Exploring County Level access to Credit in the United States of America

Peter Koumpikov and James Burnard

This project was created by Peter Koumpikov and James Burnard. Each of us contributed evenly and fairly. Peter focused on data preprocessing in pandas, statistical summarization, and visualization of the data using matplotlib and seaborn. James focused on the association analysis section as well as the predictive modeling section. We both worked together to create this writeup.

The goal of our exploratory data analysis is to explore trends in credit scores, credit balances, and debt delinquency, and to see if we can uncover relationships between these and other features in the data set. As Informatics and Economics students, we are interested in this information and analysis as it is important in uncovering shortcomings and informing policy and regulation that can help people. To create the dataset we analyzed and worked with, we averaged features across three separate datasets and then combined the datasets. All the datasets came from *opportunity insights.*

The first dataset contains 6 columns. The first is par_state (which stands for the state from which the parent is), the second is par_county (which stands for the county the parent is from), the third is kid_race (which stands for the race of the individual being examined) the fourth is par_pctile (which stands for the income percentile of the parent), the fifth is shrunk_xkidvscore2020 (which stands for the vantage credit score of the individual being observed in 2020) and the last is county_name (which is the name of the county the individual lives).

The second dataset also contains six columns. It has the same columns as the first dataset - however, instead of shrunk_xkidvscore2020, it has shrunk_xkid_brcbalance2020 (which is the credit card balance of the individual in 2020).

Similarly, the third dataset contains the same columns as the first dataset but replaces shrunk_xkidvscore2020 with shrunk_xkid_delinq90_02020 (which represents the 90 day delinquency rate of the individual). Features were averaged across counties in every dataset, then they were all joined with each other in order to be able to analyze county level trends. Column names were renamed as follows: ‘par_pctile’ became ‘par_income_pctile’, 'shrunk_xkid_vscore2020' became 'credit_score', 'shrunk_xkid_brcbalance2020' became 'credit_balance', and 'shrunk_xkid_delinq90_02020' became 'credit_delinq_rate'. The par_income_pctile was dropped as it does not help us in our EDA and our initial dataframe was complete.

For the statistical summarization task, we took two different approaches. Peter stuck to Python and summarized directly from the CSVs. He took the mean and median of credit score, credit balance, and credit delinquency rate. He compared within groups in order to determine if there was skew in the distributions - helping us theorize whether certain distributions contained outliers. He found that the mean and median of credit score were similar - so there were likely no outliers. Next, for credit balance, he found that the mean was higher than the median, indicating rightward skew in its distribution. Next, for credit delinquency rate, he found that the mean and median were similar. Then, he calculated the range of each of these variables and found that they were large for credit delinquency rate and credit balance. Finally, to wrap up the statistical summarization of our data, he took the IQR of each of these three variables. The IQR of the credit scores is relatively tight - showing that across counties credit scores are similar. The IQR

of credit balances is larger, showing that there is more spread in the average credit card balance. Finally, the IQR of credit delinquency rates shows that there is also a marked spread of average credit delinquency per county. From our statistical summarization, we uncover several insights. To begin with, credit scores do not vary much by county - meaning that geographical differences do not seem to play a role in differences in access to credit. Second - we see that the interquartile range of credit card balances is somewhat high (~$500) and that the range is much greater (~$3,000). This indicates variance in credit card balances across counties. Finally, we see a somewhat large IQR in delinquency rates (0.09) and a much higher range (0.47). This indicates that some counties struggle with debt delinquency at a much higher rate than other counties - showing variation in the data. It is possible that debt delinquency and credit card balance are related to each other - and we will explore this later.

For the data visualization section, we began by plotting the distribution of credit scores, credit balances, and credit delinquency rates across counties. The X axis of our density plots are the variable of interest and the Y axis is the density. The distribution of credit scores by county, as well as the distribution of credit balances across counties, are roughly normal. The distribution of credit delinquency has an obvious leftward skew, indicating that there are likely outliers below the median. Next, we created three scatter plots. The first plotted the average credit delinquency rate of a county against the average credit score of a county. We observed what appears to be a strong negative association between the average credit score of a county and the average credit delinquency rate of a county. The second scatter plot plotted the average credit delinquency rate of a county against the average credit balance of a county. We observed a weak yet noticeable negative association between the average credit balance of a county and the average credit delinquency rate of a county. However, to understand the true statistical associations of our data,

we must show quantitative findings we calculate. That is what we do in our association analysis section.

After data preparation and initial analysis of the shape of our data, the next appropriate step was to begin understanding the variables that display statistical association between one another. A slightly altered version of the dataset used in the previous section was created for this segment of our work. The same strategy of grouping by county and averaging key variables was used, however we also grouped by race to add a new dimension to our visualizations later on. The results have allowed us to make informed guesses and understand underlying phenomena in our data. The informed guesses in question will be addressed shortly. To start we created a correlation matrix, which calculates the r-values for every possible combination of two variables in our data. The r-value is a desirable statistic as it gives us the linearity and direction of two variables. From the matrix, three key relationships were identified: Credit_score<>Delinq_rate, credit_balance<>credit_score and credit_balance<>Delinq_rate. From here we generated a scatter plot visualization for each, this way we would have the association statistics (r-value) as well as a new layer of information in the form of visualizations. This further bolsters our understanding of associations in our data. For every data point in each scatter plot, we assigned it a color based on the race value it was given, so now not only are we able to understand the associations, but we can see the role of race within the three key relationships as well. Put simply, each dot in the scatter plot represents a county-race subgroup since the statistics of each person in a county were averaged out by race, we now have multiple data points representing a certain county, each point being the racial group for that said county. credit_balance<>credit_score and credit_balance<>Delinq_rate both initially showed weaker r-values, this weaker correlation was also identifiable in the scatter plots for both. We know this

as there was no real coherent line shape in the scatter plot, although the direction of the data was notable. What we gather from this is that credit balance and credit score are positively correlated, while credit balance and delinquency are negatively correlated. In other words higher credit balance is associated with higher credit scores and higher credit balance is associated with lower delinquency. The really interesting relationship was Credit_score<>Delinq_rate, which had the strongest r-value, as well as passing the eye test. There is a clear and visible line shape from data points in a downward sloping direction. This means we can conclude that a higher credit score is associated with a lower delinquency rate and vice versa.

Adding race as a dimension to the data uncovered new notable patterns. It is important to note that this dataset does not represent individual people or entire counties. Instead, each row corresponds to a county-race subgroup. This is because we group the merged dataset by: 'par_state', 'par_county', 'county_name', 'kid_race' and then take the mean of credit_score, credit_balance, and delinq_rate. In all three of the scatter plots, data points labeled ‘Black’ were more frequently associated with higher delinquency rates and lower credit scores. Observations labeled ‘Asian’ were more often associated with lower delinquency rates and higher credit scores. Observations labeled ‘Hispanic’ and ‘White’ were often between the two preceding groups. These patterns are notable and can be used to motivate studies of systemic racism affecting access to credit.

In the predictive modeling portion of our analysis, we constructed a classification model to identify whether a county exhibits high or low delinquency rate using a median split to create a binary class label. From here the decision tree classifier was trained on the predictors: average credit score, credit card balance, delinquency rate and county identifier. The initial (“easier”) model achieved strong performance because of the inclusion of delinquency rate. This is because

the binary classifier label is directly derived from the delinquency rate. Although there is obvious overfitting and lack of predictive value, this model is still important because it showcases how a decision tree model is created and evaluated. We then wanted to take our model to the next step by making it “harder”. By following the framework used in Lab 8 and removing the raw delinquency rate used to construct the class label, the model would be forced to rely only on the remaining predictors: credit score, credit balance and county. As we already knew, credit score and credit balance strongly correlate with delinquency, this was reflected once again when the tree used these variables mainly to create the structure of the tree. The overall accuracy declined to 0.92, understandable as the model no longer has access to the exact delinquency values and must infer class membership from other predictors. This model has more predictive importance, as it can predict whether the delinquency rate of a county is high or low with 92% accuracy. Overall the harder model still reinforces a key insight from analyses: financial characteristics remain the dominant factors associated with county-level delinquency patterns, even when more direct information is removed from the model.

The steps taken from data pre-processing to advanced predictive models felt seamless overall, however there were some bumps in the road. For example a challenge Peter faced was wrangling the data to create features that would be useful. By reading pandas documentation, he was able to combine datasets to create useful data for the analysis conducted. James faced difficulty creating informative and detailed plots in the association analysis section - however review of material and documentation helped him.

Some interesting insights we uncovered through our analysis were that delinquency and credit score were both heavily correlated. They had a strong negative relationship, meaning that a higher delinquency is associated with a lower credit score. Adding race as a dimension to the

visualizations of the data shows how access to credit (as well as delinquency and balance) is distributed across races. Additionally, we found county race subgroups and calculated the top five counties with a low credit score and high credit card balance and the top five counties with a high credit score and a low credit card balance.

The list was as follows:

5 UPPER-LEFT POINTS (Low Score + High Delinquency)

	county_name	credit_score	delinq_rate	kid_race
11825	Castro	559.450000	0.859850	Black
11950	Dickens	566.050000	0.925850	Black
12763	Wheeler	569.300000	1.049250	Black
1659	Dixie	572.150000	0.867250	Black
486	Graham	577.717857	0.712482	AIAN

5 BOTTOM-RIGHT POINTS (High Score + Low Delinquency)

	county_name	credit_score	delinq_rate	kid_race
8462	New York	769.823810	0.129559	Asian
9002	Orange	767.550379	0.172131	Asian
13472	Charlottesville City	767.500000	0.105205	Asian
5622	Norfolk	766.448188	0.147567	Asian

Unsurprisingly New York tops the list, while Norfolk county, Massachusetts, places in the top 5, while 3 of the bottom 5 counties are all in Texas.

Some ideas for future exploration of the data are determining how credit score and access to credit affects quality of life. This is important because quality of life is directly linked to crime and health outcomes. If we can determine that there is a link between access to credit and quality of life, we can find out ways to help people - thus improving societal outcomes.