Supervised Machine Learning: Classification Course Project: PPP Loan Data

Report by Jennifer Case
8/25/2022

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 all industries took advantage of the program. To qualify for loan forgiveness, borrowers were required to maintain employment and compensation levels, spend at least 60% of the loan on payroll costs, and spend any money not spent on payroll on other eligible expenses. Additionally, there were several other requirements for loan forgiveness including applying after all loan proceeds were spent, applying within a given time limit based on loan maturity, and following specific procedures based on lender participation in a direct forgiveness program[1].

In this report, the main objective is to create a classifier that filters the data based on if a loan was fully forgiven or not with emphasis on identifying unforgiven loans. This analysis serves as an attempted proof-of-concept for using machine learning to predict loan forgiveness as a way to prioritize workloads of workers who are processing loan forgiveness. The provided data is only a subset of the actual data that would be available when determining loan forgiveness.

Data Description

The PPP loan data as provided by the SBA is broken out into two categories: loans above $150k and loans equal to or below $150k. The subset of loans above $150k is considered in this analysis to make the dataset size more manageable. This analysis uses (1) geocoded PPP loan data from June 2021 provided by Geocodio[2] that contains additional columns for longitude, latitude, and 2010 Census data, (2) PPP loan data[3] from April 2022 that contains updated information about loan forgiveness, and (3) NAICS data[4] that enables linking the businesses to specific industries.

For the geocoded PPP loan data, there are 970,076 records for loans above $150k with the following 84 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

[1] "PPP loan forgiveness." *U.S. Small Business Administration*, https://www.sba.gov/funding-programs/loans/covid-19-relief-options/paycheck-protection-program/ppp-loan-forgiveness.
[2] "Geocoded PPP Loan Data." *Geocodio*, https://www.geocod.io/geocoded-ppp-loan-data/.
[3] "PPP FOIA: Data and Resources." *U.S. Small Business Administration*, 4 April 2022, https://data.sba.gov/dataset/ppp-foia.
[4] "NAICS & SIC Identification Tools." *NAICS Association*, https://www.naics.com/search/#naics.

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 - Loan Status Date is blank when the loan is disbursed but not Paid In Full or Charged Off	object
11	LoanStatus	Loan Status Description - 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	Non-Profit Indicator	object
52	ForgivenessAmount	Forgiveness Amount	float64
53	ForgivenessDate	Forgiveness Paid Date	object
54	Latitude	Latitude of borrower address	float64
55	Longitude	Longitude of borrower address	float64
56	Accuracy Score	Accuracy in matching borrower address to a known location[5]	float64
57	Accuracy Type	Description of how the borrower address was matched to map data[4]	object
58	Number	Number of the borrower address	object
59	Street	Street of the borrower address	object
60	Unit Type	Unit type of the borrower address (e.g., suite, etc)	object
61	Unit Number	Unit number of the borrower address	object
62	City	Same as BorrowerCity	object
63	State	Same as BorrowerState	object
64	County	County for the borrower address	object
65	Zip	5-digit zip code for the borrower address	float64
66	Country	Country for the borrower address	object
67	Source	Data source for location matching[6]	object
68	Census Year	Year of Census data	float64
69	State FIPS	Federal Information Processing Standard (FIPS) state code	float64
70	County FIPS	FIPS county code	float64
71	Place Name	Same as BorrowerCity	object
72	Place FIPS	FIPS city code	float64
73	Census Tract Code	Census tract code specifying a tract within a county	float64
74	Census Block Code	Census block code specifying a block within a tract	float64
75	Census Block Group	Census block group code specifying a group of blocks within a tract	float64
76	Full FIPS (block)	FIPS block code	float64

[5] "Accuracy Types & Scores." *Geocodio*, https://www.geocod.io/guides/accuracy-types-scores/.
[6] "Data Sources." *Geocodio*, https://www.geocod.io/data-sources/.

77	Full FIPS (tract)	FIPS tract code	float64
78	Metro/Micro Statistical Area Name	City and state for identified metropolitan and micropolitan areas	object
79	Metro/Micro Statistical Area Code	Area code for metro/micropolitan area	float64
80	Metro/Micro Statistical Area Type	Area type identifying if the address is in a metropolitan or micropolitan area	object
81	Combined Statistical Area Name	Combined area name for larger metropolitan areas	object
82	Combined Statistical Area Code	Area code for combined area	float64
83	Metropolitan Division Area Name	Metropolitan division area name	object
84	Metropolitan Division Area Code	Metropolitan division area code	float64

The April 2022 PPP loan data has 968,532 records for loans above $150k with the same first 53 attributes as the geocoded PPP loan data. The additional columns from the geocoded PPP loan data were joined to this data on the LoanNumber column.

The NAICS data set was used to convert the NAICSCode from the PPP loan data to a specific industry. This data set includes:

	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

Because this analysis is focused on loan forgiveness, only loans that have been paid-in-full (i.e., the loan has been paid back either through forgiveness, repayment, or a combination of the two) were considered. Of those loans (791,960 in total), only 5.6% were not fully forgiven (44,133 loans). The distribution of fully forgiven and not fully forgiven loans is shown in Figure 1. The data shows that the maximum loan approval amount for the loans that were not fully forgiven was less than those that were. This could indicate that borrowers with more financial means were better able to navigate the loan forgiveness process and, thus, were able to get their loans forgiven while others could not.

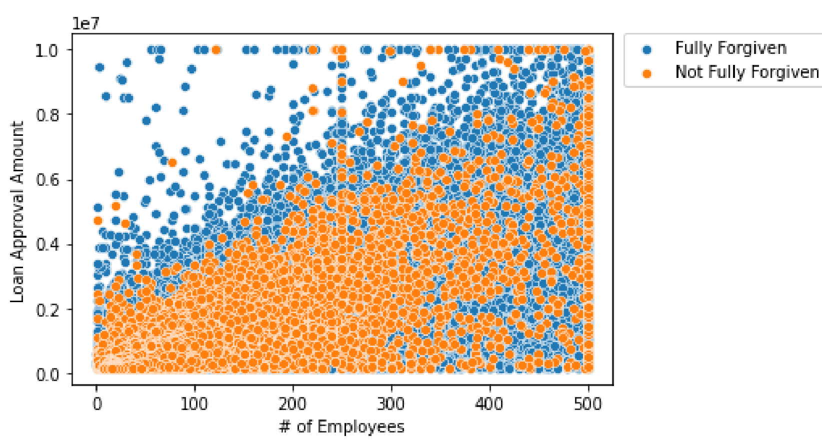

Figure 1. Number of employees versus the loan approval amount stratified by whether the loan was fully forgiven or not. The data for the loans that were not fully forgiven was plotted on top of the loans that were fully forgiven.

Of the loans that were not fully forgiven, the money owed ranged from a few cents to the entire loan amount up to $10 million, as shown in Figure 2.

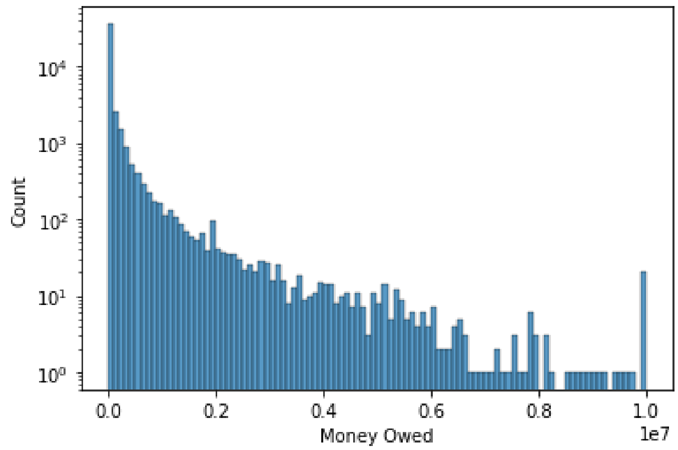

Figure 2. Histogram showing the distribution of how much money was owed on loans that were not fully forgiven. The y-axis is in log scale to show the distribution at the high end of money owed better.

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 were trimmed from the data set:

	Attribute Name	Reason for Removal
1	LoanNumber	Unique value not helpful for finding patterns.
2	DateApproved	Not factored when considering forgiveness.
3	SBAOfficeCode	Does not matter where SBA originally processed form.
5	BorrowerName	Unique value not helpful for finding patterns.
6	BorrowerAddress	Unique value not helpful for finding patterns.
7	BorrowerCity	Manual entry errors.
9	BorrowerZip	FIPS data is more usable.
8	BorrowerState	FIPS data is more usable.
10	LoanStatusDate	Not factored when considering forgiveness.
12	Term	Term is highly correlated (>90%) to ProcessingMethod and was trimmed to reduce model complexity.
13	SBAGuarantyPercentage	SBA guaranty percentage is 100% for all loans.
14	InitialApprovalAmount	InitialApprovalAmount is highly correlated (>90%) to CurrentApprovalAmount and was trimmed to reduce model complexity.
16	UndisbursedAmount	No undisbursed money for loans qualifying for forgiveness.
18	ServicingLenderLocationID	ServicingLenderName is more interpretable.
20	ServicingLenderAddress	ServicingLenderName is more interpretable.
21	ServicingLenderCity	ServicingLenderName is more interpretable.
22	ServicingLenderState	ServicingLenderName is more interpretable.

23	ServicingLenderZip	ServicingLenderName is more interpretable.
28	ProjectCity	Manual entry errors.
29	ProjectCountyName	Too individualized – could behave as unique identifier.
31	ProjectZip	Too individualized – could behave as unique identifier.
32	CD	Too individualized – could behave as unique identifier.
30	ProjectState	The correlation between BorrowerState and ProjectState is around 99%.
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 used because originating lender name is more interpretable.
46	OriginatingLender	Almost equivalent to ServicingLenderName.
47	OriginatingLenderCity	Not used because originating lender name is more interpretable.
48	OriginatingLenderState	Not used because originating lender name is more interpretable.
51	NonProfit	Data captured in BusinessType.
53	ForgivenessDate	Not factored when considering forgiveness.
57	Accuracy Type	Not used to clean the data and not relevant to forgiveness.
58	Number	Address is too unique for finding patterns.
59	Street	Address is too unique for finding patterns.
60	Unit Type	Address is too unique for finding patterns.
61	Unit Number	Address is too unique for finding patterns.
62	City	FIPS data is more usable.
63	State	FIPS data is more usable.
64	County	FIPS data is more usable.
65	Zip	FIPS data is more usable.
66	Country	The country is the same for every sample.
67	Source	Not factored when considering forgiveness.
68	Census Year	The census year is the same for every sample.
71	Place Name	FIPS data is more usable.
72	Place FIPS	Missing from significant portion of samples.
75	Census Block Group	Highly correlated to Census Block Code.
76	Full FIPS (block)	Data is captured by other columns.
77	Full FIPS (tract)	Data is captured by other columns.
78	Metro/Micro Statistical Area Name	Not every sample is part of a metro/micropolitan area.
79	Metro/Micro Statistical Area Code	Not every sample is part of a metro/micropolitan area.
81	Combined Statistical Area Name	Not every sample is part of a metro/micropolitan area.
82	Combined Statistical Area Code	Not every sample is part of a metro/micropolitan area.
83	Metropolitan Division Area Name	Not every sample is part of a metro/micropolitan area.
84	Metropolitan Division Area Code	Not every sample is part of a metro/micropolitan area.

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

	Attribute Name	Treatment of Null and N/A Values
4	ProcessingMethod	N/A
11	LoanStatus	N/A
15	CurrentApprovalAmount	N/A
17	FranchiseName	Null values left for later processing.
19	ServicingLenderName	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.
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
52	ForgivenessAmount	Null values changed to zero.
54	Latitude	N/A
55	Longitude	N/A
56	Accuracy Score	N/A
69	State FIPS	Rows removed.
70	County FIPS	Rows removed.
73	Census Tract Code	Rows removed.
74	Census Block Code	Rows removed.
80	Metro/Micro Statistical Area Type	Null values moved to a ‘None’ category.

- The data was trimmed to only include samples where Accuracy Score was above 0.8 resulting in 920,014 samples.
- The data was trimmed to only include samples where LoanStatus was “Paid in Full” resulting in 791,960 samples.
- The following columns were added:

	Attribute Name	**Attribute Descriptions**	**Data Type**
	Industry	Industry Title based on NAICSCode	object
	Franchise	1 – if there is an associated franchise; 0 – otherwise	int64
	numLoansSLender	The number of loans the servicing lender provided across the dataset	int64
	notFullyForgiven	1 – if (CurrentApprovalAmount – ForgivenessAmount) is greater than 0; 0 - otherwise	int64

Resulting in the removal of these existing columns: LoanStatus, FranchiseName, ServicingLenderName, NAICSCode, and ForegivenessAmount.

- One-hot encoding was used on the object columns for some of the classifiers resulting in a total of 83 columns:
    - ProcessingMethod became a single column where 1 indicated PPP and 0 indicated PPS.
    - RuralUrbanIndicator became a single column where 1 indicated Rural and 0 indicated Urban.
    - HubzoneIndicator became a single column where 1 indicated a HUBZone.
    - LMIIndicator became a single column where 1 indicated an LMI business.
    - BusinessAgeDescription became 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.
    - Race became 8 columns to indicate one of the following 9 options: American Indian or Alaska Native, Asian, Black or African American, Eskimo & Aleut, Multi Group, Native Hawaiian or Other Pacific Islander, Puerto Rican, White, or Unanswered.
    - Ethnicity became 2 columns to indicate one of the following 3 options: Hispanic or Latino, Not Hispanic or Latino, or Unknown/NotStated.
    - BusinessType became 25 columns to indicate one of the following 13 options: 501(c)19 – Non Profit Veterans, 501(c)3 – Non Profit, 501(c)6 – Non Profit Membership, 501(c) – Non Profit other, Cooperative, Corporation, Employee Stock Ownership Plan (ESOP), Housing Co-op, Independent Contractors, Joint Venture, Limited Liability Company (LLC), Limited Liability Partnership, Non-Profit Childcare Center, Non-Profit Organization, Partnership, Professional Association, Qualified Joint-Venture (spouses), Rollover as Business Start-Ups, Self-Employed Individuals, Single Member LLC, Sole Proprietorship, Subchapter S Corporation, Tenant in Common, Tribal Concerns, Trust, or Unanswered.
    - Gender became 2 columns to indicate one of the 3 options: Female Owned, Male Owned, or Unanswered.
    - Veteran became 2 columns to indicate one of the 3 options: Veteran, Non-Veteran, or Unanswered.
    - Metro/Micro Statistical Area Type becomes 2 columns to indicate one of the three options: metropolitan, micropolitan, or none.
    - Industry became 20 columns to indicate one of the 21 options: Accommodation and Food Services; Administrative and Support and Waste Management and Remediation Services; Agriculture, Forestry, Fishing and Hunting; Arts, Entertainment, and Recreation; Construction; Educational Services; Finance and Insurance; Health Care and Social Assistance; Information; Management of Companies and Enterprises; Manufacturing; Mining; Other Services (except Public Administration); Professional, Scientific, and Technical Services; Public Administration; Real Estate Rental and Leasing; Retail Trade; Transportation and Warehousing; Utilities; Wholesale Trade; or Unanswered.
- Label encoding was used for other classifiers, which did not add additional columns.
- Check for duplicate or highly correlated columns and remove. No duplicate columns were found and highly correlated columns have been previously noted in this report.
- The JobsReported column was transformed using the BoxCox transformation bringing the skew from 3.73 to 0.

- The CurrentApprovalAmount column was transformed using the BoxCox transformation bringing the skew from 5.64 to 0.21.
- The numLoansSLender (number of loans the servicing lender provided across dataset) column was transformed using the log transformation bringing the skew from 2.04 to -0.34.

Classification Models

Four classifiers were tested for this dataset: (1) logistic regression, (2) linear support vector machine, (3) decision tree, and (4) random forest. For the logistic regression and support vector machine, the data was scaled using a min-max scaler and one-hot encoding was used. This scaling was not used for the decision tree and random forest classifiers because it is not necessary and label encoding was used instead of one-hot encoding. A stratified sample of 30% of the data (237,588 samples) was pulled to test the various classifiers since cross-validation on the full data set was time consuming (e.g., lasting over several hours depending on the classifier).

For each classifier, the data was split into 70% training data and 30% testing data using stratification to ensure distribution of loan forgiveness remained. A stratified k-fold with a split of 4 was used for cross-validation to tune hyperparameters for all classifiers. The classifiers were scored using the f1 score since the data is heavily imbalanced (5% not forgiven vs 95% forgiven). Scoring using the Area-Under-the-Curve (AUC) score was also tested but did not perform as well.

Logistic Regression Classifier

For the logistic regression classifier, the liblinear solver was used. The tuned hyperparameters were class weight ({0: 0.1, 1: 0.9}, {0: 0.2, 1: 0.8}, {0: 0.3, 1: 0.7}), penalty (l1, l2), and C (1, 10). The best classifier was found to have a class weight of {0: 0.1, 1: 0.9}, penalty of l2, and C of 10.

Although the accuracy was 91%, the F1-score was 11% and the AUC was 53% or barely better than a coin flip. The confusion matrix for this classifier is shown in Figure 3. The recall was 11%, which means only 11% of the loans that were not fully forgiven were correctly labeled, and the precision was 12%, which means a lot of forgiven loans were identified as not forgiven.

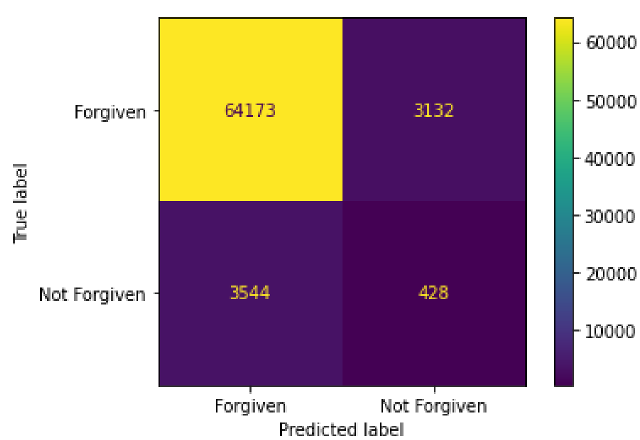

Figure 3. Confusion matrix for logistic regression classifier.

Support Vector Machine Classifier

For the linear support vector machine classifier, the tuned hyperparameters were class weight ({0: 0.05, 1: 0.95}, {0: 0.1, 1: 0.9}, {0: 0.15, 1: 0.85}, {0: 0.2, 1: 0.8}) and C (0.1, 1, 2). The best classifier was found to have a class weight of {0: 0.05, 1: 0.95} and C of 2.

While the accuracy of this classifier was only 35%, the F1-score was 64%, the AUC was 62%, the recall was 93%, and the precision was 7%. The confusion matrix for this classifier is shown in Figure 4.

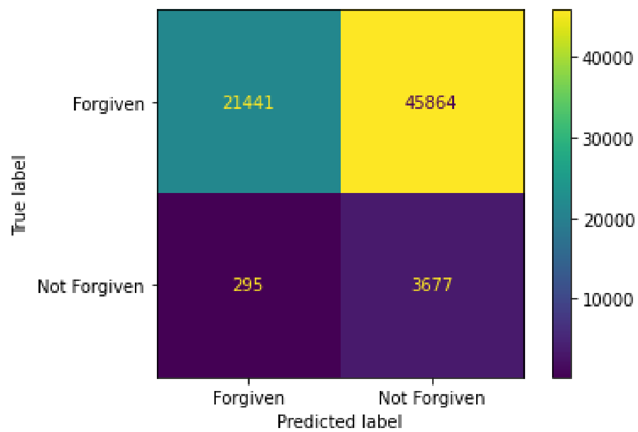

Figure 4. Confusion matrix for support vector machine classifier.

Decision Tree Classifier

For the decision tree classifier, the tuned hyperparameters were class weight ({0: 0.05, 1: 0.95}, {0: 0.1, 1: 0.9}, {0: 0.15, 1: 0.85}), max depth (10, 50), and the number of minimum samples split (2, 5, 10). The best classifier was found to have a class weight of {0: 0.05, 1: 0.95}, max depth of 10, and number of minimum samples split of 5.

While the accuracy of this classifier was only 43%, the F1-score was 61%, the AUC was 63%, the recall was 84%, and the precision was 8%. The confusion matrix for this classifier is shown in Figure 5.

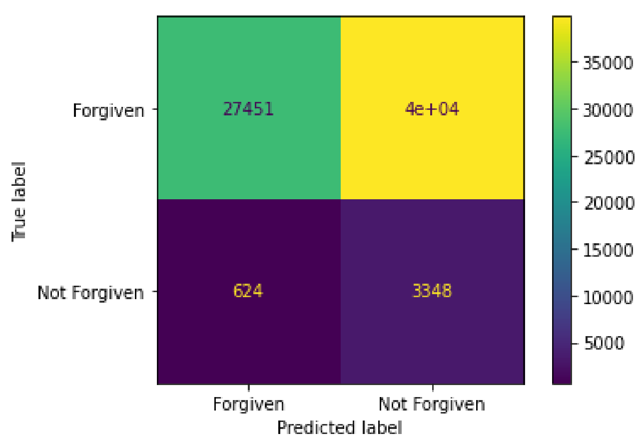

Figure 5. Confusion matrix for decision tree classifier.

Random Forest Classifier

For the random forest classifier, the tuned hyperparameters were class weight ({0: 0.05, 1: 0.95}, {0: 0.1, 1: 0.9}, {0: 0.15, 1:0.85}), max depth (10, 50), number of estimators (25, 50, 100), or the number of minimum samples split (2, 5). The best classifier was found to have a class weight of {0: 0.05, 1: 0.95}, max depth of 10, number of estimators of 25, and the number of minimum samples split of 2.

While the accuracy of this classifier was only 45%, the F1-score was 60%, the AUC was 62%, the recall was 82%, and the precision was 8%. The confusion matrix for this classifier is shown in Figure 6.

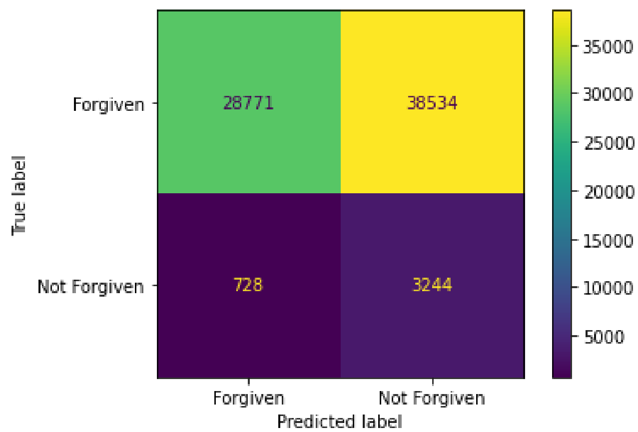

Figure 6. Confusion matrix for random forest classifier.

Model Recommendation

The accuracy and precision are low on the classifiers that have high recall. However, for establishing priority of processing loan forgiveness for workers, it may be acceptable to use a low accuracy and precision classifier since the recall is high and a considerable percentage of loans would be flagged as highly likely to be forgiven allowing them to focus on those that are less likely to be forgiven.

The random forest classifier is recommended due to its high recall score, quick training time, interpretability, and resistance to overfitting.

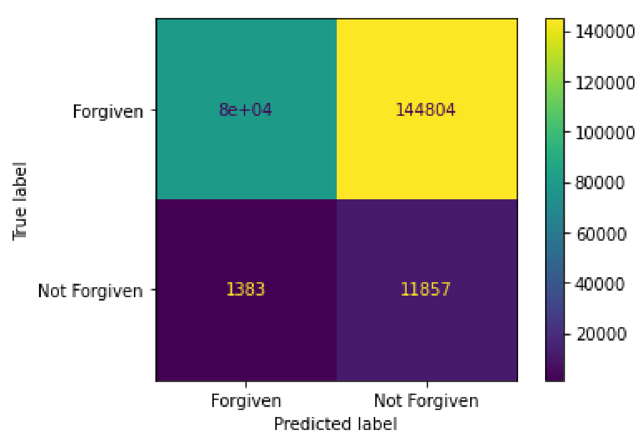

Figure 7. Confusion matrix for random forest classifier with training/testing split across all data.

The random forest classifier was re-trained across all the data set using 70/30 training/testing split using the hyperparameters found with the cross-validation testing on the sample data set. The confusion matrix for this classifier is shown in Figure 7. The accuracy was 38%, the F1-score was 63%, the AUC was 63%, the recall was 90%, and the precision was 8%.

Key Findings and Insights

To determine feature importance, an analysis using feature permutation was performed, as shown in Figure 8. This analysis showed that the processing method was the most important feature in determining if a loan was fully forgiven or not. Other important features were: BusinessType, JobsReported, CurrentApprovalAmount, Census Block Group, Veteran, Census Block Code, numLoansSLender, Census Tract Code, Metro/Micro Statistical Area Type, and RuralUrban Indicator. The business type may indicate that certain businesses may have had advantages in terms of obtaining full loan forgiveness. The Census Block Group, Block Code, and Tract Code all specify locations, which may indicate that the location of the business may influence loan forgiveness.

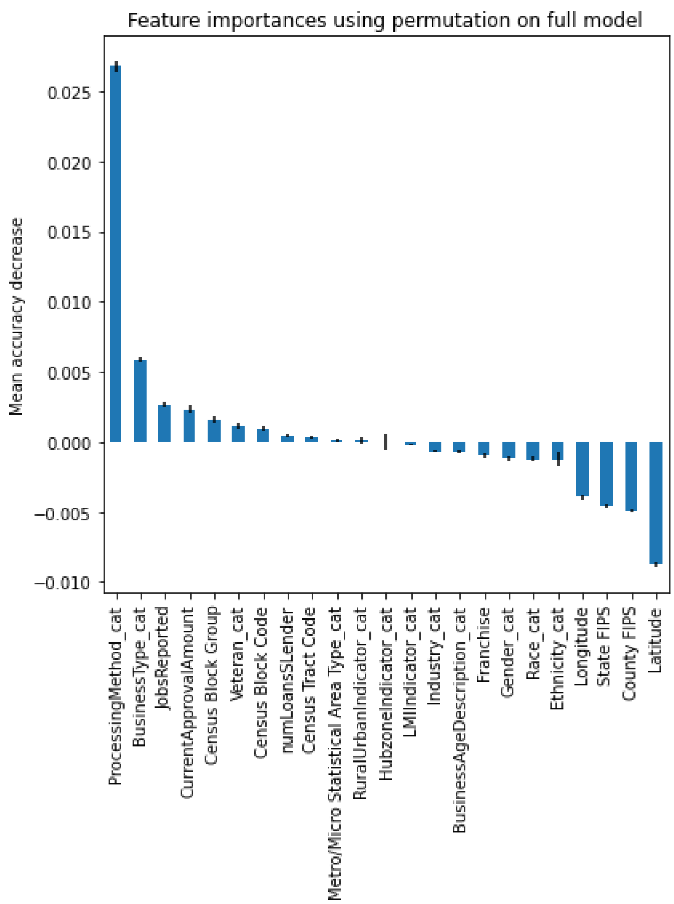

Figure 8. Feature importance measured using feature permutation.

Figure 8 also showed that the following features were found to be unimportant: HubzoneIndicator, LMIIndicator, Industry, BusinessAgeDescription, Franchise, Gender, Race, Ethnicity, Longitude, State FIPS, County FIPS, and Latitude. Note here that although latitude and longitude indicate location, this location identification was found to be less valuable than using the Census data that indicates location.

Figure 9 shows the true and predicted forgiveness of loans stratified by the processing method. The random forest classifier predicted all PPS loans were fully forgiven, which accounts for some of the model error. It also overestimated the number of PPP loans that were not fully forgiven.

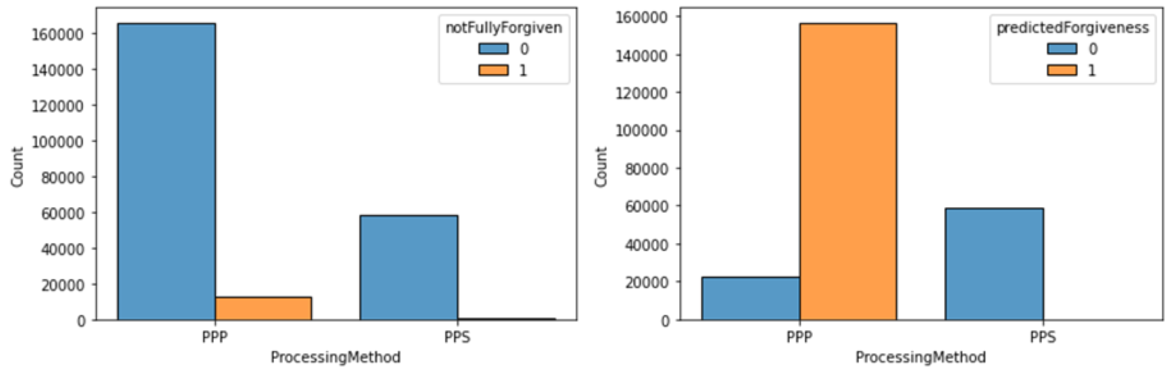

Figure 9. Number of loans in the testing data stratified by the Processing Method and (Left) true forgiveness and (Right) predicted forgiveness results where 0 indicates the loan was fully forgiven and 1 indicates the loan was not fully forgiven.

Next Steps

While this analysis shows some potential in predicting loan forgiveness, the dataset used to predict forgiveness is incomplete. To improve performance, it would be helpful to have additional features that were collected on the loan forgiveness forms[7] including Employees at Time of Loan Application and Employees at Time of Forgiveness Application. This analysis could also be re-run using the lender-reported proceeds although there are potential entry errors in these columns given manual entry errors in other columns like BorrowerCity.

[7] "PPP Loan Forgiveness Application + Instructions." *U.S. Small Business Administration*, https://www.sba.gov/document/sba-form-3508-ppp-loan-forgiveness-application-instructions.