Project 2

Scenario 2: Market research for new store locations

Aakash Tandel

1) Identify the problem:

* Write a high quality problem statement: Determine which county/city/zip code in Iowa will yield the highest sales. Then look at the relationship of various variables on the total sales for the top stores in that county.
* Describe the goals of your study and criteria for success: The goal of this study is to determine a geographic location (county) in Iowa that will yield the highest amount of liquor sales. Once we have the location specified, we can look and see if there is anything specifically that helps total volume sold. ((coupled with number of current liquor stores per area, can offer recommendations on where to open a new liquor store)). High sales and low numbers of existing stores will most likely lead to higher profits for the new liquor store, on average. The criteria for success would be determining 1-5 new locations in which the company could open new liquor stores and determining which factors effect total sales and how.

2) Acquire the Data: Iowa Liquor Data

* Jupyter Notebook

3) Explore the data

* Jupyter Notebook
  + Import data using the Pandas Library – Done
  + Perform exploratory analysis with visualizations and statistical analysis – Done – This is a large dataset and that is a great thing. When loading the full dataset, there will be upwards of 2.7 million observations. We removed the 2,973 duplicated columns from the dataframe, as well as the all of the null values. Because of our large number of observations, this should have very little effect on our analysis. For our location data, we can see that there are 100 county numbers, 99 counties, 383 cities, and 676 zip codes. It would be wise to cross-reference this data with the state's municipality records to make sure the location variables are properly matched across city, county, and zip code. We see that a large number of observations are found in Polk County, the city of Des Moines, and the zip code 50010 (Ames, Iowa). Ames is the home of Iowa State. This makes logical sense because these are the main urban centers in the state of Iowa and a larger number of people should correlate positively with a higher number of liquor sales. We have 72 different categories of alcohol. These are highly differentiated. If we were to analyze the categories further, it may be wise to group in broader categories. For example, all whiskeys and bourbons could be in one category, all vodkas in another, etc. There are 1400 unique stores in the data set. The vast majority of sales are of quantities of less than 100 bottles and of transactions less than $1,000. We continue our analysis below.
  + Risks and Assumptions: Risks and Assumptions: One major assumption we will make is that the location data and sales data is correct. There are ways to verify both of these points, but that is beyond the scope of this project. Data entry error, missing information, and other data problems could drastically affect our predictive model. We are going to assume the data is valid after we corrected for the duplicated and missing values as stated above. Another assumption we are going to make is that the errors in sales data are normally distributed and not correlated. This could be violated if one Iowa liquor surveyor consistently erred in data entry in the same manner. For example, say this person accidently increased all of the Des Moines city sales data by a factor of 10 (incorrectly placing the decimal point in the data entry process). This would systematically correlate the error across multiple observations and violating the linear model assumptions.

4) Mine the data: ((The best way to determine which locations provide the highest return, I have decided to create a new dataframe or table. This table will be organized by zip code and contain columns of the following: total sales (in dollars), price per bottle, total number of bottles, number of vendors, and volume sold (in liters). We will train our model with the total sales (in dollars) column. The zip codes with the highest predicted total sales will be the zip codes to open a store in (assuming our model is reliable.))) The majority of the data mining in this project was done with dummy variables. I have created dummy variables that indicate which county an observation is in. This will allow us to model for different locations.

* Create necessary derived columns from the data
  + Jupyter Notebook
* Format, clean, slice, and combine the data in Python

5) Refine the data

* Determine outliers, skew distribution of important variables (if any)
* Determine correlations / causations in the data: Used correlation matrices to perform feature selection. This resulted in our original model only including the location dummy variables, state bottle retail (which is similar to price per bottle), and bottles sold.
* Validate findings using statistical analysis (p-values, confidence intervals) as applicable: After running our model through statsmodels, we see a few things. The linear regression answers our first question. We see that both Bottles Sold and State Bottle Retail remain statistically significant when we add in the location parameters to our model. This model is all relative to Polk County because we have excluded one dummy variable from our regression (to prevent falling into the dummy variable trap). The county that has the highest beta coefficient (which is a higher y-intercept for this dummy variable) is Dallas County. The dummy variable for Dallas County is also highly statistically significant. Based on this regression, we would say that opening a liquor store in Dallas County could potentially lead to more sales than opening a liquor store in other counties. Dallas County is located just west of Des Moines.
* Reduce the data set
  + The location data (zip code, county, city)
  + The total sales for each month (or at least for Jan-March for the first scenario)
  + The total number of bottles sold
  + The average price per bottle (perhaps useful for distinguishing store types)

6) Build a data model

* Complete linear regressions using scikit-learn or statsmodels and interpret your findings: I first split the data into a training and test split. Then, I used lasso regularization on training data. Using L1 regularization, we see that a lot of the variables drop our (become 0) in our regression. The dummy variable for Dallas County still remains. This dummy variable also has a very large beta coefficient. This indicates that it will have sales higher than the base case (which is Polk County).
* Calculate and plot predicted probabilities and/or present tables of results
* Describe the bias-variance tradeoff of your model and errors metrics: We run the risk of overfitting with our linear model including all of those county dummy variables. That is the reason Lasso Regression is a good idea with this model. In the end, we use Lasso to drastically drop our MSE. As our MSE goes down, so should our variance (and increase out bias.
* Evaluate model fit by using loss functions, including mean absolute error, mean squared error, and root mean squared error, or r-squared: Our end model determining the county location was regularized with L1 regularization. That drastically reduced our MSE from around 2 billion in the linear model to about 170 in the lasso regression. Our end model determining which factors most effected. Using a lasso regression, we have much smaller MSE. We also see that Bottles Sold has the biggest effect on Sales Dollars in Dallas County. Thus, if we were to make a recommendation to a store opening in Dallas County, it would be that maximizing the number of bottles sold is the highest way to increase sales. Sheer number of bottles sold is a better predictor of high sales than item type, average volume of bottle, or anything else.

7) Present the results

* Create a Jupyter Notebook hosted on GitHub Enterprise that provides a dataset overview with visualizations, statistical analysis, data cleaning methodologies, and models
* Create a write-up on the interpretation of findings including an executive summary with conclusions and next steps

8) Bonus

* Regularization (Ridge & Lasso regressions)