fico

April 22, 2020

0.1 Is it sound practice to use FICO credit scores to evaluate credit risk?

0.2 Case Introduction

Business Context. Lenders, such as banks and credit card companies, use credit scores to evaluate the potential risk posed by lending money to consumers and to mitigate losses due to bad debt. The consumer credit rating company Experian PLC classifies credit standings under the following score ranges:

Rating	Credit Range
Poor	300-576
Fair	580-669
Good	670-850

You are an analyst for a large loan agency, and your role is to help your company make decisions on whether or not to approve a loan.

Business problem. Your company uses FICO credit scores, among other things, to evaluate if customers at credit risk. You are **asked to investigate if this is a sound practice**.

Analytical context. You have information about 165,000 consumers from all over the US as shown in the above video. Thus you want to get a strong understanding of how FICO scores relate to consumer credit. We will investigate this question by examining a series of tables, box plots and histograms.

0.3 Data Summary

The data set contains 7 categories. The column account status has 41 different categories including "Current Account" (the account is in good standing), "PAID SATIS" (account paid satisfactorily) to "PD 30" (account was 30 days past due), "PD 30 6" (account was 30 days past due 6 times) and "COLLACCT" (account was sent for collection). **Please look at the supporting document to get more information about different categories in account status**. The other columns in this dataset are their state, city, Zip code, and their VantageScore, an alternative to FICO score.

Id	Acct Status	State	City	ZIP	Vantage Score	FICO Score
0	CURR ACCT	GA	DOUGLASVILLE	30135	601	642
1	PAID SATIS	GA	COVINGTON	30014	659	608

Id	Acct Status	State	City	ZIP	Vantage Score	FICO Score
2	CURR ACCT	GA	GRAYSON	30017	604	600
3	CURR ACCT	FL	CHULUOTA	32766	772	769
4	CURR ACCT	TX	COPPELL	75019	762	793

0.4 Profiles of Consumers With Current Account

A natural place to start is to investigate what kind of account types does your company most often approve loans for. You have observed that most customers with current account often have their loan approved. This makes sense since current account status usually signals good credit. The following plot shows the distribution of FICO scores of consumers with current account.

Let us see if the information contained in our FICO data set resonates with Experian's classification.

0.5 Question 1:

Consider the reasoning: Most of the FICO scores of consumers with current account are above 650. Thus higher FICO scores are associated with better credit. Is this reasoning correct?

Answer: While it is true that most consumers with current account have FICO scores above 650, we cannot make this conclusion yet because we do not know the distribution of FICO scores for other account types. If FICO scores are associated with good credit, it should be differentiated across various account types. In particular it should be lower on average for account types with poor credit.

0.6 FICO Scores Across Account Types.

The following table shows the top 5 account types with highest average FICO scores.

Account Status	mean	25th percentile	75th percentile
Paid to satisfaction	721	668	790
Good Standing	702	645	776
Consumer reported as deceased	695	615	786
Paid Account	681	584	768
Current, was 180 Overdue	677	605	738

The following 5 accounts with the lowest FICO scores.

Account status	mean	25th percentile	75th percentile
Merchandise was taken	559	513	594
Account sent for collection	558	483	603
Account seriously past due	541	507	574
Paid, was 150 days Overdue	541	511	597
Account Reported	529	495	556

0.6.1 Question 2:

From the above tables, and the Experian account classification document provided to you, what can you say about the relationship between FICO scores and credit?

Answer: We see that all accounts in the Table 1 indicate good credit. Even the 25th percentile of four of these accounts have scores above 600. It is also noteworthy that FICO score penalizes an account that was past due previously even if it is current.

For Table 2, all of these 5 accounts with the lowest FICO scores indicate bad credit. Thus it is reasonable to conclude that there is some differentiation of FICO scores among various account types.

0.7 Visualizing FICO Scores across Using Boxplots

An alternative way to compare the distribution of FICO scores across account types is using boxplots. A boxplot uses five key summaries to display information about a distribution. These are median (50th percentiles), first quartile (25% percentile), third quartile (75% percentile), "minimum" and "maximum". The difference between the third quartile and the first quartile is called the "Inter Quartile Range" (IQR). The following figure (courtesy: wikipedia) shows the structure of a boxplot:

As a warm up, the following boxplot shows the distribution of FICO scores for consumers from Texas whose accounts were sent for collection.

The dots represent outliers. These are scores which are 1.5xIQR away from the 75th percentile.

Question 3: Suppose that your firm approves home loans only for those customers in Texas whose FICO scores are above 700. Based on the boxplot above, which is the following statements is most accurate for customers in Texas whose accounts were sent for collection?

1. At least 25% of them will get their home loan approved.
2. At least some of them will likely get their home loan approved.
3. At least 75% of them will get their loan approval denied.
4. There are some customers whose FICO scores are below 400.

Answer: 1. Incorrect: The 75th percentile is below 550. The maximum of these scores is below 650. Thus the percentage of customers who will likely get their home loan approved is very small. 2. Incorrect: This dataset shows that the maximum scores are below 650. However this is only a sample, and we do not know if there are any customers in Texas whose credit scores are over 720 and their accounts were sent for collection. 3. Correct! The boxplot shows that the 75% percentile is around 600 and well below 720. Thus it is reasonable to conclude that at least 75% of them will get their loan approval denied. 4. Incorrect. The boxplot does not indicate this for this sample. We have no way of knowing this without extra information.

The following distribution of FICO scores across various account types in the same plot.

Question 4: What can you infer about the relationship between credit status and FICO scores from the boxplot above? Answer: The boxplot above shows that the FICO scores

corresponding to the current account and paid account ("Paid SATIS") has the highest median (about 720 and 750 respectively) compared to all other account types. The median FICO scores for other accounts are quite a bit below 650. Even the first quartile of the current account status is above 650, whereas the first quartile of most other accounts are below 600. Thus again we see that FICO scores vary widely across accounts, with good accounts having much higher FICO scores as compared to bad ones.

Question 5: What additional information does the boxplot above give as compared to the distribution of FICO scores of consumers with current account shown earlier? Why this important to our case?

Answer: The boxplot shows how the FICO score is **differentiated** across account types. Only looking at the distribution of FICO scores of consumers with account, we cannot determine whether FICO score gives any information for differentiating good credit vs. bad.

Let us look at the distribution of consumers across various states in our dataset. The following plot shows the percentage of consumers from each state. The darker colors indicate higher percentages.

Question 6: Consider the following statement: based on the plots and tables seen so far, we can now conclude higher FICO scores are indicative of good credit for consumers in the US. Choose the correct alternative.

1.This conclusion is fair because we have found that FICO scores of current account holders on average (both from the plot and table above) are quite a bit higher than consumers with account types.

2.This conclusion is fair because both the chart and the plot show that, on average a consumer with a higher credit score is more likely to have a current account than an account that was past due.

3.We cannot quite yet generalize this to all of US yet, because the dataset we have might not be representative of entire US. The dataset could contain consumers only from one part of the US such as the east coast.

4.This conclusion is fair because both the chart and the plot show that, a consumer with a higher credit score is more likely to have a current account than an account that was past due. Even if the dataset does not contain consumers from all over the US, consumer behavior does not vary hugely from state to state.

Answer: 3 is correct. We need to ensure that we have enough data about consumers from all over the US.

From the map above, we see that the states are not represented according to their population in this dataset. For example, California accounts for about 10% of the US population, but our dataset has less than 1% consumers from California. However, we can conclude that FICO scores indeed are indicative of good credit in those states that have a large number of consumer in our dataset, including Texas, Illinois, New York etc.

1 is incorrect: Yes, this is a fair conclusion for the consumer population represented by this dataset. But we do not know yet if this dataset is representative of the entire US. Our FICO dataset could have customers mainly just from the east coast in which case we cannot generalize this conclusion

to all of US unless we have additional information 2 in incorrect: Yes, this is a fair conclusion for the consumer population represented by this dataset. But we do not know yet if this dataset is representative of the entire US. Our FICO dataset could have customers mainly just from the east coast in which case we cannot generalize this conclusion to all of US unless we have additional information. 4 is incorrect: Even if consumer behavior does not change that much from state to state, we cannot assume this while summarizing information from data. In fact this motivates us to check if this assumption is valid.

0.8 If FICO Scores are Not Available, Can We Use VantageScores Instead?

Next suppose that you have a new customer but you don't have their FICO score available, but do have their VantageScore. VantageScores are a main alternative to FICO scores. Credit bureaus Experian, TransUnion and Equifax came up with an algorithm to produce VantageScore. Thus you would like to know if VantageScore can be used to evaluate their credit status.

The following plot compares FICO scores to Vantage scores.

0.8.1 Question 7

The graph above plots FICO scores vs Vantage scores for all of the consumers in our dataset. Which of the following the statements can be made from this plot?

1. If a consumer with current account has high FICO score, on average, one would expect him/her to have a high Vantage score.
2. FICO scores are similar to Vantage scores only for consumers with good credit.
3. FICO scores are positively correlated with Vantage scores. Thus if a consumer has high FICO score, he/she is expected to have high vantage score.
4. There isn't enough information in this plot to deduce any of these statements.

Answer: 3 is Correct! It is now interesting to see if this positive correlation is preserved across all account types.

1. We don't know that yet.
2. We can't tell that yet either, because we haven't looked at this across accounts.

The following plots shows that FICO scores and VantageScores are indeed positively correlated across various account types.

Thus it seems like VantageScores are a resonable alternative for FICO scores.

0.9 Case Summary

We saw that if the FICO scores were to be associated with good credit status, the scores should be differentiated across account types. We can compare the distribution of FICO scores across account types using boxplots. By doing this, we saw that accounts with good standing had highest FICO scores and delinquent accounts had some of the lowest credit scores. Thus it seems that FICO scores are a good quantity to evaluate credit risk.

We also needed to check if our data was representative of all of the US. We did this by overlaying the percentage of consumers from each state on a map of the US. We saw from the map that a few states were over represented in our dataset; a vast majority of the states were underrepresented. When FICO scores were not available, we saw that we could use VantageScores instead. This was because the FICO scores and VantageScores were positively correlated across account types.

0.10 Takeaways

1. If we want to examine if a numerical quantity (credit score) to make a decision, we should verify if it is differentiated across categories.
2. Boxplots across categories are an important tool to see how differentiated the categories are.
3. Finding correlations between variables lets us one of them as a "proxy" for the other.