Paycheck Protection Program (PPP):
Loan Decision Analysis Led by Data

Emily Atkinson, Adam Brewer, Stephanie Leiva, Nolan Thomas
Dev10 Capstone: October 7th, 2022

I. INTRODUCTION

As the global pandemic COVID-19 impacted the U.S. Economy in 2019, the United States government implemented an unprecedented and large-scale solution to keep businesses (and by extension, their employees) afloat. The Paycheck Protection Program (PPP), which ran from March 2020 through April 2021, allowed businesses to take out loans backed by the federal government. The government intended for these loans to keep businesses solvent and employees on their payroll during the span of the pandemic. Throughout this report, we will provide information to answer the following questions:

- What types of business owners (gender, race, geography) received Paycheck Protection Program loans?
- Which industries (based on Census NAICS codes) received the largest amount of loans?
- How did businesses in each state and industry use the loans they received?
- Was the quantity of loans given proportional to the demographic breakdown of business in each state/industry?
- Which businesses in each state/industry paid off their loans? Which businesses had their loans forgiven?
- Who were the top lenders to businesses in each state and industry?
- What is the relationship between the number of PPP loans given by state and that state's unemployment numbers?
- Based on predictive modeling analysis, will a particular business pay back their loan?

II. DATASETS & ETL

We used three independent datasets to form the basis of our research. The U.S. Small Business Administration provided details on the PPP program including lenders, borrows, & loan information across 13 CSV files. From the Census, we utilized the 2020 Company Summary portion of the Annual Business Survey, which provides information on employer businesses broken down into multiple demographics. We accessed this information via an API call. Finally, the U.S. Department of Labor Employment & Training Information provides weekly Unemployment numbers at the state-level. We used this as a benchmark to view how PPP loans impacted different regions of the nation's economy. We retrieved unemployment information for the span of the PPP as an XLSX spreadsheet and then converted the data to CSV format.

After retrieving all relevant data for our research, we then began the process of transformation to populate our SQL database. Full details regarding the process of cleaning the data and populating our database are located in the RepeatableETLReport.pdf file of this GitHub repository. We utilized Python and Apache Spark to transform the data. General transformation practices included removing null values, dropping unnecessary columns, removing aggregations already present so as to not skew our results, converting data into an appropriate type, and renaming columns for clarity. For the PPP datasets, we combined the 13 CSV files into a single CSV. With transformation of the data completed, we then deployed all the data into our SQL Database through the use of Spark Databricks. We utilized a Kafka producer and consumer to stream out and in the Unemployment dataset and then deployed into our SQL Database. The Kafka producer and consumer gives us the opportunity to have Unemployment data updated independently of a user doing so manually. *Figure 1* shows a complete view of our data platform, including the cleaning and transformation process.

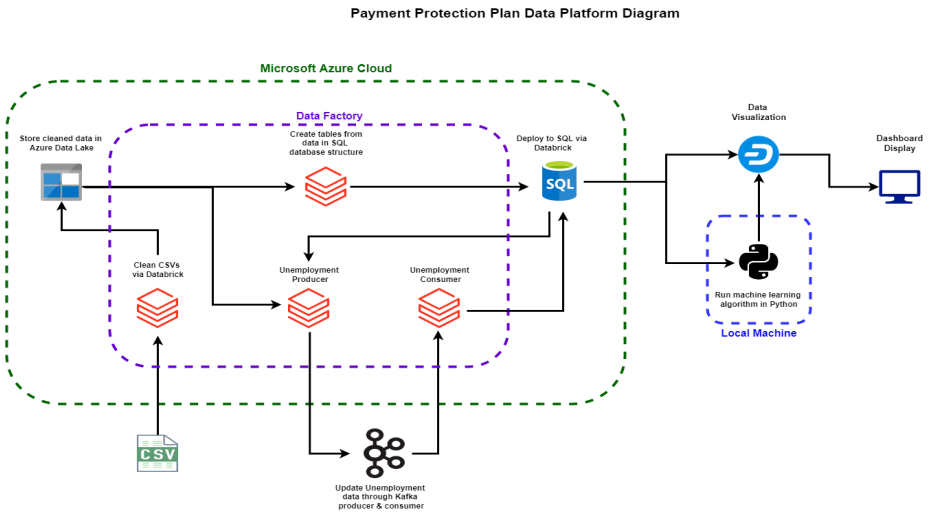

Figure 1: Payment Protection Program Data Platform Diagram

III. DATA ANALYSIS AND VISUALIZATION

For our dashboard, we utilized Plotly's Dash in combination with Bootstrapping to create a user-friendly webpage. Dropdown menus, interactive graphs, and pre-run visuals then populate using Dash's Python library. We created the visualizations in the dashboard using both Plotly Express and Plotly Graph Objects. In order to dynamically update the graphs within the dashboard, we utilize the user's dropdown input to call relevant information from the pandas DataFrames, which we then transform into Plotly visualizations.

In order to delve into our Exploratory Questions, we first looked at the general demographic information of the businesses/business owners that received PPP loans.

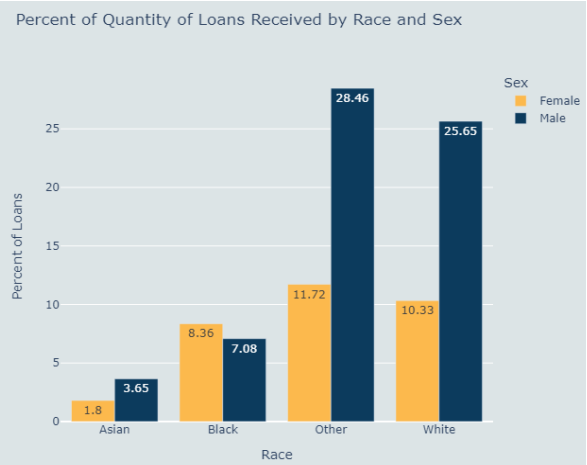

Figure 2: Percent of Quantity of Loans Recieved by Race and Sex

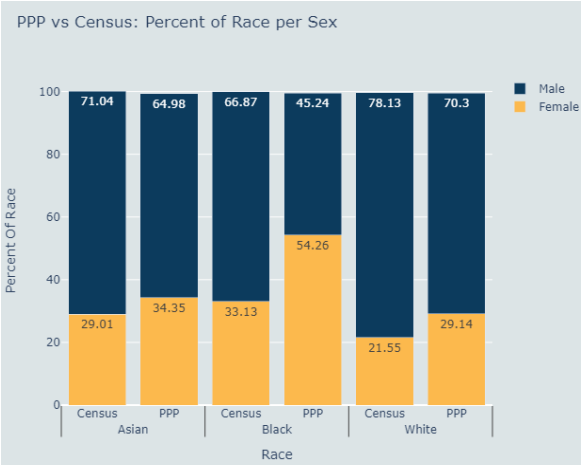

Figure 3: A graph showing the percentage of business owners broken down by race, sex, and data source (Census and PPP).

Per data from the US Census, shown in *Figure 3*, the overall percentage of female-owned businesses was lower than the overall percentage of male-owned businesses across all races depicted. However, the percentage of female-owned businesses that received PPP loans was relatively higher than the proportion of female-owned businesses overall. For example, while 33% of Black business owners are female, 54% of Black PPP loan recipients are female. We see this same trend reflected in the quantity of loans received, with businesses owned by Black women receiving a higher share of the total loans than businesses owned by Black men.

Next, we looked at the Top Industries that received funding through the PPP Loans. Within our interactive dashboard, this graph changes to reflect the Top 5 industries depending on the state that the user selects.

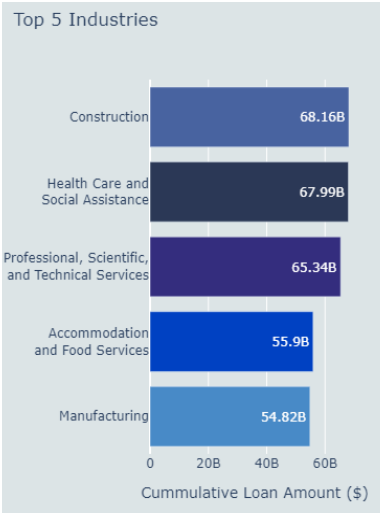

Figure 4: The top 5 industries by cumulative loan amount

Figure 4 shows the top industries that received funding at the national level. Construction, Health Care, and Science/Tech Industries received the largest amounts of funding from the PPP, with Food Services and Manufacturing trailing slightly behind. In order to view the top industries within a state, you can utilize the dropdown menus on our dashboard.

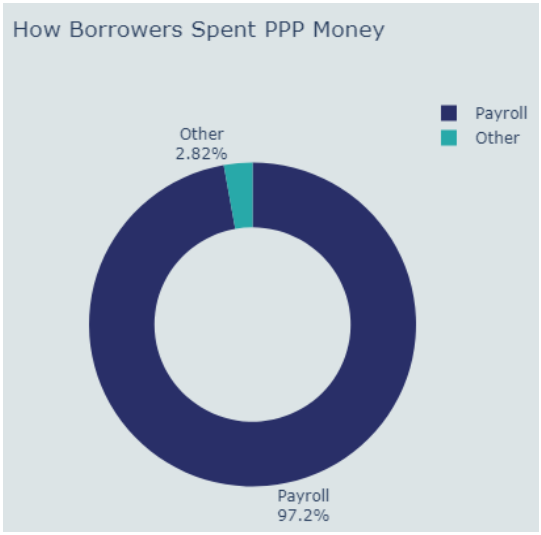

Figure 5: A breakdown of how borrowers spent PPP money, grouped by payroll and other.

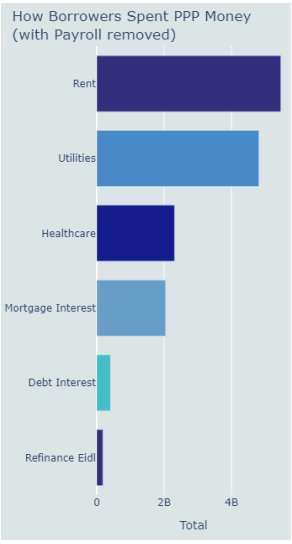

Figure 6: A breakdown of the Other category shown in Figure 5.

In order to understand how borrowers spent their funds, we looked into the different proceed categories contained with the PPP loan data. Staying consistent with the name of the program, borrowers spent 97% of the total funds on business payroll, with only 2.82% of total funds spent for other purposes. We chose to group the other categories because while the total spent for each category is still significant, and *Figure 5* belied the magnitude of each category outside of payroll.

In order to get a complete understanding of the 'Other' category, we created a bar graph. *Figure 6* provides a breakdown of the 'Other' slice within *Figure 5*. Outside of payroll, borrowers allocated the most on rent and utilities, spending over $4 billion nationwide in each category. Healthcare & Mortgage Interest trail behind at around $2 billion each.

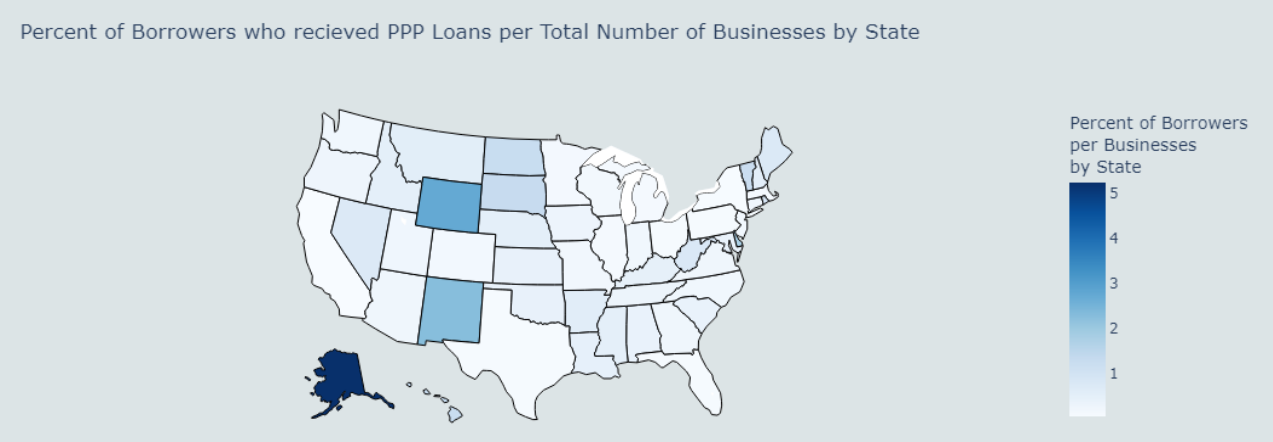

Figure 7: A map of the United States showing the Percent of Borrowers who received PPP loans per total number of businesses by state.

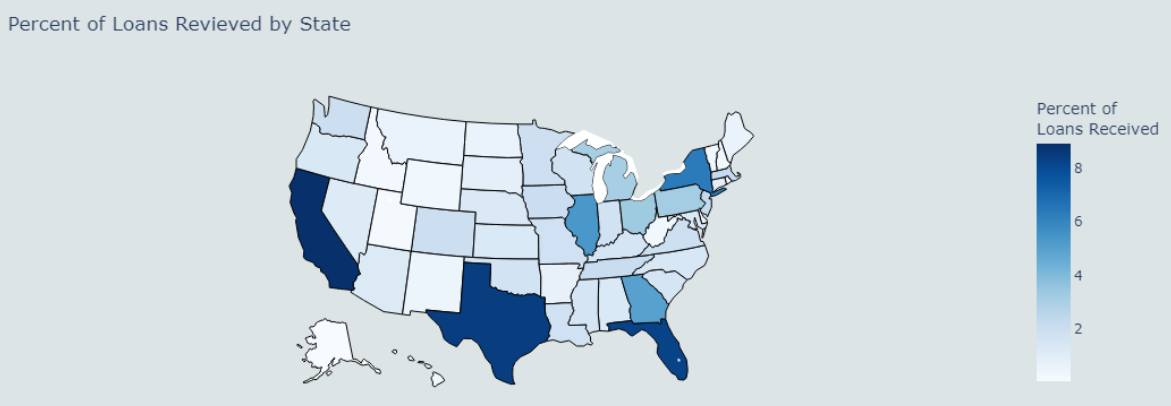

Figure 8: A map of the United States showing the Percent of PPP loans by state.

Next, we wanted to see which states received the greatest number of loans in comparison to the number of total businesses within that respective state. *Figure 7* displays the number of borrowers receiving loans divided by the total number of businesses in each state and converted into a percentage. We thought it was important to break this information down in a variety of ways to understand loans with and without the context of state population. For example, Alaska tops the nation for loans per business owner, with 5.23% of business owners receiving PPP loans. As shown in *Figure 8,* however, Alaska only accounts for 0.03% of the total national loans.

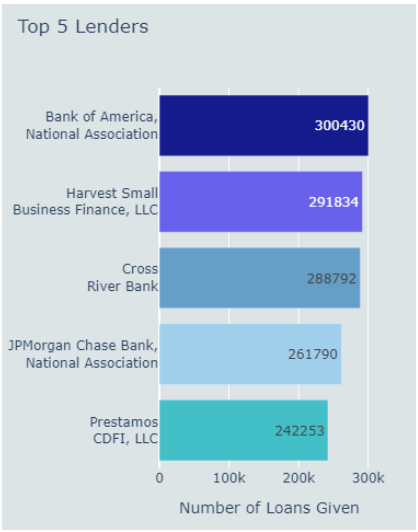

Figure 9: The top 5 PPP lenders by number of loans given

Next, we analyzed the Top 5 lenders for each respective state & industry combination through interactive dropdowns in our dashboard. For the purposes of this report, *Figure 9* represents the Top 5 lenders across all industries and at the national level. We focused on the total number of loans given by these lenders rather than the total dollar amount. The largest lenders at the national level are Bank of America, Harvest Small Business Finance, Cross River Bank, JPMorgan Chase Bank, and Prestamos CDFI.

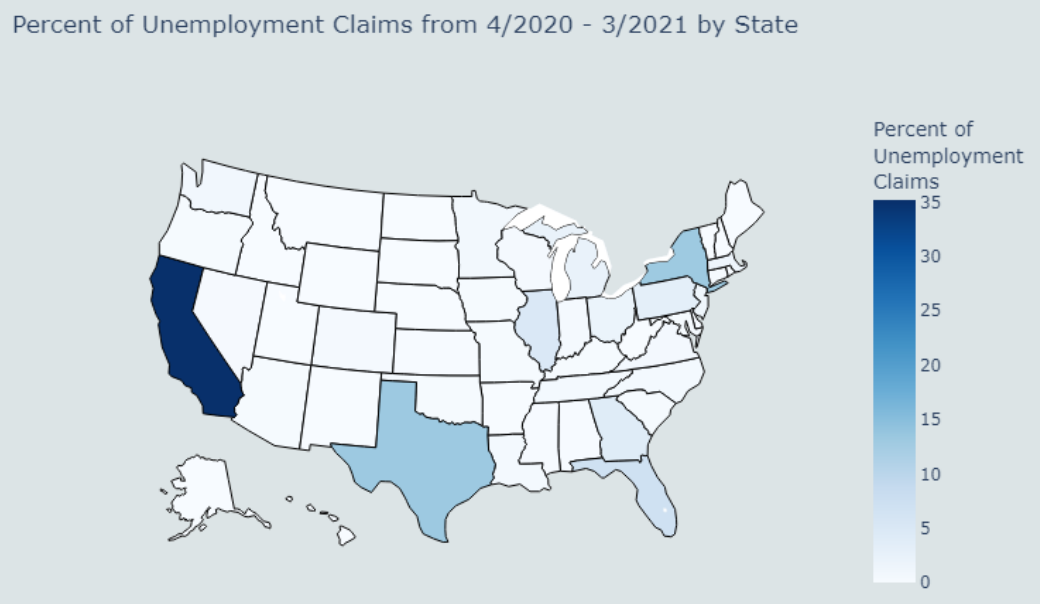

Figure 10: Percent of total unemployment claims for the duration of the PPP by state

Figure 11: A scatterplot showing the total unemployment claims on the X axis and the total number of PPP loans on the Y axis.

Lastly, we visualized the relationship between PPP loans and unemployment claims throughout the duration of the PPP. *Figure 10* represents the percentage of unemployment claims made by each state during active months of the PPP. *Figure 10* generally follows the same trends as *Figure 8,* with states with a higer population filing the largest volume of unemployment claims and applying for the largest number of PPP loans. *Figure 11* further delves into the relationship between PPP loans and unemployment claims. Our dashboard also features hover data for each point, describing the state, region, total number of unemployment claims, and total number of PPP claims.

Our goal with these visualizations was to put large-scale data into a bite-sized format, giving users the opportunity to interact with each graph through dropdowns and hover text. We hope that users leave the dashboard with an understanding of what the PPP was and what it looked like in practice. Below, we detail our machine learning process, which seeks to provide data to improve a future implementation of a program like the PPP.

IV. MACHINE LEARNING

Throughout our analysis of the PPP data, we anchored our focus on what happened during the loan program. Who received the loans, how did they use them, was there a connection to unemployment rates? For machine learning, we shifted from focusing on what the loan program was like in practice to how it can be optimized for future implementations of a similar program.

Throughout the course of the Paycheck Protection Program, the US government forgave $742 billion of the $793 billion of loans disbursed. Given the large scale of the PPP data (over 11 million total loans), this still

left a meaningful number of unforgiven loans. For our predictive modeling, we created a model that predicts whether or not a borrower will pay back their loan based on key information including the industry, age of the business, and the demographics of the business owner.

We took several steps to clean our data for the model. First, we imported target rows from our SQL database. These rows were selected based on analysis of the original PPP dataset. When importing those rows, we filtered out where business owners did not input Race, Ethnicity, Sex, Business Age Description, and/or State Name.

Once we imported our data, we created a column, *difference*, that captured the difference between the amount of money loaned to the borrower and the amount of money forgiven. We filtered our data frame so that only loans with an outstanding balance remained. After this, we created dummy variables for all columns with categorical data. After filtering out unanswered information and loans that the government forgave in their entirety, we ended up with a dataframe with 97 columns and 48,597 rows.

Our target row for this model was 'Loan Status.' According to the PPP data dictionary, a loan is either labeled as 'Paid in Full' or 'Exemption 4.' The Small Business Administration defines Exemption 4 as "when the loan is disbursed but not Paid in Full or Charged Off" (Small Business Administration 2022). Therefore, of the loans with an outstanding balance, the loan status tells us whether or not a business has fully paid their loan.

Given that K-Nearest Neighbors (the model we used) classifies a point based on its proximity to other points, we also standardized our data by running the standard scaler feature on our train and test data sets. This impacts the data by "removing the mean and scaling to unit variance" (Sklearn.preprocessing.StandardScaler).

After setting up the test and train sets, we ran ANOVA tests on all features to determine the variance. While the results are slightly harder to interpret as most features are represented as dummy variables, we used this to understand broader categories and their importance. For example, StateName_NewYork is a factor with a relatively higher ANOVA score, so that provided us with a rationale for keeping all state variables in the model. However, any variable tied to BusinessAgeDescription has a relatively low score, which led us to determine that we should remove it from the model. Based on the results, we dropped loan number and business age description from the model, and then recreated train and test sets with the new data frame.

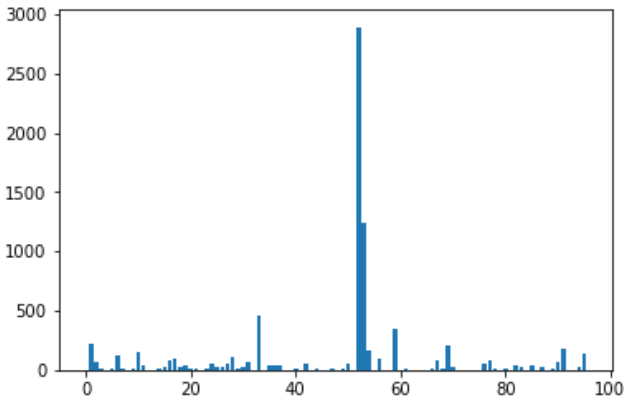

Figure 12: Results of the feature importance test. We used a printout of these results to determine the most important variables.

From there, we ran the base K-Nearest Neighbors (KNN) model before starting to tune the hyperparameters. We selected KNN because, based on preliminary tests with a subset of our data, it had the best test and train accuracy. K-Nearest Neighbors can be used for both classification and regression analysis, although the former is more common. In classification analysis, the algorithm determines a class for a target data point based on the class of the *k* nearest neighbors. *k* is a hyperparameter selected by the user. The other main hyperparameter is the formula the algorithm uses to measure the distance between points.

The base KNN model had a decent performance, with a train accuracy of 0.9 and a test accuracy of 0.87. Based on the confusion matrix, the model will predict a false positive (that someone will pay back their loan when they probably won't) about 9.6% of the time, and will predict a false negative (that someone won't pay back their loan when they probably will) about 3% of the time. Given that a lender would use this model to reduce their financial risk, we tried to decrease the false positive rate with tuning.

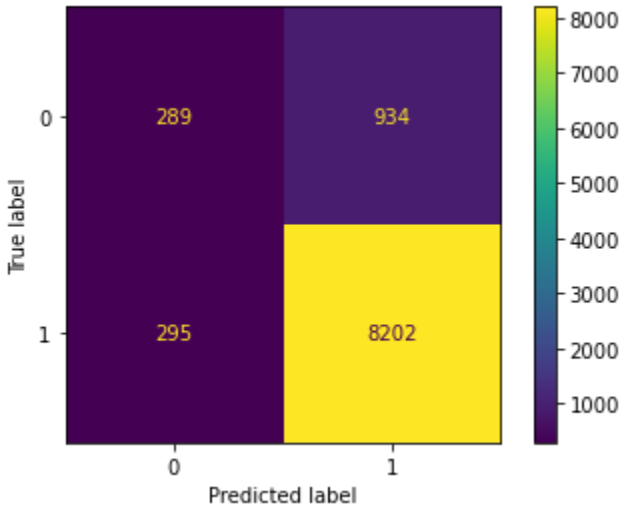

Figure 13: Confusion matrix with results from default KNN parameters.

The first tuning metric we used is a grid search. Here, SKLearn runs the model with a variety of parameters to determine the optimal parameters for a specific model. We ran the grid search with a list of neighbors between 50 and 500, increasing in increments of 50. We also tested a variety of distance equations to see which one got us the best performance. According to the grid search, the optimal parameters for our model (within the specifications provided) were k = 50 and the city block equation. When we fit the model, we saw a slight increase in accuracy (test accuracy of 0.88 and a train accuracy of 0.89) but a large increase in false positives (up to 11% of results) compared to the base model.

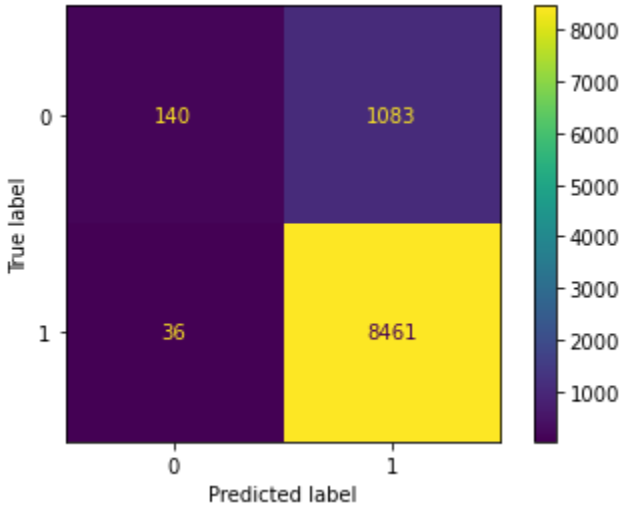

Figure 14: Confusion matrix with K = 50 and metric = 'cityblock'

Given that the grid search didn't improve the performance of the model for our purposes, we tried another method to find the optimal number for *k*. Based on time/processing restrictions, the grid search ran in iterations of 50. This search method looking only at *k* allowed us to search in increments of 1. Since the grid search chose *k* = 50, the lowest number available, we looked at *k* values from 1-50 to find the optimal number for *k*. This method runs the KNN algorithm with a list of numbers that we provide and then prints the accuracy score.

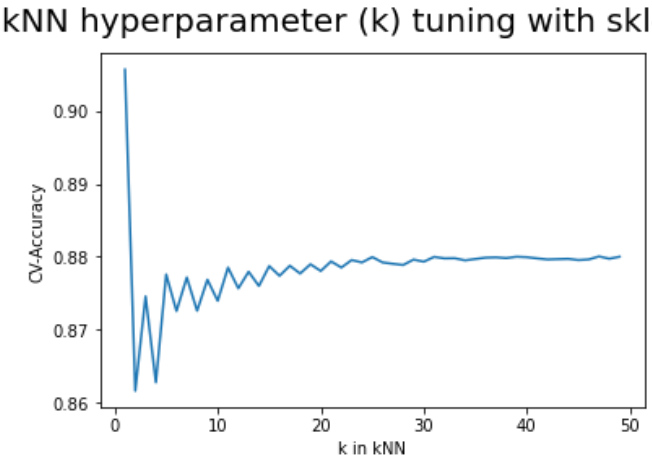

Fig 15: A printout of k parameter hypertuning.

Based on the results of the *k* accuracy test, we ran the model with $k = 1$. There are downsides to using $k = 1$, specifically that the model may overfit the data, or that one single point might be too noisy. After running the model, we determined it could potentially be overfit, with a train accuracy of 1. However, the test accuracy is lower at 0.89, showing that the model isn't necessarily overfitting the results. Looking at the confusion matrix, we reduced our rate false positives to 5.9%, which is ideal for the purposes of a lending model.

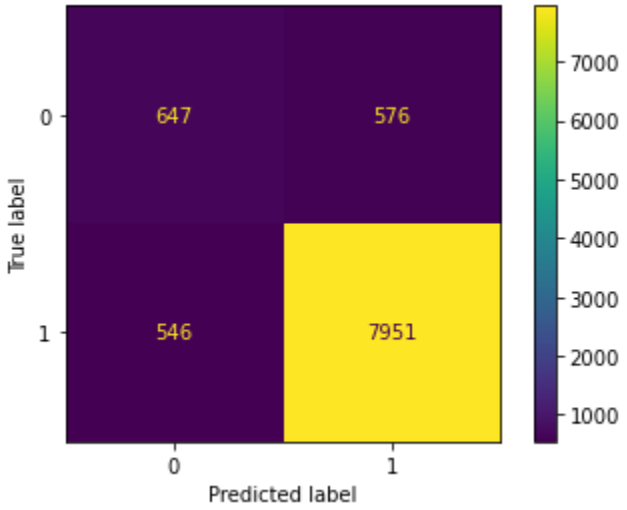

Fig 16: Confusion matrix with K = 1 and metric = 'cityblock'

With machine learning models there are always trade-offs. Our model is less accurate at predicting negatives, whether true or false. However, this is the most risk-averse model, meaning it is the most likely to be favored by lenders and financial institutions. Should the United States ever run such a program again, the implementation of our model would allow them to select borrowers most likely to pay back their loans.

V. CONCLUSIONS

The capstone project succeeds with our goal of analyzing our initial questions. Through visualizations, we explored the demographic breakdown of borrowers and census demographic breakdown of business owners. We also explored top industries that received PPP, top lenders that gave out PPP, and how businesses used their PPP. Bringing in unemployment information, we analyzed the relationship between the number of PPP loans and the number of unemployment claims for each state. Should there ever be a need to reimplement the PPP, our machine learning model serves as an effective predictive tool for whether or not a borrower will pay back their loan.

Given the emergent nature of the pandemic, the US Government implemented the PPP with relatively few regulations and little oversight. While that means that there are many things the government can improve for

future implementation, there are some potentially unintentional aspects that future loan programs can learn from. Based on our findings, we recommend the following areas for further research:

1. As seen in *Figure 3*, a relatively higher proportion of female-owned businesses received PPP loans than are represented in the Census. How and why did this happen? How can other loan programs learn from the PPP to succeed in supporting more minority groups?
2. Would it be worth demonstrating specific employee payroll requirements during the loan application process to ensure that funds are being properly allocated to keep employees on staff? As this program was initially earmarked for Payroll Protection, we see in *Figure 6* that billions of dollars of funding (even while a relatively small percentage of the total disbursed) went to items such as mortgage and debt interest payments.
3. While the PPP was implemented in a time of crisis, policy makers could use successful aspects of the PPP to prevent high unemployment rates in different circumstances. How can the relative success of the PPP assist policy makers in mitigating unemployment in the future? Specifically, it would be worth investigating the tradeoff between the cost of supporting unemployed people versus the cost of forgiving PPP loans.
4. Our machine learning model aims to predict which businesses would have their PPP loan forgiven based on key business information. Should there ever be a need to run a program such as PPP again, should businesses that are predicted to have their loans forgiven be given a grant instead? They are predicted to already have their loan forgiven; a grant would be in good faith and could ease the worries of a business owner.

VI. RESOURCES

Small Business Administration. (July 4, 2022). Paycheck Protection Program - Freedom of Information Act. Retrieved September 19, 2022 from data.sba.gov website: https://data.sba.gov/dataset/ppp-foia

United States Census Bureau. (October 28, 2021). Annual Business Survey. Retrieved September 20, 2022 from www.census.gov website: https://www.census.gov/data/developers/data-sets/abs.html

United States Department of Labor. (July 7, 2022). Unemployment Insurance Weekly Claims Data. Retrieved September 20, 2022 from oui.doleta.gov website: https://oui.doleta.gov/unemploy/claims.asp

scikit-learn developers. (n.d.). *Sklearn.preprocessing.StandardScaler*. scikit-learn.org. Retrieved October 4, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html#sklearn.preprocessing.StandardScaler