years	-6.152e-02	1.050e-02	-5.858	4.67e-09	***
emp_length2 years	-3.873e-02	1.275e-02	-3.037	0.002387	**
emp_length3 years	-1.257e-02	1.313e-02	-0.957	0.338332	
emp_length4 years	-2.369e-02	1.424e-02	-1.664	0.096161	.
emp_length5 years	-3.868e-02	1.410e-02	-2.743	0.006091	**
emp_length6 years	-5.991e-02	1.547e-02	-3.874	0.000107	***
emp_length7 years	-4.619e-02	1.568e-02	-2.947	0.003214	**
emp_length8 years	1.790e-03	1.553e-02	0.115	0.908230	
emp_length9 years	-4.418e-03	1.645e-02	-0.269	0.788285	
home_ownershipOTHER	2.088e-01	2.652e-01	0.787	0.431066	
home_ownershipOWN	1.945e-01	9.560e-03	20.349	< 2e-16	***
home_ownershipRENT	3.796e-01	6.327e-03	59.987	< 2e-16	***
annual_inc	-2.933e-06	7.571e-08	-38.742	< 2e-16	***
term60 months	4.625e-01	6.907e-03	66.960	< 2e-16	***
purposecredit_card	1.833e-01	3.049e-02	6.011	1.85e-09	***
purposedebt_consolidation	2.441e-01	3.006e-02	8.119	4.70e-16	***
purposeeducational	2.216e-01	1.831e-01	1.210	0.226281	
purposehome_improvement	2.720e-01	3.199e-02	8.501	< 2e-16	***
purposehouse	3.143e-02	4.723e-02	0.665	0.505742	
purposemajor_purchase	2.112e-01	3.543e-02	5.962	2.50e-09	***
purposemedical	3.371e-01	3.911e-02	8.620	< 2e-16	***
purposemoving	2.319e-01	4.312e-02	5.378	7.55e-08	***
purposeother	2.015e-01	3.191e-02	6.314	2.71e-10	***
purposerenewable_energy	3.317e-01	1.019e-01	3.256	0.001131	**
purposesmall_business	5.333e-01	3.725e-02	14.318	< 2e-16	***
purposevacation	2.785e-01	4.517e-02	6.167	6.97e-10	***
purposewedding	-5.126e-01	8.353e-02	-6.136	8.44e-10	***
addr_statesouth_east	1.574e-01	8.112e-03	19.403	< 2e-16	***
addr_statesouth_west	1.739e-01	9.823e-03	17.705	< 2e-16	***
addr_statenorth_east	1.455e-01	8.114e-03	17.928	< 2e-16	***
addr_statemid_west	1.263e-01	8.893e-03	14.199	< 2e-16	***

LR Model Cont. & Diagnostics

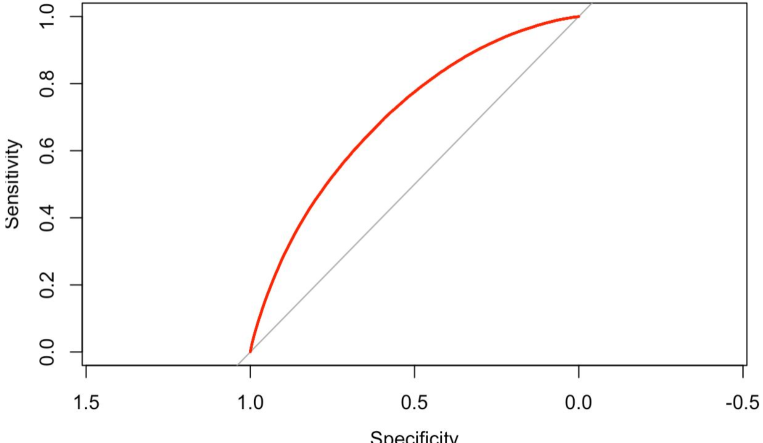



```
	Hosmer and Lemeshow goodness of fit (GOF) test

data:  testset$loan_outcome, preds
X-squared = 469.25, df = 8, p-value < 2.2e-16
```



ROC Graph



Key considerations (Confusion Matrix)

- Threshold is chosen as 0.25:

Predicted	Actual 0	Actual 1
0	168071	23791
1	77178	35826

- Threshold chosen as 0.3:

Predicted	Actual 0	Actual 1
0	198237	33255
1	46792	26582

- Less samples in Category : Loan_outcome = 1

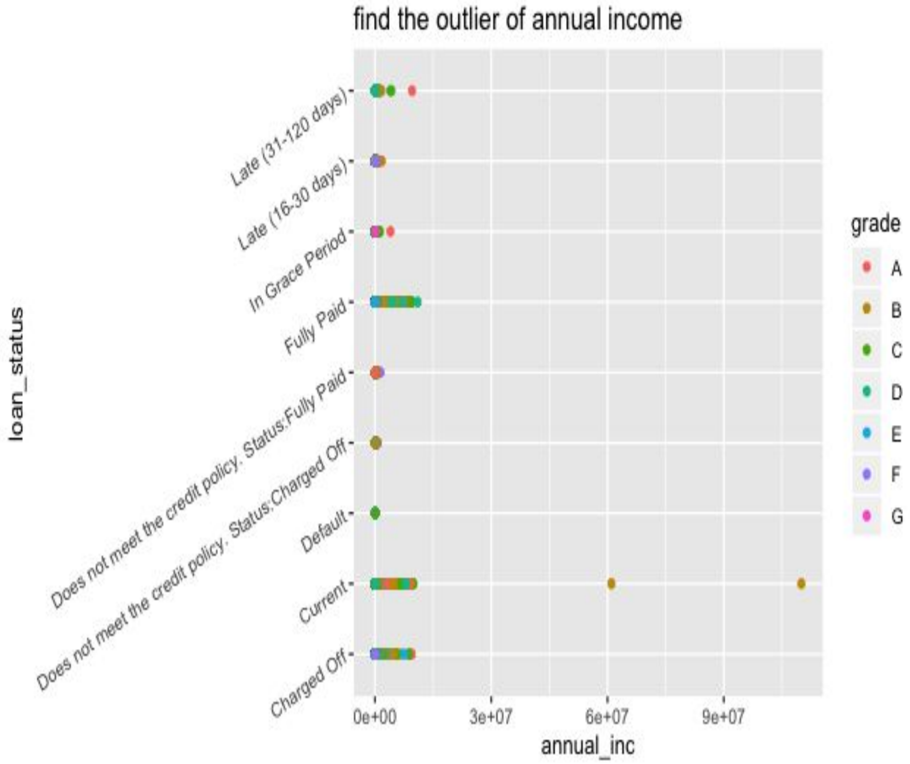
LDA & Diagnostics

- Linear Discriminant Analysis
- Outlier
- Hosmer-Lemeshow goodness of fit test
- Confusion Matrix

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  testset$loan_outcome, pred2$posterior[, 1]
X-squared = 1703434, df = 8, p-value < 2.2e-16

          Actual
Predicted      0      1
        0 304862  70500
        1   7932   7797
```

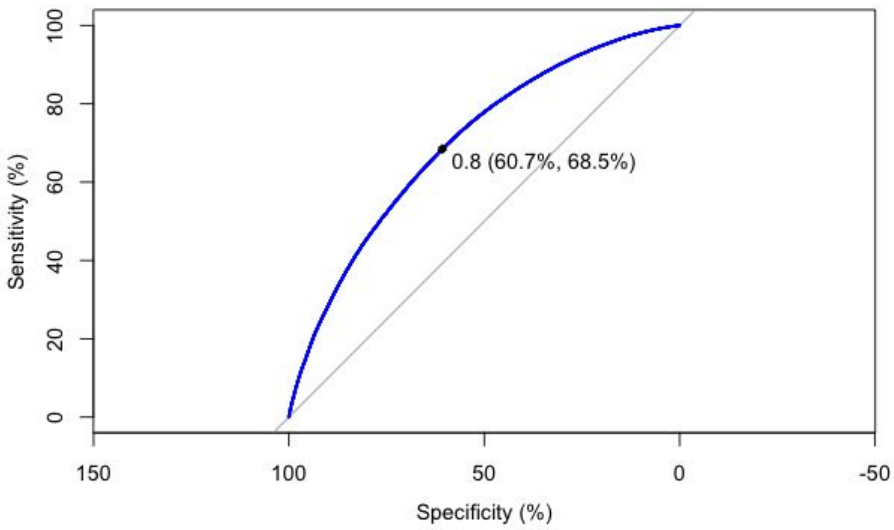


Model2 & Diagnostics

- ROC curves are a nice way to see how any predictive model can distinguish between the true positives and negatives
- the ROC curve is to the upper left corner, the higher the overall accuracy of the test

- Area under the curve: 0.7011
- Higher the AUC, better the model is at distinguishing between 0 and 1

Receiver Operating Characteristic Graph



Conclusion

- Best accuracy: 77.995%, threshold = 32%
- Most influential variable: loan amount, interest rate & annual income
- Problem existing
  - insufficient amount of data for bad loans
- Potential Improvement
  - More data
  - Better classification machine learning model (regularization)
  - Smarter us?