

Machine Learning Project: Credit Risk

FANG HSUAN (ADAM) TSENG

Agenda

1. Introduction
2. Data Source and Review
3. Explanatory Analysis
4. Model Result
5. Conclusion

Introduction

The focus of this project is to try find different machine learnings models – other than logistic regression model - that can make good predictions with binary outcome.

The motivation is based on that realizing the wide application of logistic regression in industry and the trend moving towards machine learning, I want to research on different models that can make binary binary predictions and compare them to see which one should be applied under different circumstances.

Data Source and Review

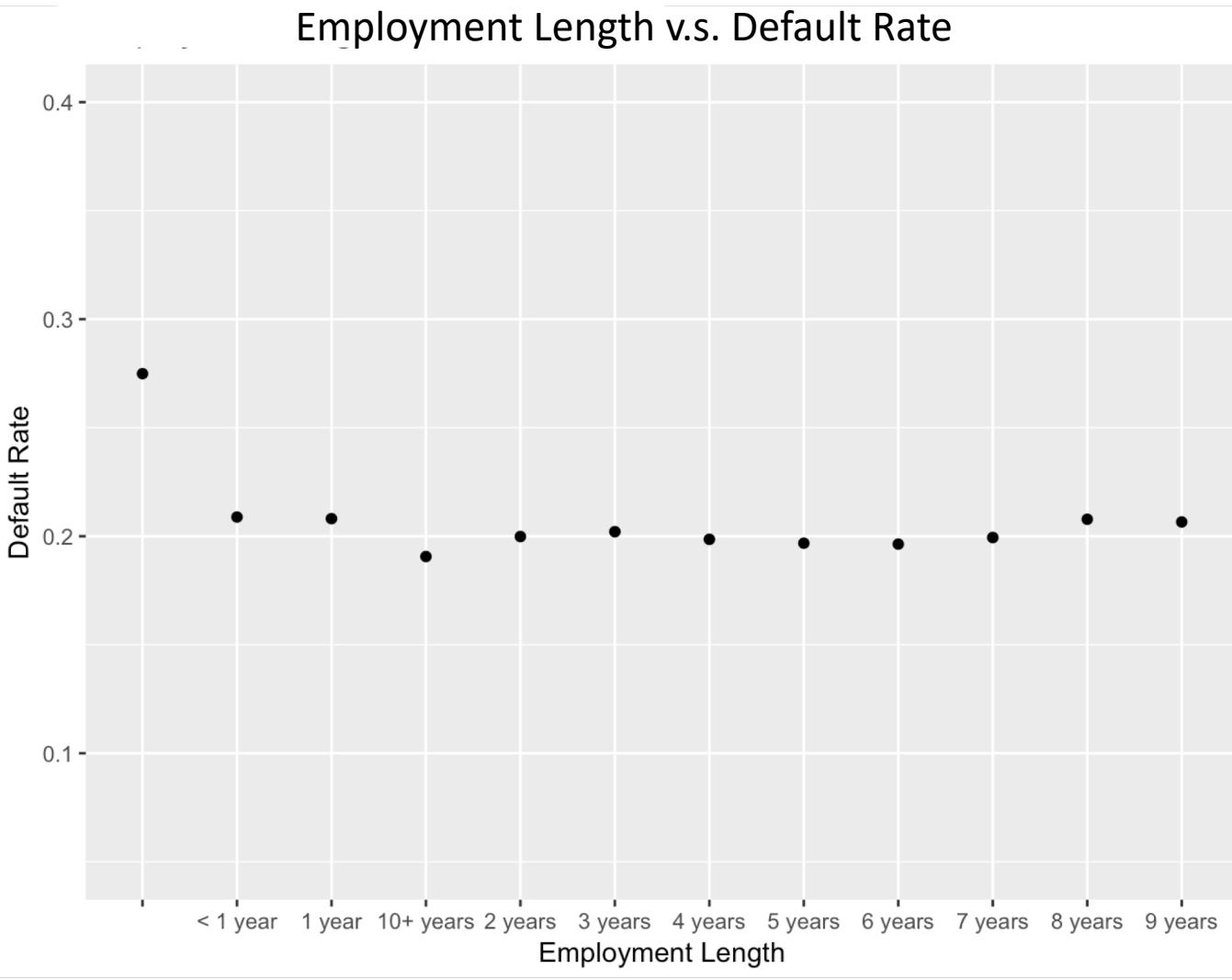
1. Introduction
2. Data Source and Review
3. Explanatory Analysis
4. Model Result
5. Conclusion

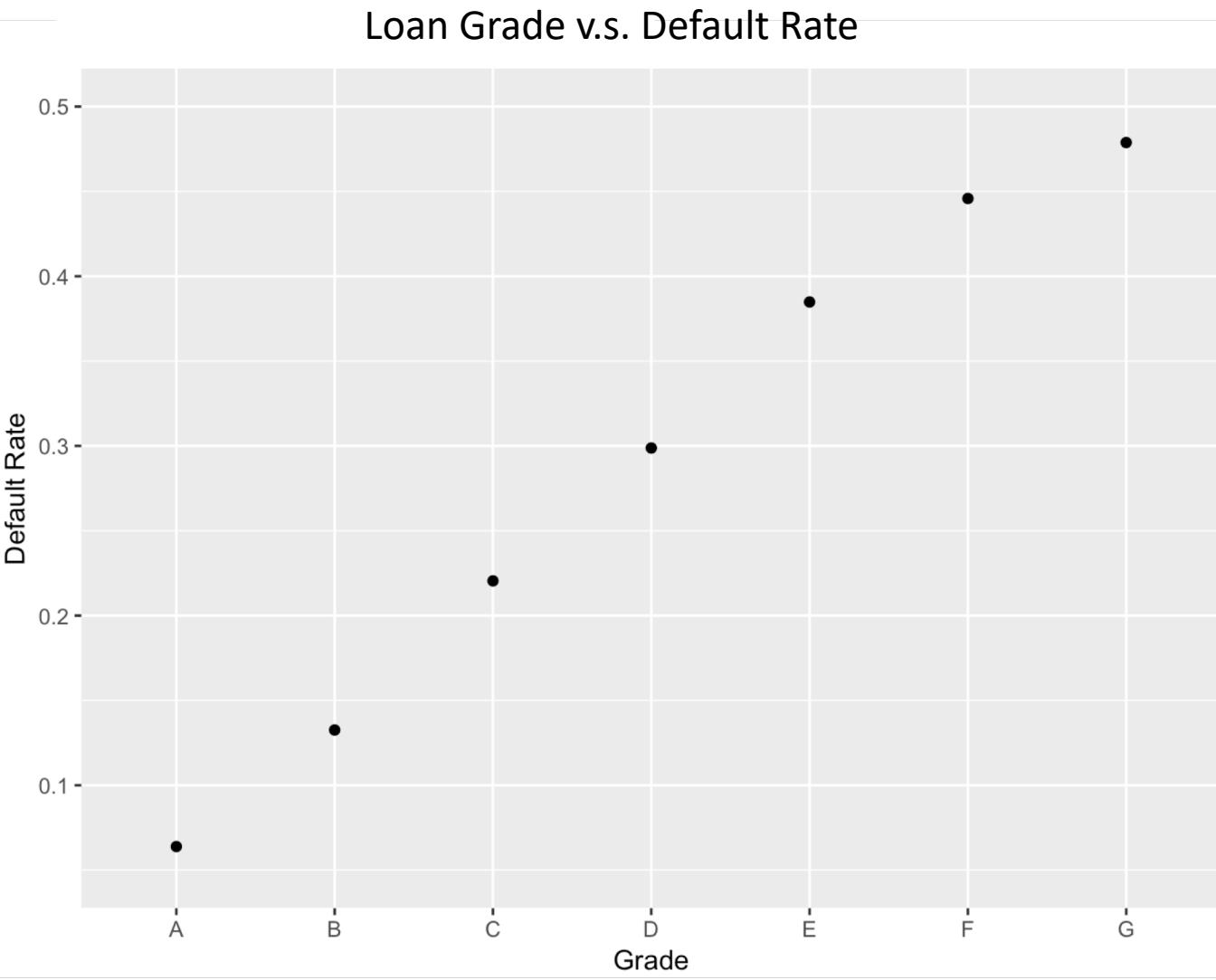
Data Source and Review

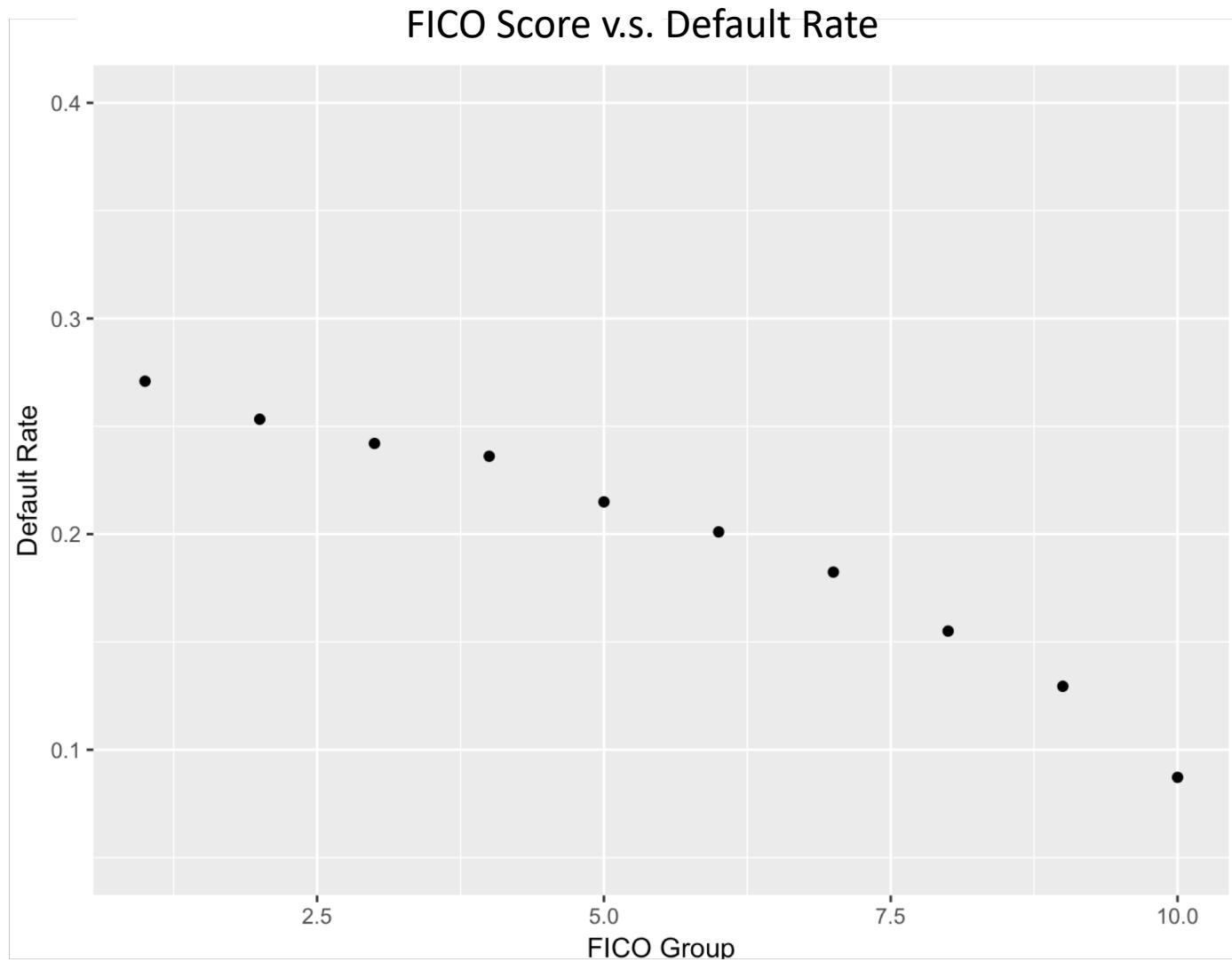
- Data Source: Lending Club Official Website
- Dates of Data: 2007 – 2018Q2
- Total number of observations: 843,934
- Independent Variables: 151
- Default Percentage: 0.2024116

Explanatory Analysis

1. Introduction
2. Data Source and Review
3. Explanatory Analysis
4. Model Results
5. Conclusion



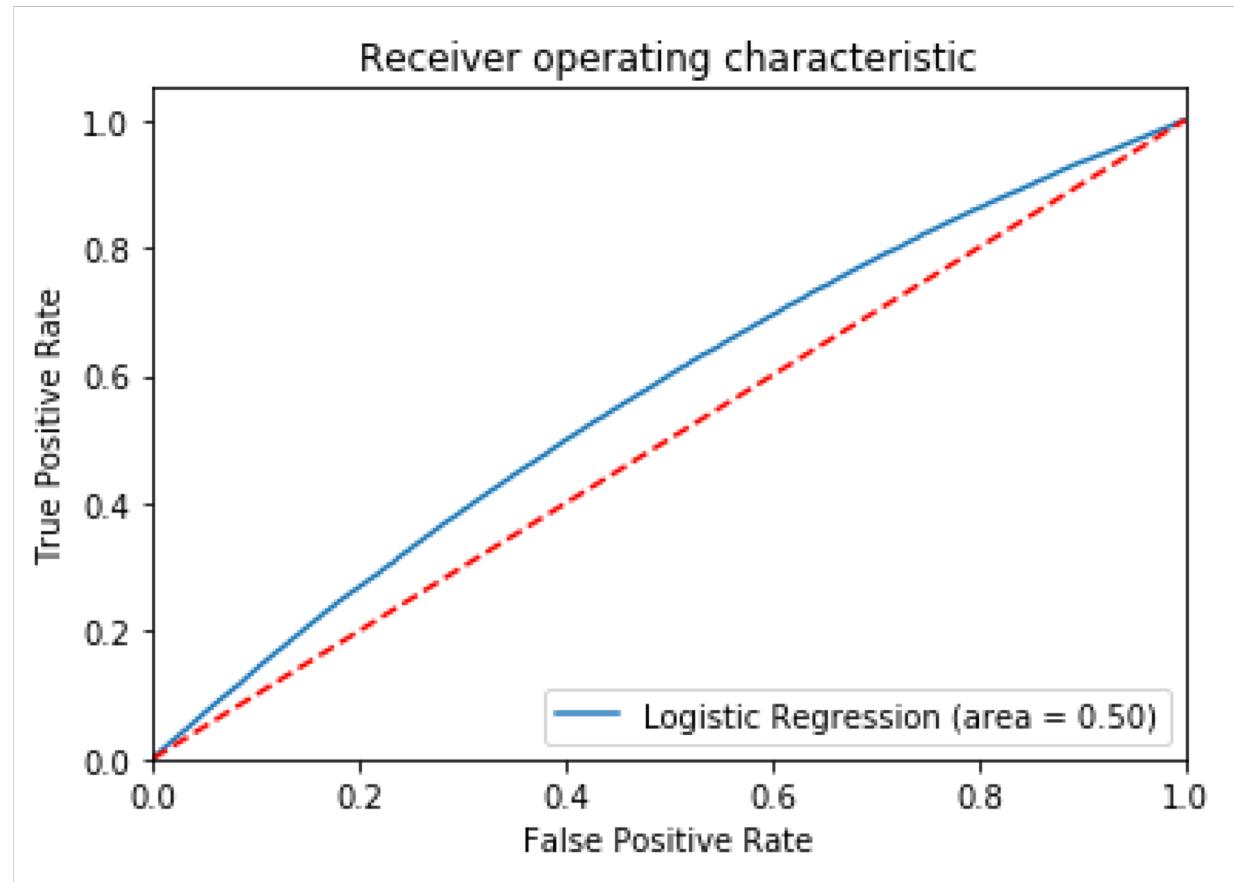




Model Result

1. Introduction
2. Data Source and Review
3. Explanatory Analysis
4. Model Result
5. Conclusion

Logistic Regression – Simple Model



Logistic Regression Full Model

| Logit Regression Results | | | | | | |
|--------------------------|------------------|-------------------|-------------|-------|-----------|-----------|
| Dep. Variable: | loan_status | No. Observations: | 590753 | | | |
| Model: | Logit | Df Residuals: | 590693 | | | |
| Method: | MLE | Df Model: | 59 | | | |
| Date: | Tue, 13 Nov 2018 | Pseudo R-squ.: | 0.9841 | | | |
| Time: | 23:48:18 | Log-Likelihood: | -4739.0 | | | |
| converged: | True | LL-Null: | -2.9737e+05 | | | |
| | | LLR p-value: | 0.000 | | | |
| | coef | std err | z | P> z | [0.025 | 0.975] |
| loan_amnt | 0.0032 | 4.53e-05 | 70.050 | 0.000 | 0.003 | 0.003 |
| int_rate | -0.0632 | 0.010 | -6.020 | 0.000 | -0.084 | -0.043 |
| annual_inc | -2.187e-06 | 4.21e-07 | -5.197 | 0.000 | -3.01e-06 | -1.36e-06 |
| dti | 0.0127 | 0.001 | 9.364 | 0.000 | 0.010 | 0.015 |
| delinq_2yrs | 0.1573 | 0.049 | 3.240 | 0.001 | 0.062 | 0.252 |
| inq_last_6mths | 0.0475 | 0.038 | 1.267 | 0.205 | -0.026 | 0.121 |
| open_acc | -0.6462 | 0.057 | -11.372 | 0.000 | -0.758 | -0.535 |
| pub_rec | 0.0551 | 0.139 | 0.396 | 0.692 | -0.218 | 0.328 |
| revol_bal | -2.494e-05 | 4.22e-06 | -5.905 | 0.000 | -3.32e-05 | -1.67e-05 |
| revol_util | -0.0071 | 0.002 | -3.251 | 0.001 | -0.011 | -0.003 |
| total_acc | -0.3563 | 0.022 | -16.355 | 0.000 | -0.399 | -0.314 |
| total_pymnt | -0.0035 | 4.75e-05 | -73.504 | 0.000 | -0.004 | -0.003 |
| total_rec_int | 0.0039 | 7.33e-05 | 52.836 | 0.000 | 0.004 | 0.004 |
| total_rec_late_fee | 0.0389 | 0.002 | 16.527 | 0.000 | 0.034 | 0.044 |
| recoveries | 120.3509 | 5.377 | 22.384 | 0.000 | 109.813 | 130.889 |
| collection_recovery_fee | 1.9541 | 10.707 | 0.183 | 0.855 | -19.031 | 22.939 |
| last_pymnt_amnt | -0.0007 | 2.7e-05 | -26.027 | 0.000 | -0.001 | -0.001 |
| last_fico_range_high | -0.0142 | 0.001 | -27.813 | 0.000 | -0.015 | -0.013 |

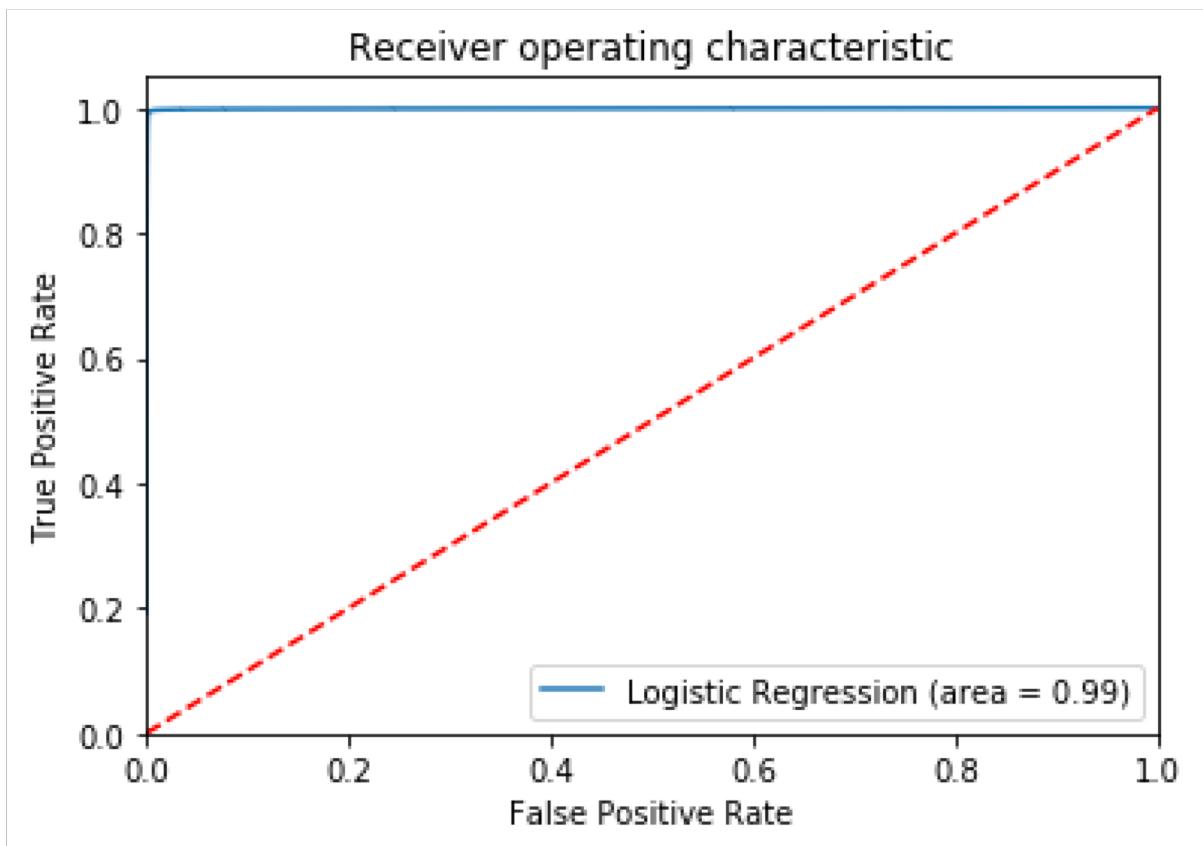
Logistic Regression Full Model

| | | | | | | |
|----------------------------|------------|----------|--------|-------|-----------|-----------|
| last_fico_range_low | -0.0012 | 0.000 | -4.396 | 0.000 | -0.002 | -0.001 |
| collections_12_mths_ex_med | 0.1187 | 0.123 | 0.963 | 0.335 | -0.123 | 0.360 |
| acc_now_delinq | -0.5231 | 1.102 | -0.475 | 0.635 | -2.682 | 1.636 |
| tot_coll_amt | -1.647e-05 | 2.47e-05 | -0.667 | 0.505 | -6.49e-05 | 3.19e-05 |
| tot_cur_bal | -3.562e-06 | 7.09e-07 | -5.025 | 0.000 | -4.95e-06 | -2.17e-06 |
| total_rev_hi_lim | 2.349e-06 | 9.49e-07 | 2.477 | 0.013 | 4.9e-07 | 4.21e-06 |
| acc_open_past_24mths | 0.0239 | 0.017 | 1.437 | 0.151 | -0.009 | 0.057 |
| avg_cur_bal | 1.92e-05 | 4.37e-06 | 4.391 | 0.000 | 1.06e-05 | 2.78e-05 |
| bc_open_to_buy | -2.902e-06 | 6.77e-06 | -0.429 | 0.668 | -1.62e-05 | 1.04e-05 |
| bc_util | 0.0122 | 0.002 | 5.165 | 0.000 | 0.008 | 0.017 |
| chargeoff_within_12_mths | 0.1567 | 0.264 | 0.593 | 0.553 | -0.361 | 0.675 |
| delinq_amnt | 7.706e-06 | 4.81e-05 | 0.160 | 0.873 | -8.66e-05 | 0.000 |
| mo_sin_old_il_acct | 0.0010 | 0.001 | 1.580 | 0.114 | -0.000 | 0.002 |
| mo_sin_old_rev_tl_op | 0.0020 | 0.000 | 5.277 | 0.000 | 0.001 | 0.003 |
| mo_sin_rcnt_rev_tl_op | -7.358e-05 | 0.003 | -0.024 | 0.981 | -0.006 | 0.006 |
| mo_sin_rcnt_tl | 0.0004 | 0.005 | 0.076 | 0.940 | -0.010 | 0.010 |
| mort_acc | 0.3947 | 0.031 | 12.818 | 0.000 | 0.334 | 0.455 |
| mths_since_recent_bc | 0.0026 | 0.001 | 2.113 | 0.035 | 0.000 | 0.005 |
| mths_since_recent_inq | 0.0159 | 0.006 | 2.591 | 0.010 | 0.004 | 0.028 |
| num_accts_ever_120_pd | 0.1775 | 0.027 | 6.597 | 0.000 | 0.125 | 0.230 |
| num_actv_bc_tl | -0.0934 | 0.038 | -2.433 | 0.015 | -0.169 | -0.018 |
| num_actv_rev_tl | 0.0266 | 0.024 | 1.119 | 0.263 | -0.020 | 0.073 |
| num_bc_sats | 0.1252 | 0.027 | 4.580 | 0.000 | 0.072 | 0.179 |
| num_bc_tl | -0.0407 | 0.018 | -2.244 | 0.025 | -0.076 | -0.005 |
| num_il_tl | 0.3399 | 0.023 | 15.050 | 0.000 | 0.296 | 0.384 |
| num_op_rev_tl | -0.0772 | 0.024 | -3.205 | 0.001 | -0.124 | -0.030 |

Logistic Regression Full Model

| | | | | | | |
|----------------------------|------------|----------|--------|-------|-----------|-----------|
| num_rev_accts | 0.3870 | 0.025 | 15.680 | 0.000 | 0.339 | 0.435 |
| num_sats | 0.6898 | 0.057 | 12.104 | 0.000 | 0.578 | 0.802 |
| num_tl_120dpd_2m | 1.2717 | 1.280 | 0.994 | 0.320 | -1.236 | 3.780 |
| num_tl_30dpd | 1.0913 | 1.183 | 0.922 | 0.356 | -1.228 | 3.411 |
| num_tl_90g_dpd_24m | -0.1670 | 0.078 | -2.142 | 0.032 | -0.320 | -0.014 |
| num_tl_op_past_12m | 0.0126 | 0.027 | 0.463 | 0.643 | -0.041 | 0.066 |
| pct_tl_nvr_dlq | 0.0542 | 0.004 | 13.340 | 0.000 | 0.046 | 0.062 |
| percent_bc_gt_75 | -0.0049 | 0.002 | -3.135 | 0.002 | -0.008 | -0.002 |
| pub_rec_bankruptcies | 0.0005 | 0.157 | 0.003 | 0.997 | -0.308 | 0.309 |
| tax_liens | -0.0087 | 0.150 | -0.058 | 0.953 | -0.302 | 0.284 |
| tot_hi_cred_lim | 9.915e-07 | 2.64e-07 | 3.760 | 0.000 | 4.75e-07 | 1.51e-06 |
| total_bal_ex_mort | 1.852e-05 | 3.17e-06 | 5.845 | 0.000 | 1.23e-05 | 2.47e-05 |
| total_bc_limit | 1.253e-05 | 4.82e-06 | 2.601 | 0.009 | 3.09e-06 | 2.2e-05 |
| total_il_high_credit_limit | -1.682e-05 | 3.33e-06 | -5.044 | 0.000 | -2.34e-05 | -1.03e-05 |
| fico_average | 0.0014 | 0.001 | 1.866 | 0.062 | -7.21e-05 | 0.003 |
| fico_group | -0.0873 | 0.016 | -5.376 | 0.000 | -0.119 | -0.055 |

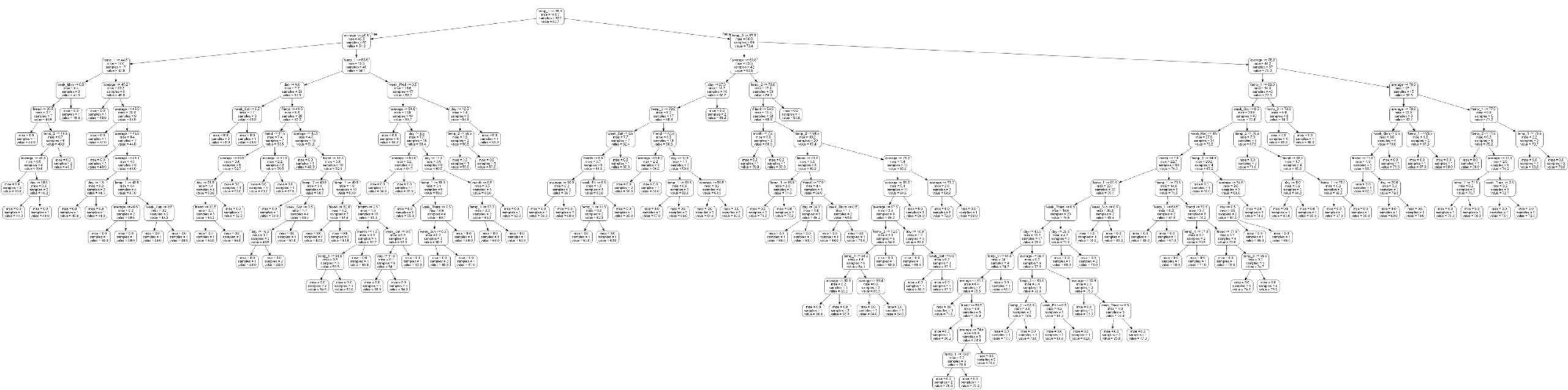
Logistic Regression – Complex Model



Logistic Regression – Complex Model Confusion Matrix

| prediction | 0 | 1 |
|-------------------|---------------|--------------|
| y_test | | |
| 0 | 201846 | 268 |
| 1 | 1265 | 49802 |

Random Forest Model– Decision Process



Random Forest Model– Confusion Matrix

| prediction | 0 | 1 |
|-------------------|---------------|--------------|
| y_test | | |
| 0 | 202058 | 7 |
| 1 | 241 | 50875 |

Conclusion

1. Introduction
2. Data Source and Review
3. Explanatory Analysis
4. Model Result
5. Conclusion

Conclusion

1. We see huge improving in using random forest model.
2. We can continue to work on different machine learning models on different data set and different scenarios that need to predict binary outcomes.