Classification Credit Card Project

By: Andrew Acosta

Introduction to the project

I have been given a prompt as a “risk analyst” of a bank. This particular bank offers a credit card service to its clients. This is an important part of revenue for the bank. For this reason, they want to know what truly affects the clients to accept or decline the offer they give out. They have given me a complete dataset of records for all their customers that might be useful as to seeing why they accepted or rejected the offer given. I must go through this dataset and perform data operations to simplify these factors that really play a role on if the offer is accepted or not.

Exploring the data

Like I mentioned, I must truly understand this dataset. If I don’t explore all the columns and recognize biases and correlations, than performing what is expected of me will be much harder. When the dataset was first loaded I could see that the headers were wrong and contained what should have been the first value in each column. I fixed the headers to display what they should be, and got a basic idea of the columns by looking at they’re first few values and their data types. I dropped any unrelated data columns and got to exploring the details. I used the describe and isnull function to check basic correlations and empty values for the columns. After dropping the few nulls the columns had, I was ready to start to visualize the data at hand.

Visualizations

This is a straightforward way of understanding the data. I used pairplots, distplots, count plots, and box plots to visualize the data and see what the “offer accepted” column was being affected by. For example in this image here I plotted a countplot that showed the amount of people who accepted and declined the offer based off of their income.

Features and Target

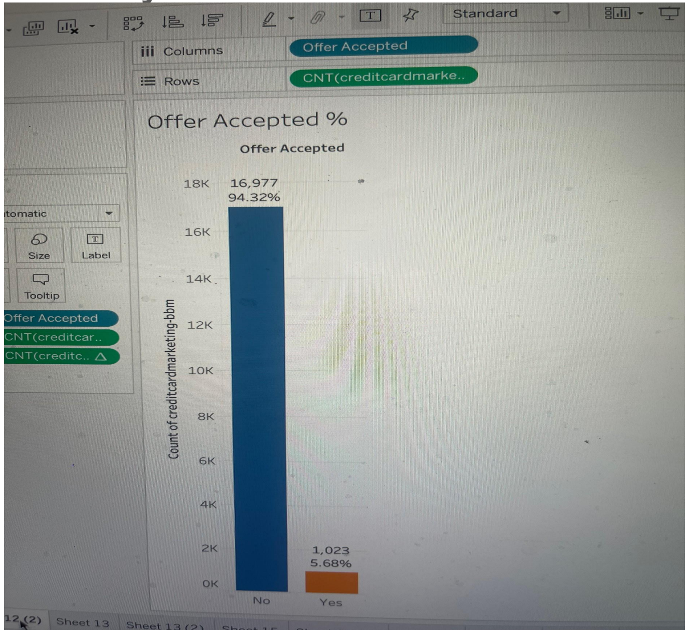
I now split the features and target into x and y variables. After splitting the x values into categorical and numerical features, I scale the numericals and encode the categoricals. While I do this I also am looking for correlations in the values of the columns. At this point I’m very familiar with all the value and value counts of the columns.

Logistic Regression and KNN

I now set up the data for fitting models. The target column “Offer Accepted” showed a big imbalance of values yes and no. Since it was a large imbalance, I used RandomOverSampler to avoid losing useful data. I then performed train test split and fit the values into the logistic regression model. I got about a 7.1 accuracy score and examined the confusion matrix. I think the amount of false negatives would affect this data the most. This is because of someone accepts the offer but I falsely predicted negatively, then the company is losing money when they should have another client for the credit card service. Unlike if I predict a false positive, the company will most likely reach out again and again until they respond. If anything it can produce a positive outcome. I got little false negatives and more false positives. Now when I used the knn model it predicted a .91 accuracy score but the confusion matrix showed a very biased outcome toward true positive.

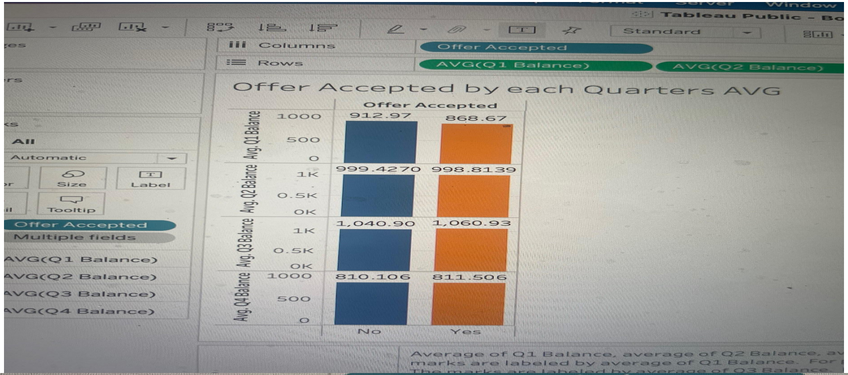
Important visualizations

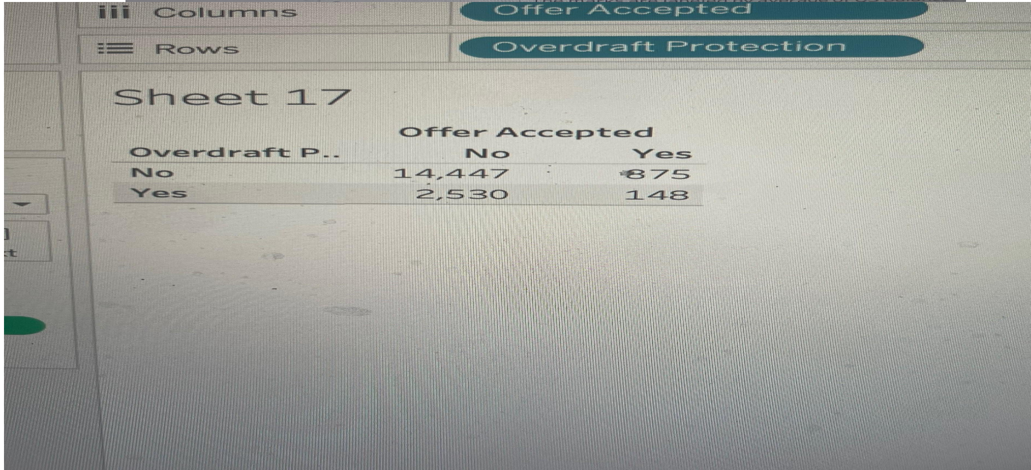
On tableau I observed some promoted questions that really helped visualize the important data. This first graph shows the count and percentage of the “offer accepted” column. This shows the true imbalance shows how if I just performed the model without balancing the database, it will be truly biased.



Important Tableau visualizations

This next graph shows the balance of the customers paired with if they accepted the offer or not. It shows that there’s no significant signs of a customers average balance being a reason someone accepts the offer or not. The second image shows that overdraft protection doesn’t truly indicate a correlation either.





Overdraft P..	Offer Accepted No	Offer Accepted Yes
No	14,447	875
Yes	2,530	148

Important tableau visualizations

This last person I'll show is an important one. It shows the customers value of income, being low medium and high. This graph shows that the lower the income, the more the customer will accept the offer. It clearly shows people who have a higher income probably just don't feel the need for that credit card service which makes a lot of sense. Overall this dataset showed that income definitely has a play in if someone accepts the offer. I also observed that a lot of these customers could just never see the offer especially if it was sent digitally. If the offer was presented by every employee while a client makes a transaction or payment, then more clients will be made known of the offer compared to however they have been offering the service.