Hanna Co
CS M148
Discussion 1A

Introduction

With weed being legalized only fairly recently, and slowly gaining acceptance with the general public, it's not surprising that the retail market has grown too. With this in mind, we had chosen to utilize various data analysis techniques to predict sales, but more importantly, see what factors influence sales the most. For some, they believe that convenience of the service is more important, especially in times of pandemic. The article "Best Dispensaries in LA" lists the best as ones that provide services such as delivery, having multiple locations, or carrying certain brands that are owned by BIPOC. However, we are more interested in seeing what factors influence the sales of a certain product, rather than what influences the success of a certain store. This is where BDSA comes in. According to their research, the flavor and brand recognition appear to be the most important factors. Additionally, the co-founder of Cookies, a prominent CBD brand, emphasizes the importance of branding, advertising, and exposure.

For this project in particular, I wanted to investigate what are the most important factors that influence sales, taking into account how many products a brand carries, brand, as well as previous sales data. Additionally, I wanted to see if season and year are important factors–with stigma around weed slowly lifting, will we see sales increase? Using data collected from various sources, we will analyze it and attempt to find answers.

Methodology

Before we analyzed our data, I wanted to add specific features that I thought would be the most influential on total sales. I imported data from BrandTotalUnits, BrandTotalSales, BrandDetails, BrandAverageRetailPrice, and modified it to get more information from it. For example, I computed the rolling average from the past three months for units, ARP, and sales to observe past trends. If there wasn't data for the past three months, I took the past two months, or the past month's data. If there were no data points to observe, I simply set the rolling average to 0. I also added inhalables and edibles as categorical variables, because I thought it would be important. Additionally, I also added how many products a brand carries, thinking that a brand with more products is likely to be better known. From the month, I extracted the year and season, curious to see if these factors were important.

To augment and categorize the data, I calculated the average ARP of each brand, and split the brands into twenty categories. I did this to replicate how many industries are split into high-end and generic brands. I also added previous months sales, divided by the rolling average. I dropped features I didn't think were relevant, such as previous month's data, or features that I already extracted the necessary data from, such as Brands.

After fitting certain models, I utilized GridSearch to find the best parameters for those models. Additionally, I also used some other models to see if I could achieve better performance, such as Random Forest Regressor, and ran a Grid Search on that as well. I compared the RMSE and R^2 of my models to find the best one.

Results

I first fitted my data to a linear regression model, which achieved an R-squared score of 0.676 and a root mean square error of 107421.23, neither of which are ideal. To try and improve these scores, I used BaggingRegressor, with Linear Regression as the base estimator. With this, we achieved an R-squared score of 0.6819 as well, and a root mean square error of 106473.57.

To try and improve further, I performed cross validation on both methods and got R-squared values of 0.677 and 0.982 for linear regression and BaggingRegressor respectively. Additionally, I performed a GridSearch to find the best parameters for BaggingRegressor, which I found to be {bootstrap=True, max_samples=25, n_estimators=85}. Fitting the model with these parameters, I achieve a R-squared of -1.15 and root mean squared error of 276814.73. The negative R-squared means that this model, unfortunately, performs worse than a horizontal line.

Finally, just to test the performance of other models, I used RandomForestRegressor. Although I got a decent R-squared value of 0.75, I got a root mean square error of 94390.11. I did a GridSearch for this too, and the R-squared decreased to 0.7317, the root mean squared error stayed high at 97789.44. I found it strange that RandomForestRegressor turned out to be the best model for my data.

Using my linear regression model, I calculated the p-values for each feature to determine which were significant. I considered features with a p-value greater than 0.05 to be insignificant. These included Average ARP and Season. These results surprised me, as I thought that Average ARP would be more important. From this, we can conclude that the other features are significant.

Discussion

The R-squared values measure the correlation, so the closer it is to 1, the better. Although the RMSE is very high, I believe that this could be because the values for Total Sales are high, and have a wide range. From my analysis, it seems that RandomForestRegressor is the best for this dataset, which is strange considering it's usually used for classifying data. Thus, I could recommend Munchies gather further data, perhaps data that is more product specific, or based on consumer surveys, and re-evaluate. However, one conclusion from this is that features such as Brand, Average Retail Price, Inhalables, Edibles, ProdCount are all important when determining sales. With this in mind, Munchies can focus their efforts on providing a wide range of products at an affordable price, and focus on brand presentation.