Credit Card Application Risk Assessment and Prediction

ANNA F HARVEY

Bellevue University

DSC 680: Applied Data Science

Prof. Catherine Williams

May 30, 2021

Abstract

Credit card companies are businesses. As such, their purpose is to make money from their customers. In order to do that, they must attract and maintain a customer base that will be loyal to them as well as provide a regular source of profit. Offering appealing credit card interest rates, rewards, and accessible balance amounts can encourage customers to sign up with a credit card company. However, due to the nature of the business, attracting customers who will contribute to the financial bottom-line is of primary importance.

One of the ways to assess the risk of a consumer is to analyze financial and demographic information. Certain variables can indicate whether someone will be able to maintain a credit card or if they will default. Although credit card companies benefit from a customer paying interest on a balance, they benefit less if that person cannot pay at all. Therefore, it is important to a credit card company to analyze the risk for potential customers. Understanding customers' potential risk can also help determine factors such as interest rates and introductory lines of credit. Machine learning techniques can be employed well for this task by using information from credit score cards. By analyzing different types of machine learning models, we can discover reliable methods of prediction to understand the financial situation of potential credit card customers.

Introduction

` Credit cards are an inevitable part of financial culture, particularly in the United States. We have created a system that virtually requires an individual to build credit in order to partake in life's main financial goals, such as buying a car or a house. One of the easiest ways to do this is by opening a credit card. Credit cards are used as safety nets by many and when used responsibly, they can be a reliable source of assistance.[1] However, a credit card that is used responsibly does not benefit credit card issuers and networks. These companies make money from the calculated risk of providing someone with a line of credit that they will not be able to immediately pay back.

Credit card companies make money from three main sources: interest, cardholder fees, and transaction fees.[2] Interest accrued from a balance carried over month to month and cardholder fees make up the bulk of a credit card company's revenue with only a small portion coming from transaction fees that are paid by businesses where the cardholder is using their credit card. Therefore, the hope of a credit card company is that a cardholder will maintain a balance on their credit card and subject themselves to a certain amount of fees. The balance for the credit card company is assessing what kind of customer will need to maintain a balance but will not end up in default on the card.

When a customer defaults on a credit card, the outcome is worse for the customer than the company. Credit scores suffer and the customer continues to be buried in debt. However, the company will lose money as well if the customer is never unable to pay the debt back. Therefore, credit card companies will often be willing to negotiate various options with the customer to receive some amount of payment in exchange for outcomes such as reducing the card's effect on the customer's credit score.[3]

Problem Statement

Credit card companies need to assess the risk-level of their potential customers before entering a contract with them. Companies want to attract and maintain customers that will bring in revenue without risking the customer completely defaulting on the loan. Several different demographic and financial variables can be used to predict the likelihood of a customer defaulting on a credit card. Therefore, using the correct machine learning predictive algorithms along with the correct variables, a customer's risk level should be able to be predicted before the customer enters a contract with the company.

Methods

Two datasets were analyzed which focus on demographic and financial variables of credit card customers.[4] The first dataset (hereafter referred to as "application") includes seventeen different variables for potential credit card customers gathered from a financial application. These variables include marital status, income, home ownership, and employment information, among others. The second dataset (hereafter referred to as "credit") shows the credit status of various credit card customers. The dataset indicates whether or not a credit card loan has been repaid, and if not, how long it has been overdue.

The application dataset required cleaning. There were many duplicate customers represented in the data and the dataset ended up being reduced from 438,557 customers (rows) to 90,084 before combining the datasets together. Since it was unknown which of the seventeen variables might affect the overall outcome of a customer, only two variables were dropped: mobile phone ownership (every customer owned a mobile phone) and occupation type (vague descriptions of job types).

Several variables were visualized to gain understanding about the distribution of data. There did not seem to be significant outliers in any of the variables. There did seem to be several variables that were imbalanced, such as gender. Variables for “days employed” and “days birth” needed to be converted to years for visualization purposes but were maintained in their original format for modeling.

The credit dataset contained 45,985 different customer IDs but only 5,426 IDs overlapped with the application dataset after duplicate rows were removed. The credit dataset was grouped and counted to represent the number of months that each customer was in a specific status on their credit card loan. It was then pivoted and merged with the application dataset to form a new dataset that included all the demographic and financial information for the 5,426 common customers as well as their credit statuses.

In order to use predictive models on the data, a target variable needed to be created to indicate whether a customer was considered high-risk. Using information about the minimum and maximum values of the numeric variables as well as the quartile ranges, I created a function to assign a risk variable of 0 for low-risk and 1 for high-risk based on six variables and their potential outcomes. Customers were labeled high-risk in the following scenarios: the customer’s income was less than the 25th percentile, the customer was unemployed and not a pensioner, the customer had been overdue on credit card payments for 9 months or more (75th percentile), or they had never paid off the credit card at any time (if the months where the loan was paid off and the months there was no loan both equaled zero). Customers were labeled low-risk in the following scenarios: the number of months the credit card was paid off equaled the total number of months with the company, the number of months without a loan equaled the total number of months with the company, or any other scenario that did not fit the high-risk requirements.

PyCaret was utilized to compare various regression model performance as the target variable was binary. Bayesian Ridge (BR), Ridge Regression (Ridge), and Extra Trees Regressor (ET) were chosen as models to build, tune, and deploy. The initial R2 value for Bayesian Ridge was 0.2357, Ridge Regression was 0.2296, and Extra Trees Regressor was 0.9747.

Results and Conclusions

BR, Ridge, and ET models were created, tuned, evaluated, and predicted on the hold-out sample from the dataset. Each model, once evaluated, each had different features that were most important for the model (Figure 1). Common variables of most importance between all three models were: income type – pensioner, gender – female, and no work phone. Output from the prediction models showed improvement only from the Ridge model with an increased R2 value of 0.2613 (Figure 2). Interestingly, the Extra Trees model's R2 value decreased significantly to 0.2087 (Figure 3).

Although the R2 values are low, this may indicate a high level of variation between the customers rather than a poorly fit model[5]. This would be a strong possibility considered the wide variation in the important features between each model as well as the general unpredictable nature of this type of data. Someone with a high-income may default on a credit card because they have overextended their investments. Someone with a low-income may continue to make timely payoffs because they are using their credit card to earn rewards.

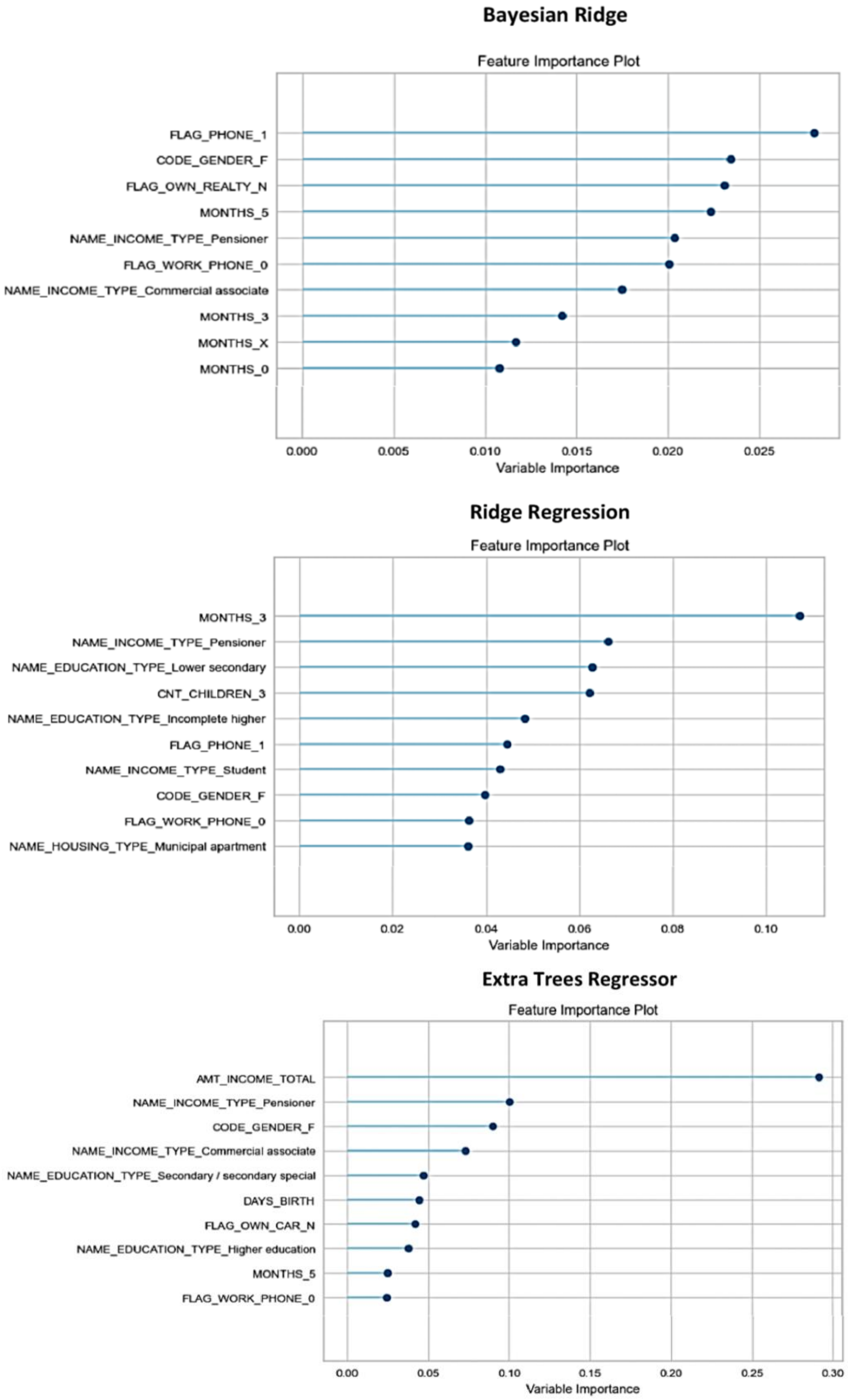
However, the choice of variables to use as indicators of risk may have also played a role in the low prediction accuracy.

Future Investigations

This project would have benefitted from a larger number of customers in the final dataset. There are also variables that could be considered that may contribute to a person's credit payment ability. An important one would be whether the potential customer has any other debts (student loans, other credit cards, mortgages). Aggregated debt arguably plays a more significant factor in debt default than many other variables. In the current era, where much of the working class is made up of millennials with significant student loan debt, this is a factor that should not be ignored if a company wants an accurate prediction of future debt default.

APPENDIX

Figure 1. Top 10 features for each model



Bayesian Ridge

Ridge Regression

Extra Trees Regressor

Figure 2: Prediction error visualizations for Regression models

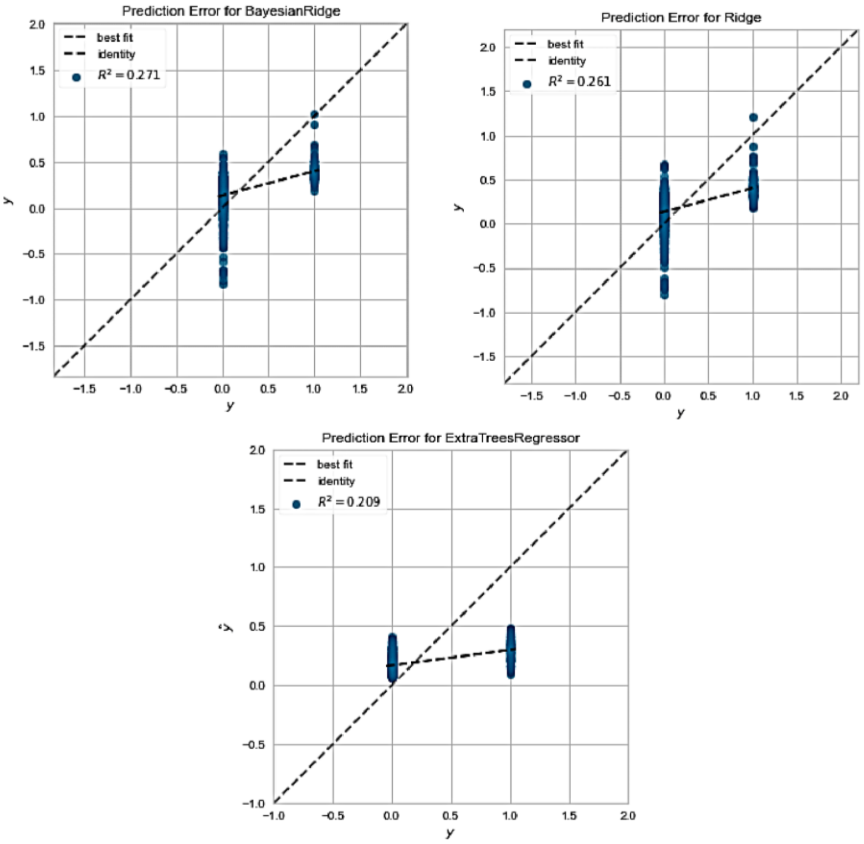
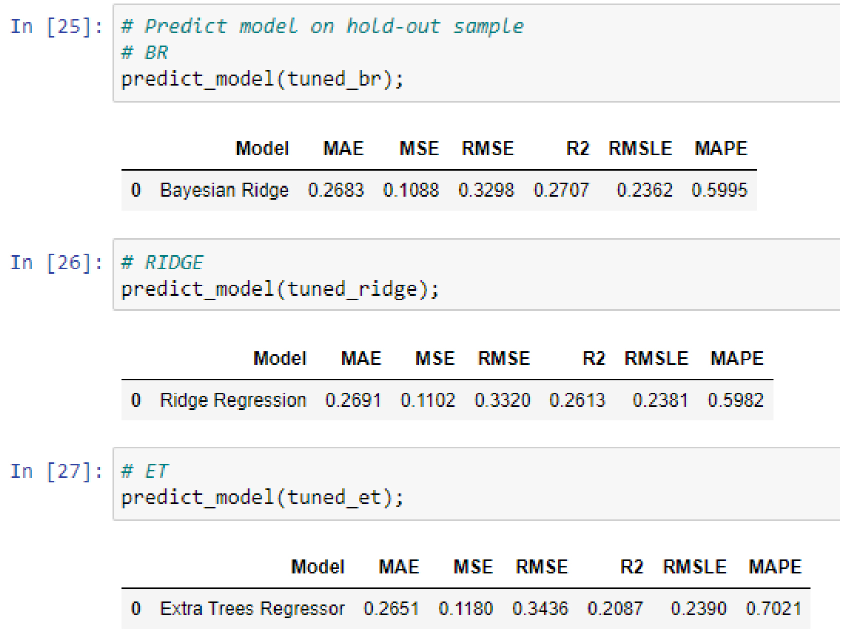


Figure 3: PyCaret Regression model predictions



In [25]:
```
# Predict model on hold-out sample
# BR
predict_model(tuned_br);
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Bayesian Ridge	0.2683	0.1088	0.3298	0.2707	0.2362	0.5995

In [26]:
```
# RIDGE
predict_model(tuned_ridge);
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Ridge Regression	0.2691	0.1102	0.3320	0.2613	0.2381	0.5982

In [27]:
```
# ET
predict_model(tuned_et);
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Extra Trees Regressor	0.2651	0.1180	0.3436	0.2087	0.2390	0.7021

RESOURCES

1. Kurt, D. (2021, May 7). Should I Get a Credit Card? Investopedia. https://www.investopedia.com/should-i-get-a-credit-card-4589811.

2. Lambarena, M. (2021, May 17). How Do Credit Card Companies Make Money? NerdWallet. https://www.nerdwallet.com/article/credit-cards/credit-card-companies-money.

3. Irby, L. T. (n.d.). Learn How Credit Card Default Happens and What You Can Do About It. The Balance. https://www.thebalance.com/what-is-credit-card-default-960209.

4. Credit Card Approval Prediction. Kaggle. (n.d.). https://www.kaggle.com/rikdifos/credit-card-approval-prediction.

5. Frost, J. (2021, April 23). How To Interpret R-squared in Regression Analysis. How to Interpret R-squared in Regression Analysis. https://statisticsbyjim.com/regression/interpret-r-squared-regression/.

https://github.com/anhar421/Portfolio

10 Questions

1. *What issues did you have with either/both dataset(s)?*

   The application dataset had an incredible number of duplicates. It is probable that some of the duplicate IDs were actually different people and those entries were just an error. However, the large number of identical rows with ID being the only difference are concerning. This dataset presumably came from real information, but it seems problematic that so many rows would have identical information.

2. *What variables did you choose to focus on to assess risk and why?*

   I chose to focus on income type, payment behavior, income total, and employment. With my limited domain knowledge, it seemed likely that the amount of income and type and length of employment would directly affect the ability to pay back a loan. Payment behavior was important to look at because it helps indicate if someone has or will default on a loan.

3. *What variables were not included in the datasets that would be helpful in risk assessment?*

   Other types of debt currently owned by the customer would be incredibly helpful. Someone with a high income may have so many different debts that they can still default on a credit card.

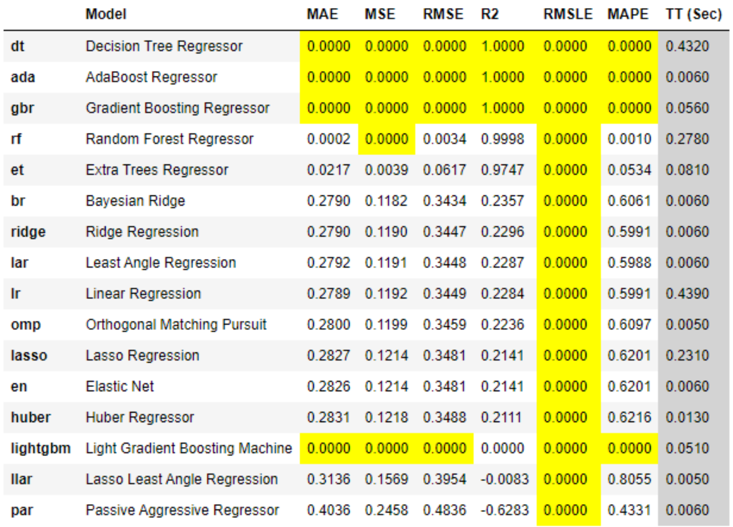
4. *What type of risk balance do credit card companies look for in a customer?*

   Companies ideally want a customer who is going to maintain a credit balance so that they can charge interest and fees. However, if the customer’s financial situation prevents them from paying at all and having to renegotiate or default, the company will lose money.

5. *What makes regression models suitable for credit card risk assessment?*

   Regression models work well for credit card risk assessment because the target variable for these predictions is binary. Generally, the target variable will be “default – yes or no” or “risk – high or low”.

6. *Did you encounter overfitting during model assessment?*



	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
dt	Decision Tree Regressor	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.4320
ada	AdaBoost Regressor	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0060
gbr	Gradient Boosting Regressor	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0560
rf	Random Forest Regressor	0.0002	0.0000	0.0034	0.9998	0.0000	0.0010	0.2780
et	Extra Trees Regressor	0.0217	0.0039	0.0617	0.9747	0.0000	0.0534	0.0810
br	Bayesian Ridge	0.2790	0.1182	0.3434	0.2357	0.0000	0.6061	0.0060
ridge	Ridge Regression	0.2790	0.1190	0.3447	0.2296	0.0000	0.5991	0.0060
lar	Least Angle Regression	0.2792	0.1191	0.3448	0.2287	0.0000	0.5988	0.0060
lr	Linear Regression	0.2789	0.1192	0.3449	0.2284	0.0000	0.5991	0.4390
omp	Orthogonal Matching Pursuit	0.2800	0.1199	0.3459	0.2236	0.0000	0.6097	0.0050
lasso	Lasso Regression	0.2827	0.1214	0.3481	0.2141	0.0000	0.6201	0.2310
en	Elastic Net	0.2826	0.1214	0.3481	0.2141	0.0000	0.6201	0.0060
huber	Huber Regressor	0.2831	0.1218	0.3488	0.2111	0.0000	0.6216	0.0130
lightgbm	Light Gradient Boosting Machine	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0510
llar	Lasso Least Angle Regression	0.3136	0.1569	0.3954	-0.0083	0.0000	0.8055	0.0050
par	Passive Aggressive Regressor	0.4036	0.2458	0.4836	-0.6283	0.0000	0.4331	0.0060

The first four models assessed by PyCaret were overfit.

7. *What might have caused overfitting of the models?*

My target variable creation may have contributed to the overfitting as well as the number of customers being assessed from the final dataset.

8. *How could you avoid overfitting for this type of assessment in the future?*

Better domain knowledge of the dataset would likely help me create a more accurate target variable without forcing the model into too accurate predictions.

9. *What methods can be used to collect data for this type of assessment?*

Credit card applications, bank account applications, and other information given during other loan applications could all be used to collect data.

10. *How could this type of risk assessment be used to benefit customers as well as creditors?*

Credit card companies are most thought of when considering the benefits of default prediction. However, if that information was made available to customers as well, a customer could potentially make an educated decision on whether or not to take out a credit card.