Credit and Bankruptcies

By Yeon Hwa Choi

Questions

Will a customer file a bankruptcy in the future?

Will it be safe to provide a loan to the customer?

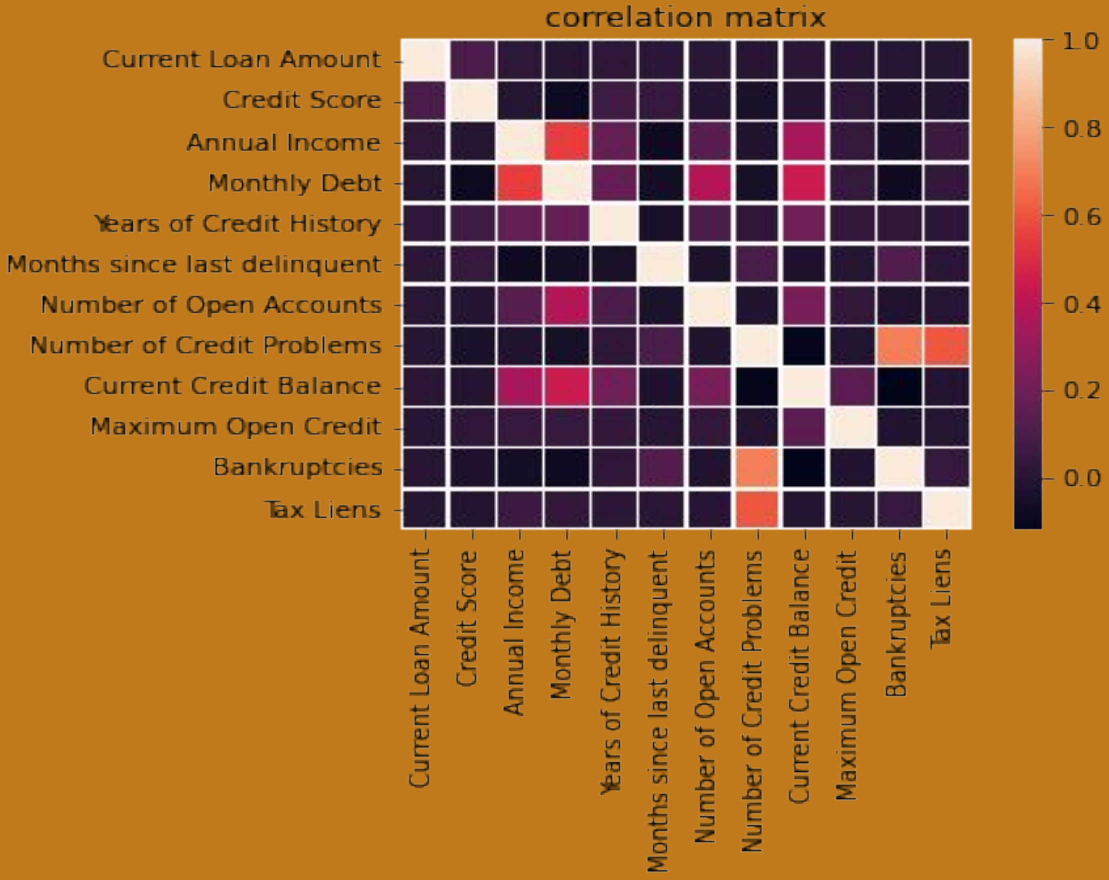
Dataset

Total number of data: 100,000

Number of Columns: 17

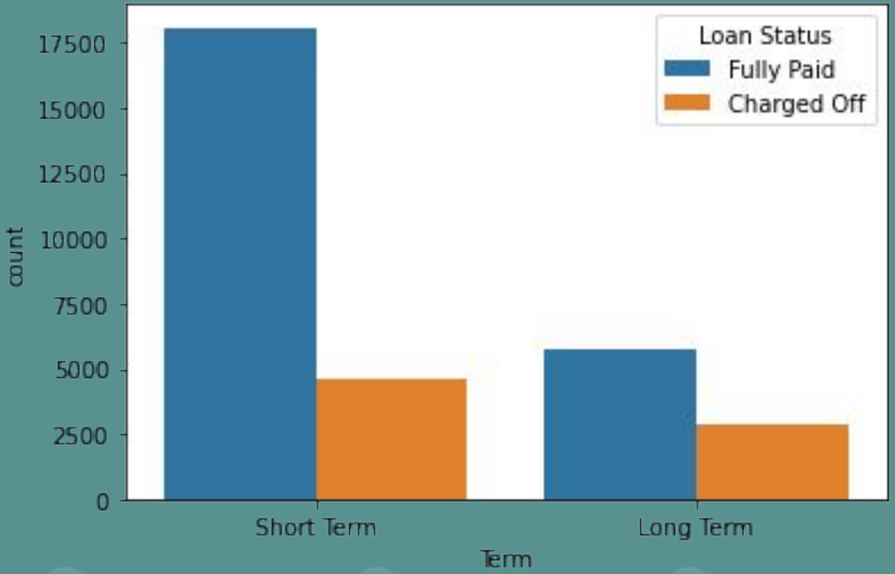
Columns: Loan Status, Current Loan Amount, Term, Credit Score, Annual Income, Years in Current Job, Home Ownership, Purpose, Monthly Debt, Years of Credit History, Months Since Last Delinquent, Number of Open Accounts, Number of Credit Problems, Current Credit Balance, Maximum Open Credit, Bankruptcies, Tax Liens

Correlation Table Using Heatmap



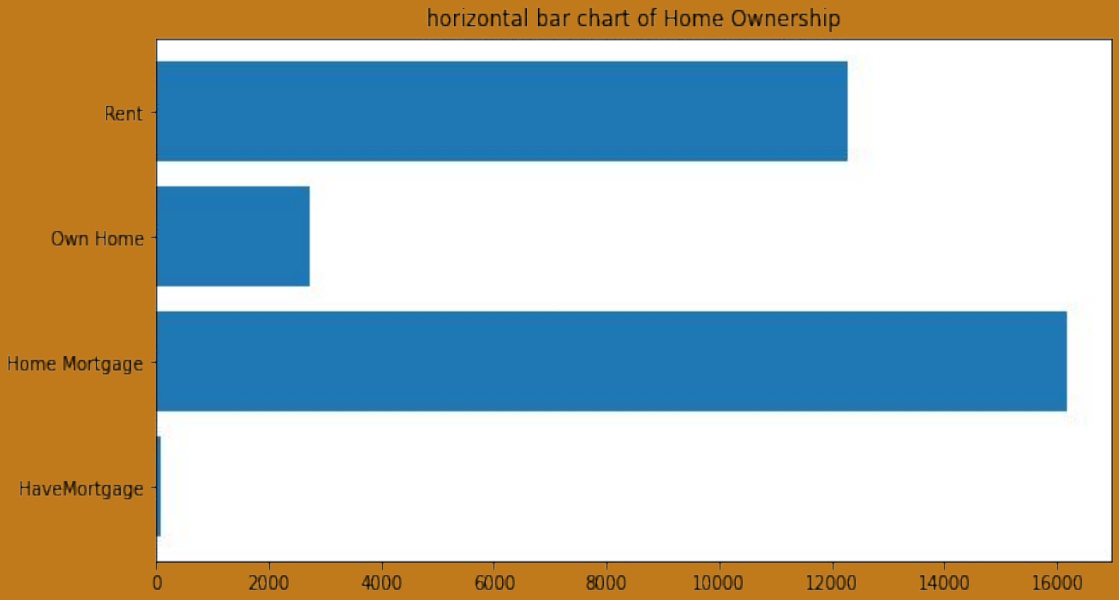
- Monthly Debt and Annual Income have comparably correlated.
- Number of Credit Problems is comparably correlated with Tax Liens
- Bankruptcies and Number of Credit Problems have the highest correlation.

Loan Status Grouped by Term



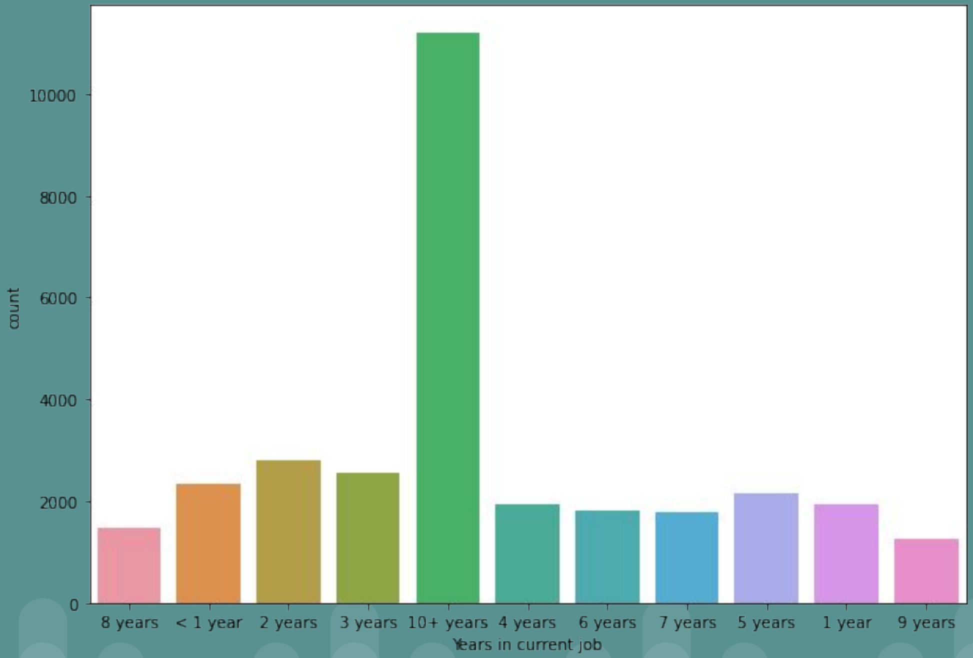
- More short term loans than long term loans are borrowed.
- The number of fully paid short term loans are significantly greater than fully paid long term.

Home Ownership



- Not a lot of people own their houses.
- A lot of people rent their current residence.
- The largest percentage of home ownership is having mortgage.

Years in Current Job

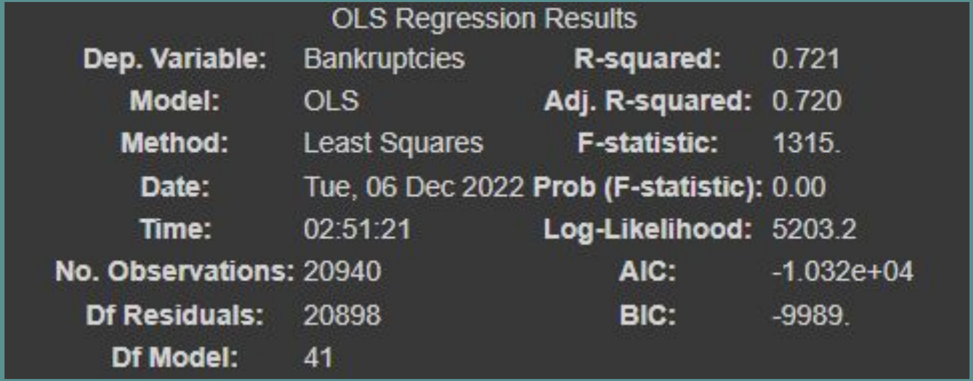


- Surprisingly significant number of people have been in their current job more than 10 years.
- Rest of the data are comparably the same.

X = all columns except Bankruptcies

Y = Bankruptcies (Dependent Variable)

Ordinary Least Squares



```
                      OLS Regression Results
   Dep. Variable:    Bankruptcies          R-squared:       0.721
       Model:        OLS              Adj. R-squared:       0.720
      Method:        Least Squares         F-statistic:     1315.
       Date:         Tue, 06 Dec 2022 Prob (F-statistic):   0.00
       Time:         02:51:21         Log-Likelihood:       5203.2
No. Observations:    20940                   AIC:           -1.032e+04
  Df Residuals:      20898                   BIC:           -9989.
    Df Model:        41
```

- R-squared value is 0.721. Means 72.1% of the variance for Y variable is explained by X variables.
- F-statistics is 1,315.

Decision Tree

Cross validation of 5 folds

Score Average: 0.895675786

Total time to read in raw data: 0.07 seconds.

More favorable result compared to OLS.

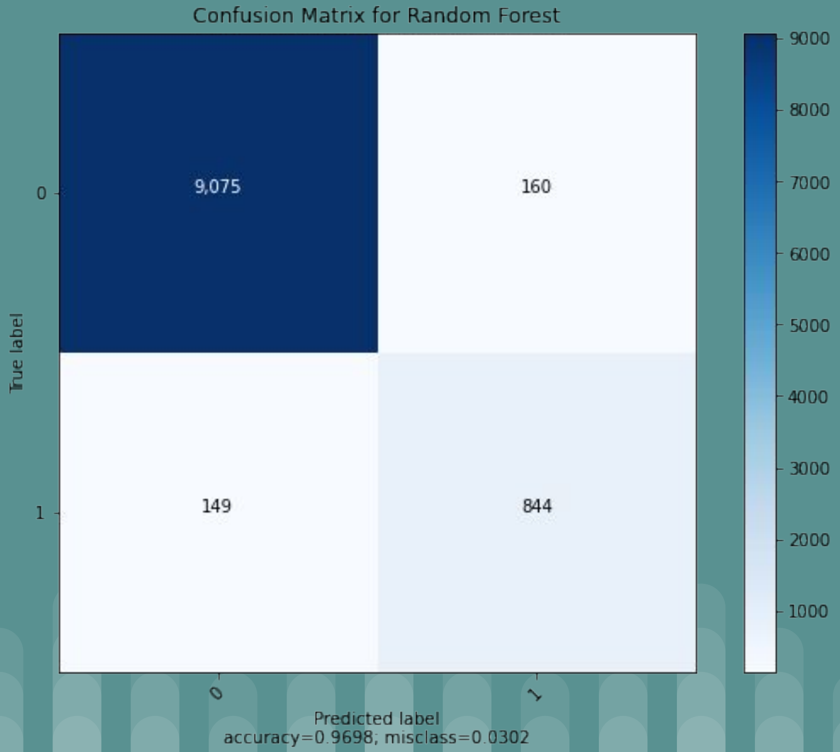
Random Forest

Cross validation 5 folds

Average Score: 0.952977196

Total time to read in raw data: 2.38 seconds.

Longer time but better score

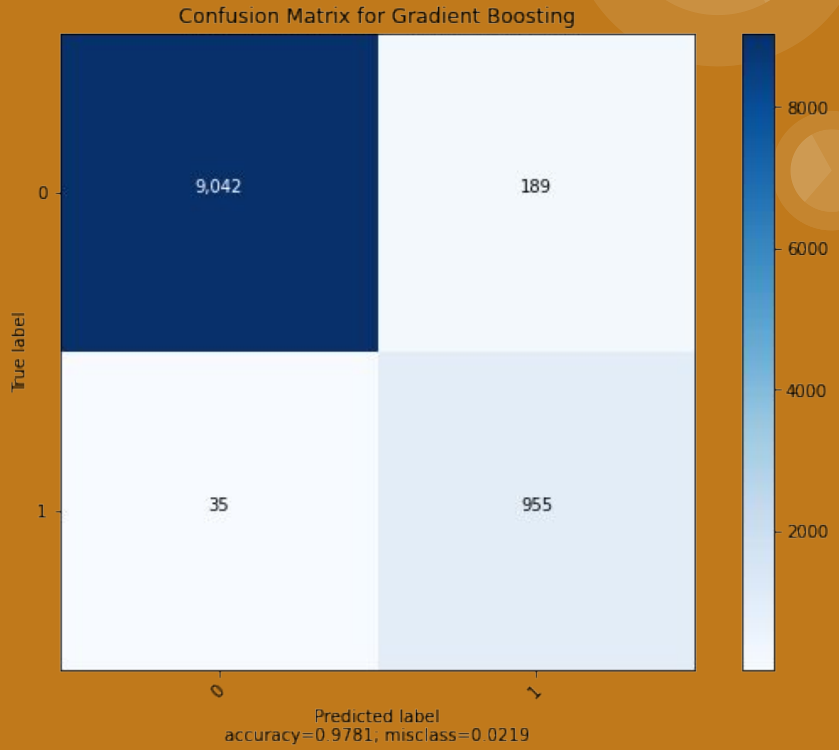


Gradient Boosting

Score: 0.9705254993213108

Total time to read in raw data: 35.21 seconds.

Longest time to run but the best result.



Confusion Matrix for Gradient Boosting

Conclusion

Four different machine learning models has been used for the analysis: **Ordinary Least Squares**, **Decision Tree**, **Random Forest**, and **Gradient Boosting**. Based on my research, **Gradient Boosting** has performed the best result with the highest accuracy among all four models.

Recommendation: Gradient Boosting with Number of estimator 100, Max Depth 2, and Loss Deviation to predict if the person will bankrupt in near future. It no bankruptcy was predicted, then it will be safe to provide the loan.