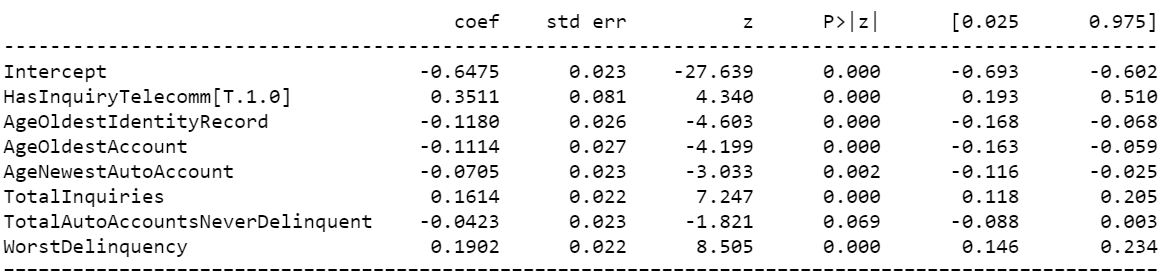
2)

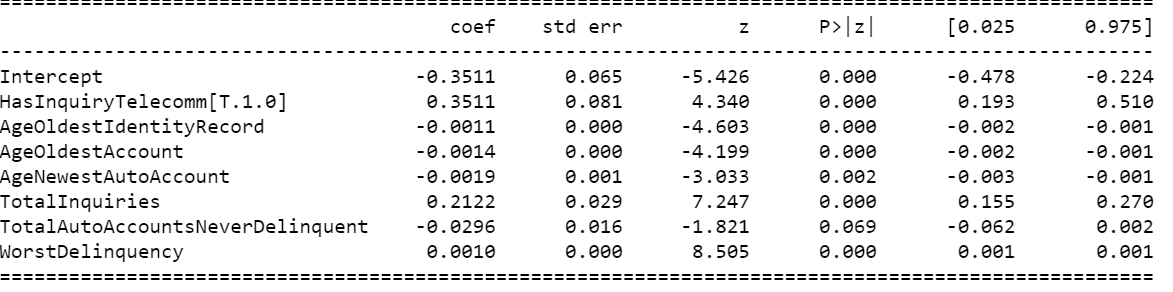
1. Identify the strongest predictor variables and provide interpretations

**All Interpretation are derived considering P-value =0.05**



The above picture provides the summary of logistic Regression model with Normalized Variables i.e stand\_model and hence the coefficients values can be used to compare impact of each variable.

From the above results we can see the strongest predictor is “HasInquiryTelecom” which denotes whether one or more telecommunications credit inquiries are on record within the last 12 months.



The above model summary is from corrected\_model which was generated after removing multicollinearity, but the variables are not normalized here so it cannot be used for comparing between parameters for their relative importance but can reasonably utilized for drawing interpretations.

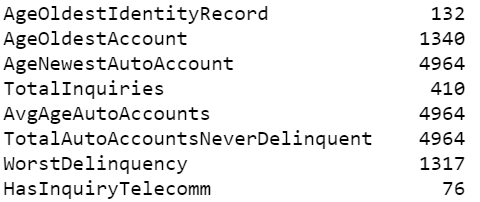
(Exp(coef)-1)\*100 gives the % impact of unit increase in continuous variable on the odds of a loan going delinquent within the first year of the loan's life.

Interpretations:

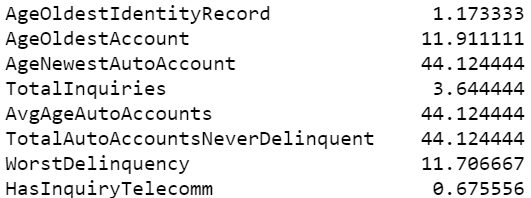
1. TotalAutoAccountsNeverDelinquent has no impact on the FirstYearDelinquency
2. Loans with “HasInquiryTelecomm” = 1 has 42% more chances of FirstYearDelinquency as compared to “HasInquiryTelecomm” = 0
3. 100 months increase in “AgeOldestIdentityRecord” will decrease the odds of FirstYearDelinquency by ~11%
4. 100 months increase in “AgeOldestAccount” will decrease the odds of FirstYearDelinquency by ~14%.
5. 100 months increase in “AgeNewestAutoAccount” will decrease the odds of FirstYearDelinquency by ~19%
6. A unit increase in “TotalInquiries” will result in increase of odds of FirstYearDelinquency by ~24%.
7. 100 days increase in “WorstDelinquency” will result in increase of odds of FirstYearDelinquency by ~10%
8. Identify and explain any issues or caveats with the data and the model(s)

* The data has a missing value in all the variables.

Count of missing values in each variable.



% of missing values



* The target variable is imbalanced. We have more observations of “0” class as compared to “1” class.

% of observation in each class.



* “WorstDelinquency” feature is right censored with 400 upper limit which can cause bias.
* Max data has “HasInquiryTelecomm” under “class 0” which will bias the model learning towards “class 0”.
* The data is not sufficient to capture the information and generalize it. That is why, the accuracy of both the models is very low.
* The variables “AgeNewestAutoAccount” and “AvgAgeAutoAccounts” have high correlation.
* For prediction purpose, high correlation is not a problem but for interpretations of parameter, I dropped “AvgAgeAutoAccounts” variable out of the Regression.
* The parameter importance provided by Random Forest is not reliable.
* The pseudo-R-square of logistic regression model is 0.021 which denotes that predictor variables are not providing enough information with regards to target variable.

1. Calculate predictions and show model performance on out-of-sample data?

Train data- 9000 observations, 8 features

Test data – 2250 observations, 8 features

Class 0: Lower Risk accounts (Probability<0.5)

Class 1: Higher-Risk accounts (Probability>=0.5)

Random Forest:

Train Accuracy: 0.6486

Test Accuracy: 0.648

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
|  | Prediction | |
| Actual | 0 | 1 |
| 0 | 1450 | 5 |
| 1 | 787 | 8 |

AUC: 0.58

Logistic Regression:

Train Accuracy: 0.6493

Test Accuracy: 0.655

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
|  | Prediction | |
| Actual | 0 | 1 |
| 0 | 1412 | 43 |
| 1 | 733 | 62 |

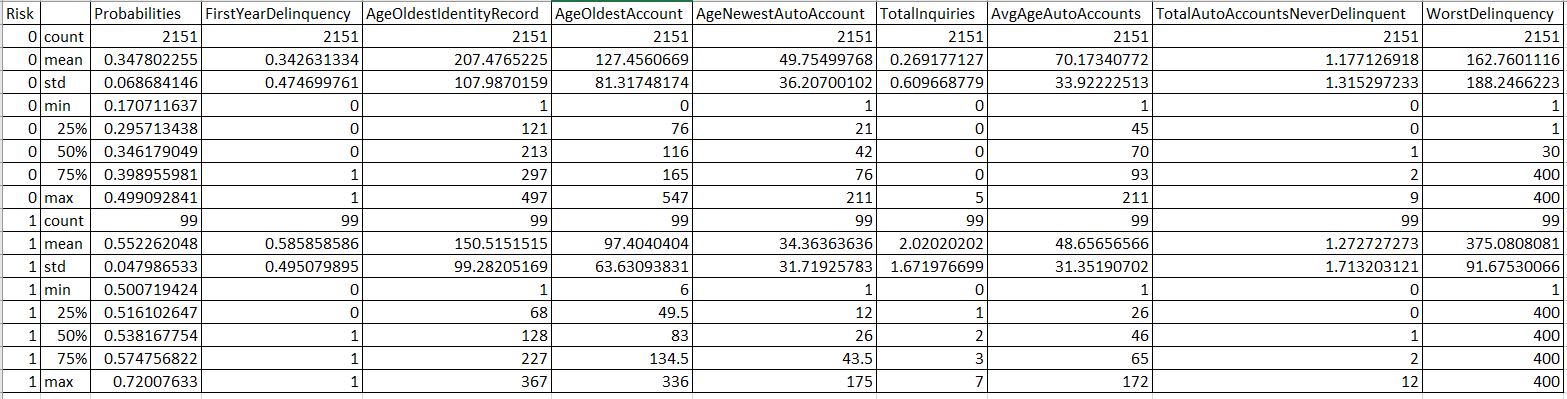
AUC: 0.52

1. Summarize results in a fashion that differentiates lower-risk accounts from higher-risk accounts

Risk

Class 0: Lower Risk accounts (Probability<0.5)

Class 1: Higher-Risk accounts (Probability>=0.5)



If above chart image int not clear please refer to “Risk\_Analysis” csv file in the zip folder generated by python script.

