# Supervised Machine Learning: Regression Course Project: PPP Loan Data

Report by Jennifer Case
7/12/2022

## Objective

At the start of the COVID-19 pandemic, the US government offered Paycheck Protection Program (PPP) loans through the Small Business Administration (SBA) to help pay business costs and to increase job retention due to the pandemic-related shutdown. Many businesses across industry took advantage of the program.

In this report, the main objective is to understand what factors had the largest influence on the loan approval amount received by a business. If there are strong correlations in how most businesses requested loans, this knowledge may assist in developing fraud detection models in the future. To study this problem, a subset of the total PPP loan data released by the SBA is analyzed using interpretation-focused regression models.

## Data Description

The PPP loan data[1] is broken out into two categories: loans above $150k and loans equal to or below $150k. Since fraud patterns are likely to differ based on loan size (e.g., does an individual apply for one large fraudulent loan or a series of small fraudulent loans?), only the subset of loans above $150k is considered in this analysis.

There are 986,532 records for loans above $150k with the following 53 attributes:

| | Attribute Name | Attribute Descriptions | Data Type |
|---|---|---|---|
| 1 | LoanNumber | Loan Number (unique identifier) | int64 |
| 2 | DateApproved | Loan Funded Date | object |
| 3 | SBAOfficeCode | SBA Origination Office Code | int64 |
| 4 | ProcessingMethod | Loan Delivery Method (PPP for first draw; PPS for second draw) | object |
| 5 | BorrowerName | Borrower Name | object |
| 6 | BorrowerAddress | Borrower Street Address | object |
| 7 | BorrowerCity | Borrower City | object |
| 8 | BorrowerState | Borrower State | object |
| 9 | BorrowerZip | Borrower Zip Code | object |
| 10 | LoanStatusDate | Loan Status Date<br>- Loan Status Date is blank when the loan is disbursed but not Paid In Full or Charged Off | object |

[1] "PPP FOIA: Data and Resources." *U.S. Small Business Administration*, 4 April 2022, https://data.sba.gov/dataset/ppp-foia.

| | | | |
|---|---|---|---|
| 11 | LoanStatus | Loan Status Description<br>- Loan Status is replaced by 'Exemption 4' when the loan is disbursed but not Paid in Full or Charged Off | object |
| 12 | Term | Loan Maturity in Months | int64 |
| 13 | SBAGuarantyPercentage | SBA Guaranty Percentage | int64 |
| 14 | InitialApprovalAmount | Loan Approval Amount (at origination) | float64 |
| 15 | CurrentApprovalAmount | Loan Approval Amount (current) | float64 |
| 16 | UndisbursedAmount | Undisbursed Amount | float64 |
| 17 | FranchiseName | Franchise Name | object |
| 18 | ServicingLenderLocationID | Lender Location ID (unique identifier) | int64 |
| 19 | ServicingLenderName | Servicing Lender Name | object |
| 20 | ServicingLenderAddress | Servicing Lender Street Address | object |
| 21 | ServicingLenderCity | Servicing Lender City | object |
| 22 | ServicingLenderState | Servicing Lender State | object |
| 23 | ServicingLenderZip | Servicing Lender Zip Code | object |
| 24 | RuralUrbanIndicator | Rural or Urban Indicator (R/U) | object |
| 25 | HubzoneIndicator | Historically Underutilized Business zone (Hubzone) Indicator (Y/N) | object |
| 26 | LMIIndicator | Low- and Moderate-Income (LMI) Indicator (Y/N) | object |
| 27 | BusinessAgeDescription | Business Age Description | object |
| 28 | ProjectCity | Project City | object |
| 29 | ProjectCountyName | Project County Name | object |
| 30 | ProjectState | Project State | object |
| 31 | ProjectZip | Project Zip Code | object |
| 32 | CD | Project Congressional District | object |
| 33 | JobsReported | Number of Employees | float64 |
| 34 | NAICSCode | North American Industry Classification System (NAICS) 6 digit code | float64 |
| 35 | Race | Borrower Race Description | object |
| 36 | Ethnicity | Borrower Ethnicity Description | object |
| 37 | UTILITIES_PROCEED | Note: Proceed data is lender reported at origination. On the PPP application the proceeds fields were check boxes. | float64 |
| 38 | PAYROLL_PROCEED | | float64 |
| 39 | MORTGAGE_INTEREST_PROCEED | | float64 |
| 40 | RENT_PROCEED | | float64 |
| 41 | REFINANCE_EIDL_PROCEED | | float64 |
| 42 | HEALTH_CARE_PROCEED | | float64 |
| 43 | DEBT_INTEREST_PROCEED | | float64 |
| 44 | BusinessType | Business Type Description | object |
| 45 | OriginatingLenderLocationID | Originating Lender ID (unique identifier) | int64 |
| 46 | OriginatingLender | Originating Lender Name | object |
| 47 | OriginatingLenderCity | Originating Lender City | object |
| 48 | OriginatingLenderState | Originating Lender State | object |
| 49 | Gender | Gender Indicator | object |
| 50 | Veteran | Veteran Indicator | object |

| | | | |
|---|---|---|---|
| 51 | NonProfit | 'Yes' if Business Type = Non-Profit Organization or Non-Profit Childcare Center or 501(c) Non Profit | object |
| 52 | ForgivenessAmount | Forgiveness Amount | float64 |
| 53 | ForgivenessDate | Forgiveness Paid Date | object |

In addition to the PPP loan data, this analysis will also consider NAICS data[2] that specifies the number of businesses that fall under various industries to enable proper scaling of the number of businesses that received PPP loans over $150k per industry given the total number of businesses in that industry. To join the data, the NAICSCode attribute that contains a 6 digit NAICS code, which indicates a specific industry, in the PPP loan data will be converted to a 2 digit NAICS code, which indicates a general industry, and connected to the NAICS data on the Code attribute. The goal of identifying industries is to enable the stratification of the data by industry. That data set will include:

| | **Attribute Name** | **Attribute Descriptions** | **Data Type** |
|---|---|---|---|
| 1 | Code | NAICS 2 digit code | int64 |
| 2 | IndustryTitle | Industry Title | object |
| 3 | NumBusinesses | Number of Business Establishments | object |

## Data Exploration

During data exploration, it appears as though each industry has its own trends for the loan approval amount (CurrentApprovalAmount). For this reason, a single industry, Mining, was targeted for the analysis. Mining was chosen because it appeared to be disproportionately affected by shutdown compared to the other industries, as shown in Figure 1.

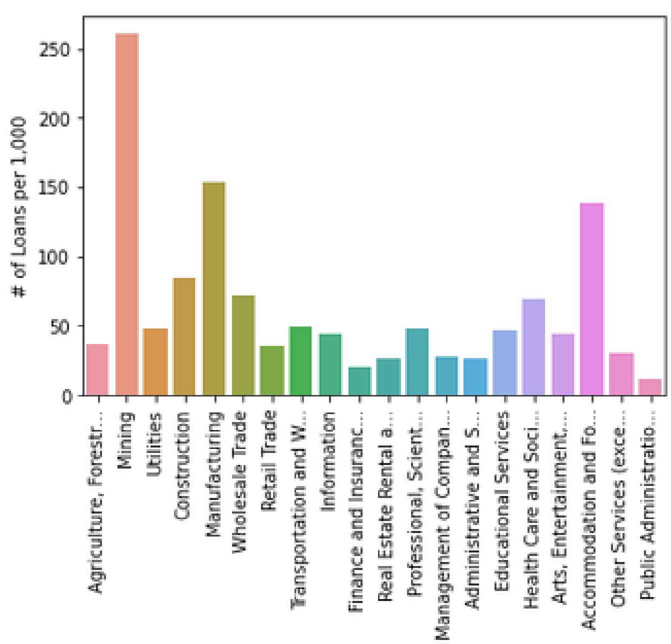


*Figure 1. Distribution of the number of loans per 1,000 businesses by industry.*

[2] "NAICS & SIC Identification Tools." *NAICS Association*, https://www.naics.com/search/#naics.

## Data Cleaning and Feature Engineering

The following steps were taken to clean the data:

- Evaluated for duplicates via ensuring that the loan numbers were unique. No duplicate entries were found.
- The following columns are being trimmed from the data set:

| | Attribute Name | Reason for Removal |
|---|---|---|
| 1 | LoanNumber | Unique value not helpful to find patterns. |
| 3 | SBAOfficeCode | Does not matter where SBA originally processed form. |
| 5 | BorrowerName | Unique value not helpful to find patterns. |
| 6 | BorrowerAddress | Unique value not helpful to find patterns. |
| 7 | BorrowerCity | Manual entry errors. |
| 9 | BorrowerZip | Too individualized for mining companies. |
| 10 | LoanStatusDate | Not factor considered when loan was approved. |
| 11 | LoanStatus | Not factor considered when loan was approved. |
| 13 | SBAGuarantyPercentage | SBA guaranty percentage is 100% for all loans. |
| 14 | InitialApprovalAmount | Almost equivalent to CurrentApprovalAmount and will dominate the analysis. |
| 16 | UndisbursedAmount | Only 17 samples have undisbursed amounts, so the sample is not large enough to learn anything meaningful. |
| 17 | FranchiseName | Only a single entry listed a franchise. |
| 18 | ServicingLenderLocationID | Not considering service lender information in analysis. |
| 19 | ServicingLenderName | Not considering service lender information in analysis. |
| 20 | ServicingLenderAddress | Not considering service lender information in analysis. |
| 21 | ServicingLenderCity | Not considering service lender information in analysis. |
| 22 | ServicingLenderState | Not considering service lender information in analysis. |
| 23 | ServicingLenderZip | Not considering service lender information in analysis. |
| 28 | ProjectCity | Manual entry errors. |
| 29 | ProjectCountyName | Too individualized for mining companies. |
| 31 | ProjectZip | Too individualized for mining companies. |
| 32 | CD | Too individualized for mining companies. |
| 37 | UTILITIES_PROCEED | The lender provided data is inconsistent |
| 38 | PAYROLL_PROCEED | The lender provided data is inconsistent |
| 39 | MORTGAGE_INTEREST_PROCEED | The lender provided data is inconsistent |
| 40 | RENT_PROCEED | The lender provided data is inconsistent |
| 41 | REFINANCE_EIDL_PROCEED | The lender provided data is inconsistent |
| 42 | HEALTH_CARE_PROCEED | The lender provided data is inconsistent |
| 43 | DEBT_INTEREST_PROCEED | The lender provided data is inconsistent |
| 45 | OriginatingLenderLocationID | Not considering lender information in analysis |
| 46 | OriginatingLender | Not considering lender information in analysis |
| 47 | OriginatingLenderCity | Not considering lender information in analysis |
| 48 | OriginatingLenderState | Not considering lender information in analysis |
| 51 | NonProfit | Data captured in BusinessType |
| 52 | ForgivenessAmount | Not factor considered when loan was approved |
| 53 | ForgivenessDate | Not factor considered when loan was approved |

- The null values for the remaining columns were treated as follows:

| | Attribute Name | Treatment of Null and N/A Values |
|---|---|---|
| 2 | DateApproved | N/A |
| 4 | ProcessingMethod | N/A |
| 8 | BorrowerState | Rows removed |
| 12 | Term | N/A |
| 15 | CurrentApprovalAmount | N/A |
| 24 | RuralUrbanIndicator | N/A |
| 25 | HubzoneIndicator | N/A |
| 26 | LMIIndicator | N/A |
| 27 | BusinessAgeDescription | Null values moved to the pre-existing 'Unanswered' category |
| 30 | ProjectState | Rows removed |
| 33 | JobsReported | Rows removed |
| 34 | NAICSCode | Null values changed to zero to be translated to 'Unanswered' in the Industry column |
| 35 | Race | N/A |
| 36 | Ethnicity | N/A |
| 44 | BusinessType | Null values moved to an 'Unanswered' category |
| 49 | Gender | N/A |
| 50 | Veteran | N/A |

- The following column was added and the NAICSCode column was removed:

| | Attribute Name | Attribute Descriptions | Data Type |
|---|---|---|---|
| | Industry | Industry Title based on NAICSCode | object |

- The data was reduced to only include samples where the industry was Mining leaving 8,411 samples.
- The DateApproved column was broken into three columns: ApprovalDay, ApprovalMonth, and ApprovalYear.
- One-hot encoding was used on the object columns resulting in a total of 141 columns:
    - ProcessingMethod becomes a single column where 1 indicates PPP and 0 indicates PPS.
    - BorrowerState becomes 52 columns to indicate one of the following 53 options: AK, AL, AR, AZ, CA, CO, CT, DC, DE, FL, GA, HI, IA, ID, IL, IN, KS, KY, LA, MA, MD, ME, MI, MN, MO, MS, MT, NC, ND, NE, NH, NJ, NM, NV, NY, OH, OK, OR, PA, PR, RI, SC, SD, TN, TX, UT, VA, VI, VT, WA, WI, WV, WY.
        - Note that there are more than 50 options because DC indicates District of Columbia, PR indicates Puerto Rico, and VI indicates Virgin Islands[3].
    - RuralUrbanIndicator becomes a single column where 1 indicates Rural and 0 indicates Urban.
    - HubzoneIndicator becomes a single column where 1 indicates a HUBZone.
    - LMIIndicator becomes a single column where 1 indicates an LMI business.

[3] "Two-Letter State and Territory Abbreviations." *Federal Aviation Administration*, https://www.faa.gov/air_traffic/publications/atpubs/cnt_html/appendix_a.html.

- BusinessAgeDescription becomes 4 columns to indicate one of the following 5 options: Change of Ownership, Existing or more than 2 years old, New Business or 2 years or less, Startup, Unanswered.
    - ProjectState becomes 52 columns to indicate one of the following 53 options: AK, AL, AR, AZ, CA, CO, CT, DC, DE, FL, GA, HI, IA, ID, IL, IN, KS, KY, LA, MA, MD, ME, MI, MN, MO, MS, MT, NC, ND, NE, NH, NJ, NM, NV, NY, OH, OK, OR, PA, PR, RI, SC, SD, TN, TX, UT, VA, VI, VT, WA, WI, WV, WY.
    - Race becomes 5 columns to indicate one of the following 6 options: American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, or Unanswered.
    - Ethnicity becomes 2 columns to indicate one of the following 3 options: Hispanic or Latino, Not Hispanic or Latino, or Unknown/NotStated.
    - BusinessType becomes 12 columns to indicate one of the following 13 options: Cooperative, Corporation, Employee Stock Ownership Plan (ESOP), Limited Liability Company (LLC), Limited Liability Partnership, Non-Profit Organization, Partnership, Professional Association, Self-Employed Individuals, Sole Proprietorship, Subchapter S Corporation, Tribal Concerns, or Unanswered.
    - Gender becomes 2 columns to indicate one of the 3 options: Female Owned, Male Owned, or Unanswered.
    - Veteran becomes 2 columns to indicate one of the 3 options: Veteran, Non-Veteran, or Unanswered.
- Identify and remove duplicated columns. There were 48 duplicated columns where the BorrowerState columns matched the ProjectState columns.
- The JobsReported data was transformed using a log transformation to reduce skewness of the data.
- The target variable, CurrentApprovalAmount, was transformed using a log transformation to reduce skewness and improve the performance of the regression models.
- The non-target variable data was scaled with a MinMaxScaler.

# Linear Regression Models

Three models were evaluated for interpreting the data: (1) a simple linear regression model, (2) a Ridge regression model with polynomial effects, and (3) a LASSO regression model with polynomial effects. To train and test the models, the data was split into 70% training data and 30% testing data.

For the simple linear regression model, no cross-validation was done since there were no hyperparameters for this model. For the other two models, cross-validation was done to select hyperparameters, specifically the degree of polynomial effects (degree = {1,2}) and the lambda ($\lambda$={0.001, 0.01, 0.1, 1, 10, 100}) for regularization. The cross-validation used a K-fold approach with 4 subsamples across the testing data.

## Simple Linear Regression Model

The simple linear regression model failed to predict the approval amount for loans. The $R^2$ score for the testing data was negative indicating that the model performed worse than assuming every loan is equal to the average of the dataset.

Adding polynomial affects to a simple linear regression model also resulted in a negative $R^2$ score indicating that regularization is needed to improve model accuracy.

## Ridge Regression Model with Polynomial Effects

A cross-validated grid search found that polynomial effects of degree = 2 and λ = 10 were the best parameters for a Ridge regression model. The $R^2$ score was 0.79 and the distribution of true vs. predicted approval amounts is shown in Figure 2.

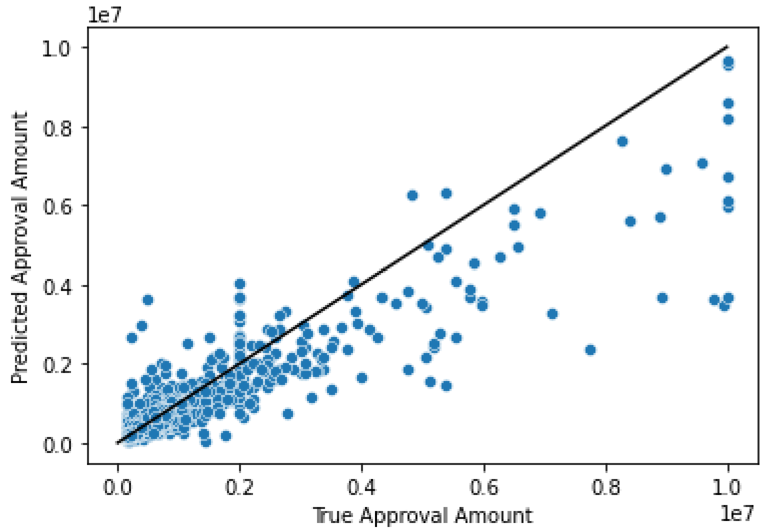


*Figure 2. True vs. Predicted Approval Amount for a Ridge regression model with 2nd degree polynomial effects and a λ = 10.*

## LASSO Regression Model with Polynomial Effects

A cross-validated grid search found that polynomial effects of degree = 2 and λ = 0.001 were the best parameters for a Ridge regression model. The $R^2$ score was 0.80 and the distribution of true vs. predicted approval amounts is shown in Figure 2.

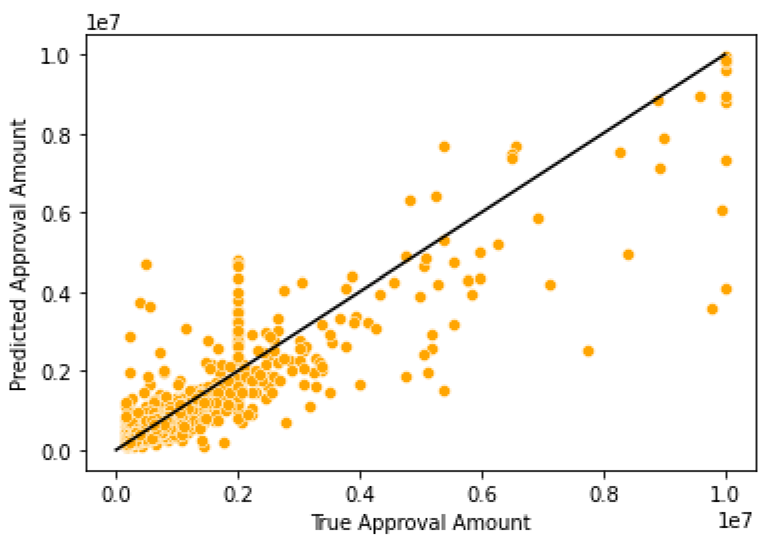


*Figure 3. True vs. Predicted Approval Amount for a LASSO regression model with 2nd degree polynomial effects and a λ = 0.001.*

## Model Recommendation

**The LASSO regression model is recommended for this analysis.** One of the benefits of the LASSO regression model is its ability to eliminate unnecessary coefficients, which increases interpretability of the data. In the case of these models, the Ridge regression model had 1,718 non-zero coefficients while the LASSO regression model had only 87 non-zero coefficients, which is a more manageable number to work with. In terms of accuracy, the difference between the Ridge and LASSO regression models is minimal, although the LASSO regression model did marginally outperform the Ridge regression model.

# Key Findings and Insights

- Of the original 91 parameters of input data into the model, 53 of those parameters were eliminated entirely from the model, leaving only 36 parameters:

    1. Term
    2. JobsReported
    3. ApprovalDay
    4. ApprovalMonth
    5. ApprovalYear
    6. ProcessingMethod_PPP
    7. BorrowerState_AR
    8. BorrowerState_CA
    9. BorrowerState_CO
    10. BorrowerState_FL
    11. BorrowerState_GA
    12. BorrowerState_LA
    13. BorrowerState_MO
    14. BorrowerState_NC
    15. BorrowerState_ND
    16. BorrowerState_NM
    17. BorrowerState_OK
    18. BorrowerState_PA
    19. BorrowerState_TX
    20. BorrowerState_UT
    21. RuralUrbanIndicator_R
    22. HubzoneIndicator_Y
    23. LMIIndicator_Y
    24. BusinessAgeDescription_Existing or more than 2 years old
    25. BusinessAgeDescription_New Business or 2 years or less
    26. ProjectState_OK
    27. ProjectState_TX
    28. Race_White
    29. Ethnicity_Hispanic or Latino
    30. Ethnicity_Not Hispanic or Latino
    31. BusinessType_Corporation
    32. BusinessType_Limited Liability Company(LLC)
    33. BusinessType_Partnership
    34. BusinessType_Sole Proprietorship
    35. BusinessType_Subchapter S Corporation
    36. Gender_Female Owned
    37. Gender_Male Owned
    38. Veteran_Non-Veteran

  These remaining terms show only 14 remaining states indicating that they either differed from the rest of the states or they hold most of the mining companies across US States and territories. Additionally, both ethnicity and gender parameters remain. Further insights from these parameters are shown later in the report.

- Looking at the top ten coefficients with the largest magnitudes, it becomes evident that the most important factor for the CurrentApprovalAmount for mining companies is the number of jobs reported by the company with the coefficients for JobsReported^2 = 3.9 and JobsReported = 0.6. It is expected that the number of jobs at a company would be the largest driving factor for the PPP loan approval amount. The small coefficient values of the remaining parameters may indicate minor variability on the CurrentApprovalAmount for mining companies based on those parameters.


*Figure 4. Coefficient values for the top ten coefficients with the largest magnitudes.*

- Looking at ethnicity, the results clearly show that being Hispanic or Latino negatively impacted the CurrentApprovalAmount. This effect is likely due to Hispanic or Latino business owners requesting smaller loan amounts, but further analysis is needed.

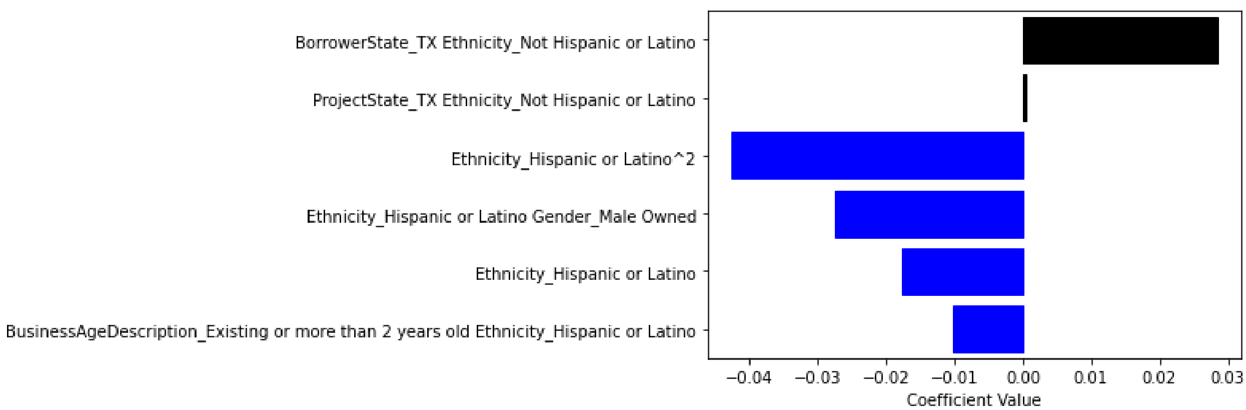


*Figure 5. Coefficient values for coefficients related to ethnicity.*

- Looking at how gender influenced the CurrentApprovalAmount, the parameters related only to Male Owned and Female Owned businesses were eliminated by the model. However, looking at the polynomial features related to gender reveals that features related Female Owned businesses were all negative while Male Owned businesses varied. This effect is likely due to some female business owners requesting smaller loan amounts, but further analysis is needed.

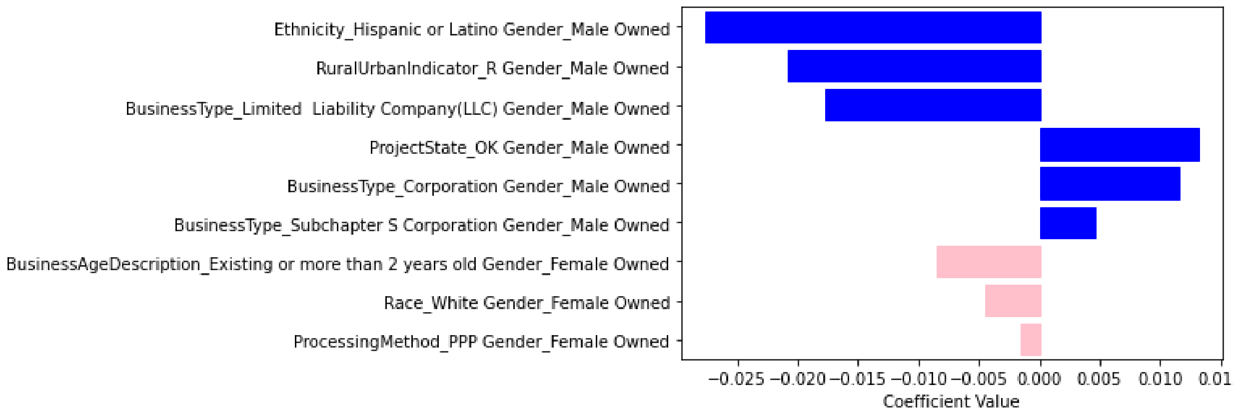


*Figure 6. Coefficient values for coefficients related to gender.*

- Large disparity between the predicted and true values for the CurrentApprovalAmount could indicate fraud when the true CurrentApprovalAmount is significantly larger than the predicted value. These data points could be flagged for auditing by an agent. Using machine learning to flag suspicious data can reduce the workload of an agent examining the data.

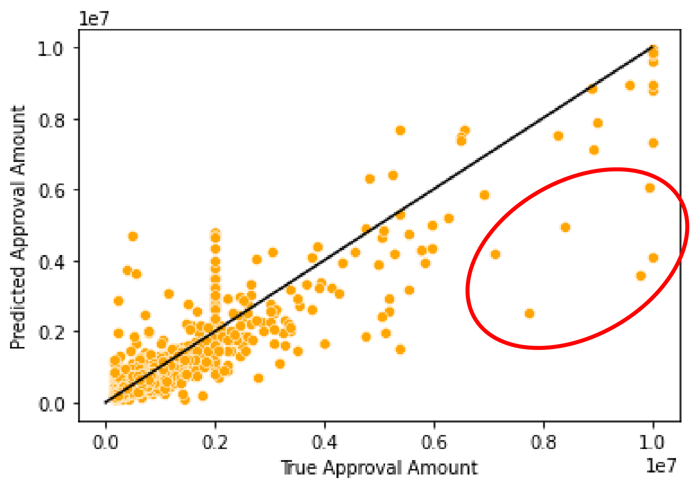


*Figure 6. True vs. Predicted Approval Amount for a LASSO regression model with 2nd degree polynomial effects and a λ = 0.001. The data in the red circle represents example data that could be flagged for agent auditing.*

## Next Steps

Next steps for this analysis include revisiting the model to clean up the polynomial features before they are fed into the model. Many of the columns were filled with zeros and ones, so there is no difference between BorrowerState_FL and BorrowerState_FL^2, but they were both retained by the model. Additionally, a second sweep of lambda values for the regularization models closer to the initially-selected lambda value may have yielded a higher $R^2$ score, improving the model performance.

To achieve better explanation of the results, it would be useful to have data on what each loan initially requested.

This analysis can also be repeated for other industries to see if their loan approval amounts vary in alternative ways.