Delivery Tip Prediction Final Evaluation

Introduction

A delivery driver from a Chinese restaurant in Charleston, South Carolina took the time to collect data related to food deliveries that he made in the local area. By evaluating this data, we can determine if there are certain factors that influence a customer's tip amount.

A potential problem related to this data could present itself if the tips that delivery drivers receive vary in amounts. It could then be determined if it is due to location of services or any of the categorical data related to the delivery like race, age, place of delivery or gender. Another potential problem could be if the restaurant owner is considering splitting tips equally between delivery drivers because drivers are receiving an unequal tip amount from the service. The owners could decide on whether to continue to let the delivery drivers possess their own related tips.

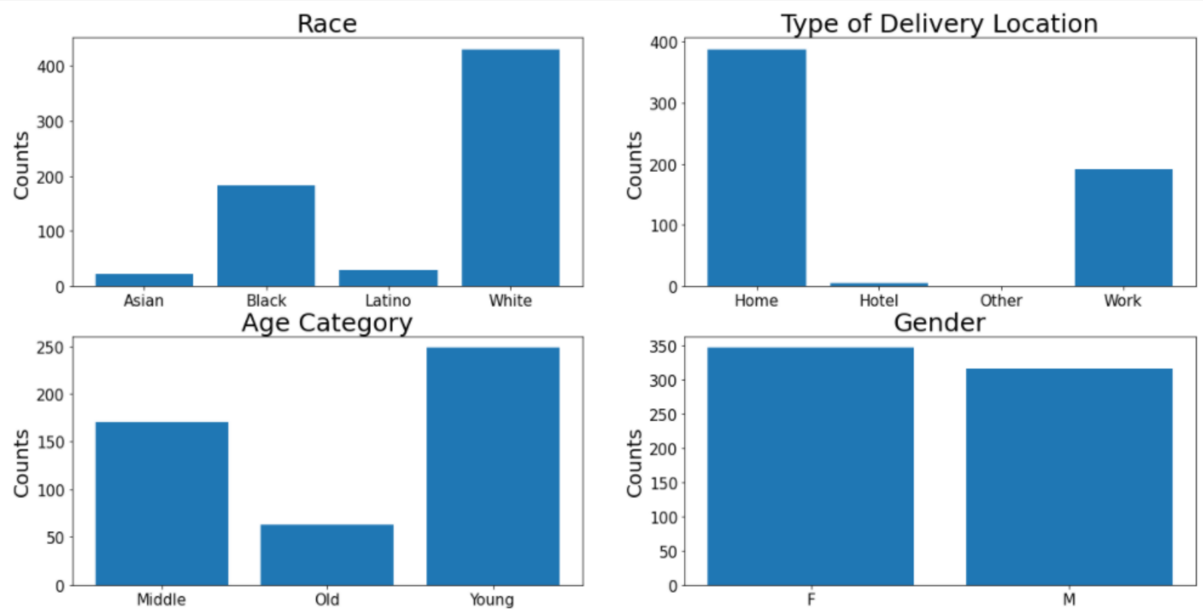
The data was obtained from https://www.kaggle.com/dustincm/chinese-delivery-drive. The data was collected during a three-month period. Some data was derived from the U.S. Census Bureau related to the Charleston area based on race to identify average income.

Detailed Summary

EDA

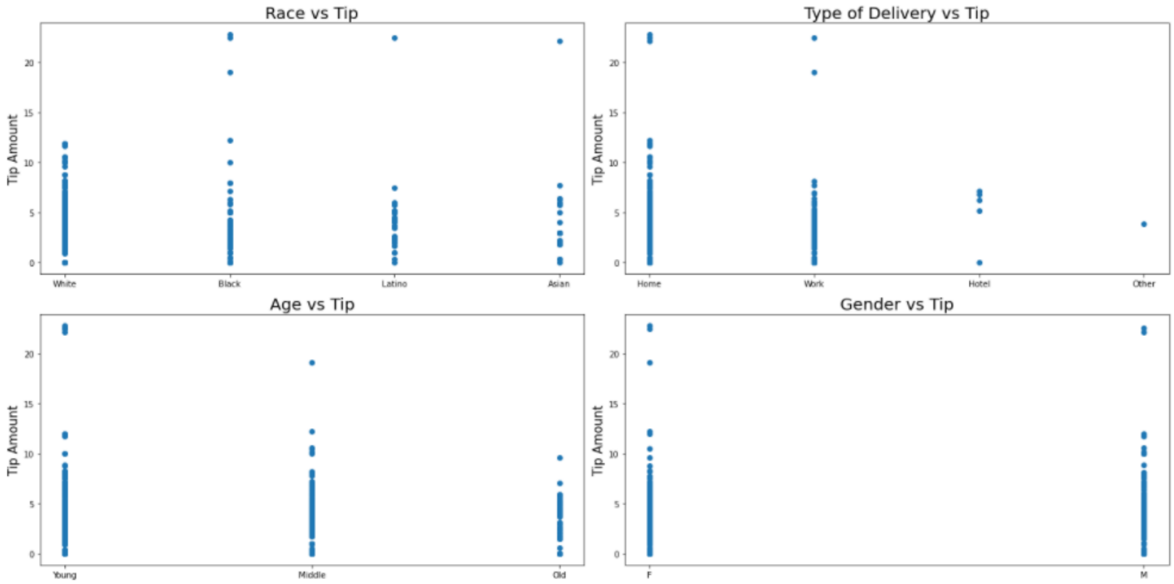
The initial evaluation of the data identified 772 rows and 15 columns. Columns identified as not useful were Date, Phone_Number, Payment_Type and Street_Address2. The focus will be on the tip amount column and the price of the food.

If the values in the tip or price columns have no value, those rows will be replaced with a median (average) value. Any categorical data to include race, gender, type of delivery, or age category that has no value will be replaced with the mode (most) value.



Looking at the visualization summarizing the evaluation of the categorical data, the race column is highly dominated by white people, with black people next, representing about half of the white people. The type of delivery location identifies where the food is being delivered. This category shows most deliveries being made to a home and the next delivery place is work, which represents about half of the home deliveries. The age category identifies most deliveries made to young people, with the next category being middle aged then older people. Lastly, I looked at the gender division which is evenly divided.

The next visualization looks at the tip amount in relation to the categories and where we may see larger tip values vs the categories determined.



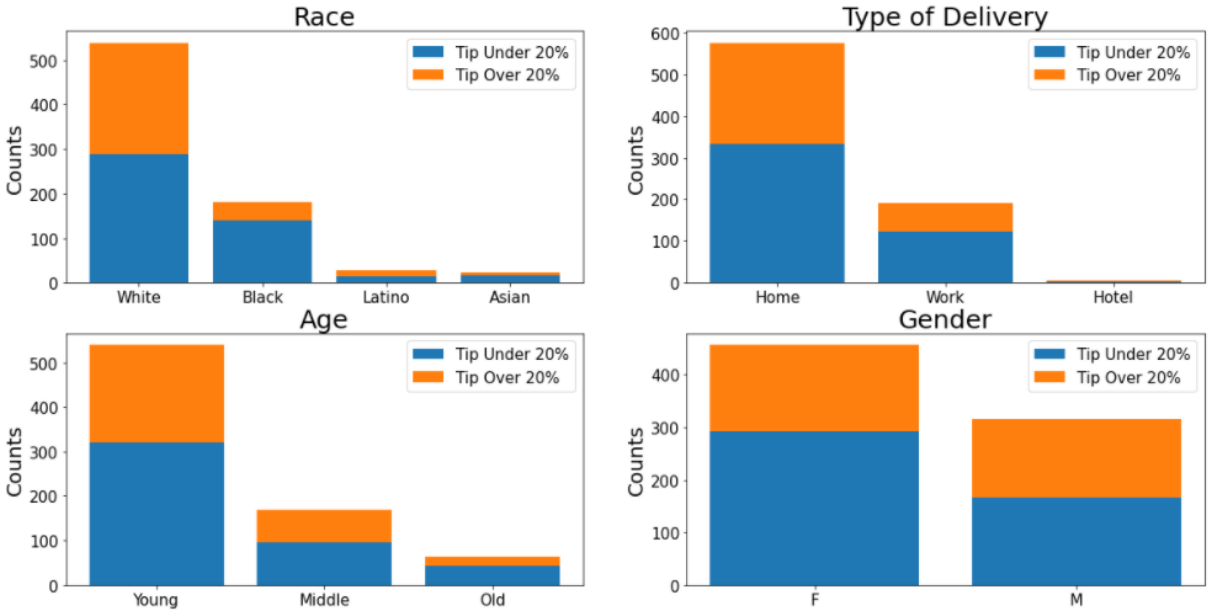
This analysis shows the general area where the tip amount is for a particular category. There are some instances where the tip is much higher. I do not consider these outliers as much as I do exceptions to the average tip amount.

Data Preparation

Data preparation steps included the removal of columns that were unnecessary. By creating median values of the tip and price amounts and mode values for the categorical columns, I was able to maintain the same number of records I started with in the original data.

I created a scatter plot of tip vs price and added a regression line. The regression line was not steep as I would have expected but there is an indication that when the food price increases so does the tip amount.

After looking over my data, I decided to create a column to identify any tips that are more than 20 percent of the food price and those tips that are less than 20 percent of the food price. I can use this in the data modeling.

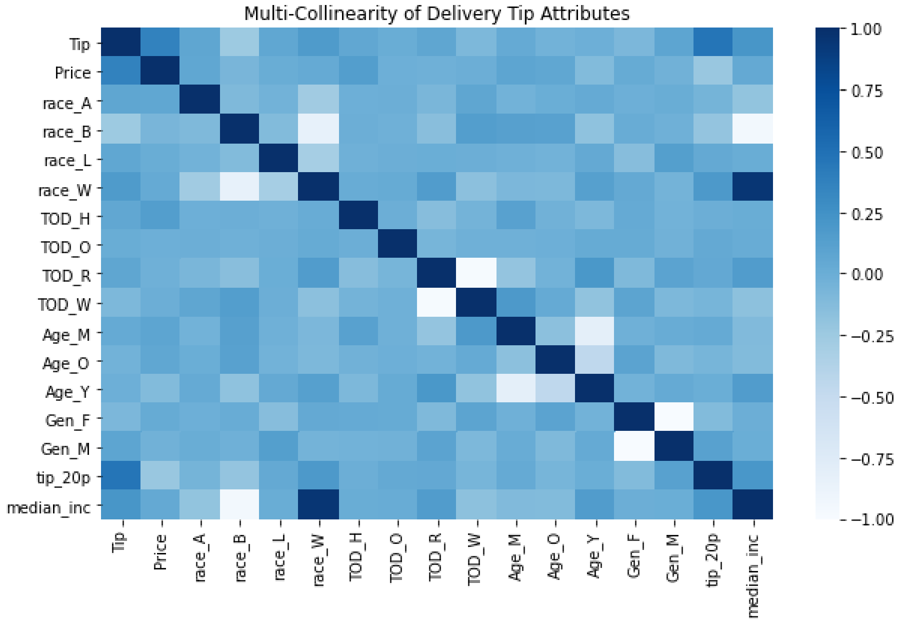


Once I created the tip over and under 20% column, I created stackable bar plots to compare each categorical column in relationship to the tip amount. Visibly, the division is close to even in all categories except in the race columns, blacks tend to tip lower than 20% and the same with the work location.

I reviewed information on the internet for income levels in the Charleston area. My intentions were to use the street address to determine a general income level. This proved to be difficult as the street data was incomplete and hard to update without looking up every single address. I was able to find on the U.S. Census Bureau website an identifier for that area of income by race. I then created a column in my data, based on race, of a median income amount.

Next, I worked on datasets for my dummy variables from the categorical columns. I removed the Street_Address, Time_In, Time_Out and Time_Elapsed fields as they were not complete and had no value for the study at this time.

By creating dummy variables, I was able to evaluate my data using a heat map, as shown below:



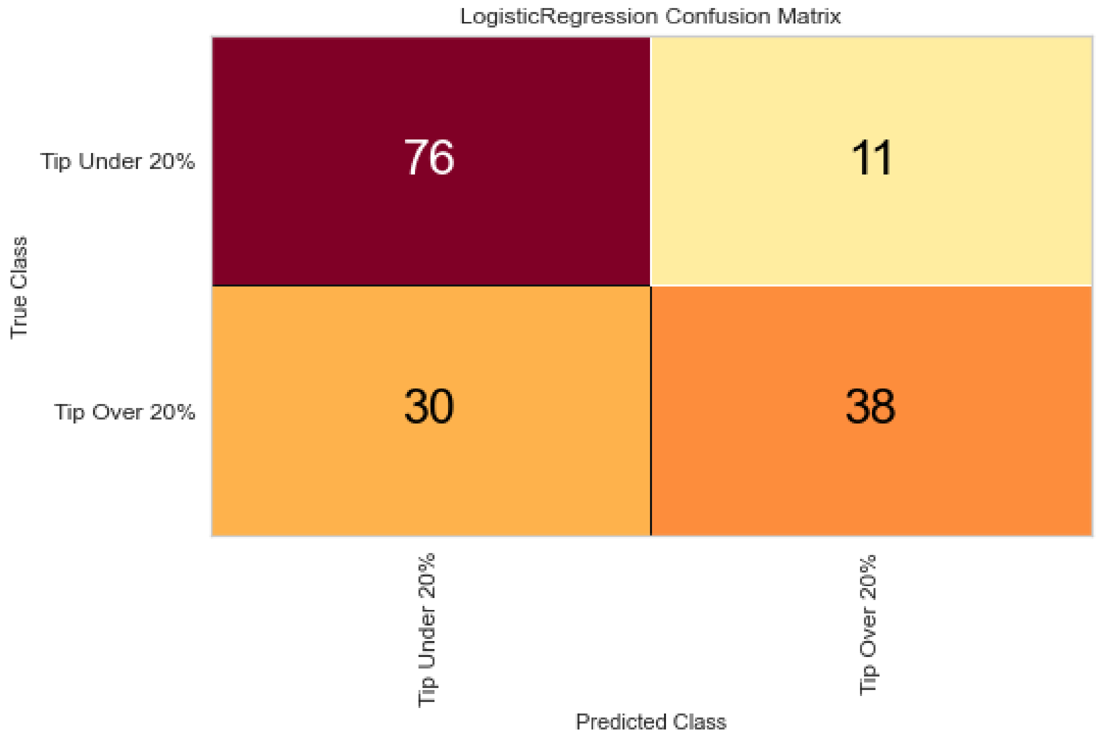
Positive correlation exists where the squares are darker in shade. The white race has a positive correlation with median income, and it also shows a positive correlation with the tip and the tip 20 percent calculation.

Model Building and Evaluation

To prepare the data for modeling it was split into two separate data frames with an 80/20 split. The first split I created, to run a linear regression model, compared food price to tip amount. With this model, the coefficient of determination was 14% which means the price is not a good predictor of tip amount and it is not a good fit for this model.

I modified the data frames to include all the variables and perform a multiple linear regression model using all the dummy variables and the target was the tip amount. This data model proved to be a better predictor of the tip amount with a coefficient of determination of 42%.

Next, I decided to use the model logistic regression and look at the predictions relative to a tip over 20% or under 20%. This data comparison provided a much better accuracy of 73.55%. The following is the confusion matrix output of this analysis.



This confusion matrix shows the results of logistic regression of my test data frame.

Last, I decided to work my data into the KNN classifier, and the results were so much better. KNN supports non-linear solutions. While my data has a linear appearance, it could be considered non-linear. The KNN classifier, when using the prediction of the test data shows an accuracy of 89%.

Conclusion

Street addresses would need to be cleaned up or presented in a more functional way to be used for analysis. If it were decided to use them, I would take the time to determine the value it would provide for the process.

Some of the take aways from the data modeling is to understand the linear vs non-linear appearance of your data. This can help you determine the best possible model to use. I was happy with stepping through a variety of models to see the results.

The model may not be ready to be deployed if the idea is that it is ideal to have accuracy percentages in the 90s. In relation to the project and reason to identify the values if a tip is over 20% or under 20% to decide if we want to split tips among delivery drivers, this may be good enough to apply to new data and give us enough information to make some basic business decisions.

I would recommend gathering more data. Gaps can be realized within modeling by simply adding that volume. Predictions made can be scrutinized when looking at factors like age, race, and gender. Careful consideration of these categories must be evaluated if used on a much larger scale with public entities of any kind for ethical reasons.