Karina Santoso

COMSCI M148

Professor Baharan Mirzasoleiman

16 March 2022

Project 3 Final Report

Background/Introduction

The newly emerging and rapidly growing cannabis market is becoming more prominent in many states. The medical use of cannabis is now legal in 37 states, while recreational use has been legalized in 18 of those states. 13 states in addition to these 18 have decriminalized the use of cannabis, though it remains illegal by federal law. New innovations and products, such as edibles, beverages, and concentrates, are quickly emerging in the industry as demand continues to increase, with the US driving the global growth of the cannabis industry. BDSA predicts a global legal cannabis market of $56 billion by 2026.

Since this is such a novel industry, it is difficult to know how consumers may react to a new cannabis product or company and how successful it will be. Using four different datasets that include financial and product details for a number of cannabis brands, we attempt to hone in on the factors that make a specific brand successful. To do this, we model the monthly total sales of a brand based on its other financial indicators and product inventory. This will allow us to not only model the previous history of the brands but also predict the revenue of these different companies in the future. Thus, this analysis is important not only for prospective new companies but also for existing companies to see what factors do and do not make them successful. I chose to model the monthly total sales of a company as this is an objective measure that is not based on how long the company has existed.

Due to the difference in legislation regarding the drug in various states, we focus our analysis on the cannabis market in the state of California, where marijuana has been legal since 2016. In 2021, California alone has a cannabis market of $4.2 billion according to BDSA. This allows us to draw insights that may be useful in predicting the growth of the industry in states that legalize the drug in the future. Additionally, we also only look at legal/licensed sales as we only want to look at regulated products and sales that are safer for consumers and will contribute to positive growth of the market.

Methodology

A number of different steps were taken in order to effectively put together, clean up, and then model this data. These steps are outlined below:

Dataset Merging: As mentioned previously, four different datasets were joined together to conduct this analysis. The information that these datasets contained were: 1) monthly average retail price over time per brand, 2) monthly total units sold over time per brand, 3) monthly total sales over time per brand (in US dollars), and 4) details about the various products sold by each brand. Since the information in all four of these datasets seemed extremely relevant to the problem we want to model, I merged them into a single dataframe to analyze.

The first three datasets containing average retail price, total units, and total sales information were able to be joined by brand and by month. However, the fourth dataset containing brand details did not have a time series element and had multiple rows per brand. To deal with this issue, I filtered out the columns of interest and got a list of the unique entries for each column for each brand. The columns chosen give information about the type of products sold by the brand, their contents and features, and price, all of which a buyer will take into

account. Then, I merged this dataset with the previous one by brand so that each row for the same brand had the same brand detail information.

I then did some minor data cleaning, dropping rows that had NA values for all three of the columns: average retail price, total units, and total sales, as well as formatting the total sales and total units columns into a numeric rounded value. The data related to the brands' finances is essential to our modeling approach, so rows that do not contain values for all three of these columns will not be useful.

Time Series Features: Next, in order to take advantage of the chronological data we are given through the brands' financial information, I created a number of time series features to be used in the later modeling. First, I added year and month columns to the entire dataset to track trends both over time and seasonal.

The next set of time series features were added for each brand separately. Filtering the dataset for each brand, I added a 'Time in Market' feature that denotes what number entry the row is after ordering the dataset chronologically. I also added a 'Total Units Rolling Average' feature that took the average of the total units sold for the past three months. Finally, I added a 'ARP Percent Change' feature that calculates the percent change in average retail price compared to the previous month. These features allow us to take into account the age of a brand which allows it to build up its reputation and products as well as its sales history which is a good predictor of whether or not the company is doing well and will do so in the near future.

Feature Engineering: To prepare the dataset for modeling, some further data cleaning and feature engineering was necessary. To remove the columns of the dataset that were of type list, I created a number of binary categorical features based on whether a certain string was in the list or not. These columns were: Edibles, Inhaleables, Contains CBD, Pax Filter, Mood Effect, Generic

Items, and Under 20. These columns have a 1 if the brand sells a product of the type specified and a 0 if not. The Under 20 column denotes whether or not the brand sells products of price under $20 or not. Edibles and Inhalables are the two most popular forms of cannabis consumption, and $20 was chosen as a somewhat arbitrary cutoff to try to determine whether a brand only sells expensive products or not. Whether or not a product contains CBD, a pax filter, or a mood effect heavily impacts a user's experience of the product and their desire to use it again. As responses to different products may vary heavily from person to person, many consumers may be more likely to buy generic items as they know the exact effects of these products better. All of these factors are likely important to a consumer when they make a decision whether or not to buy from a brand, which is why we incorporate them into our dataset.

After dropping the unused columns as well as our response variable column which we save separately, we see that the only columns with NA values are the Rolling Average and Percent Change columns. This makes sense as rows that are the first of their brand type in the dataset do not have previous data to calculate this information from. NA values cannot be in the dataset to run various models on it, so I imputed NA's and infinite values in the Percent Change column with 0's as there would be no change in the first month of a brand's history and imputed NA's in the Rolling Average column with the median of that column.

I additionally augmented the features ARP and Time in Market by creating a new feature called 'ARP over Time in Market' which shows the brand's success relative to the amount of time the brand has existed, which is an important indicator to see whether or not the brand continues to grow over time even if it is successful. The data was run through a pipeline to encode categorical features and scale numerical values as the final step to prepare it for modeling.

Modeling: A number of different models and techniques were used to model the data, including linear regression, random forest regression, K-Nearest Neighbors Regression (KNN), and Support Vector Regression (SVR). A variance inflation factor analysis was done on the predictor variables to filter out features to use for our linear regression to avoid collinearity, and the p-values from the linear regression were used to determine features of importance in predicting total sales.

A 10-fold cross validation was then run on the linear regression and random forest regression models to confirm our results were not a result of overfitting, and a grid search was run on the random forest regression to find its optimal parameters. Finally, experiments were also run to find the optimal k value for KNN and to see what the optimal parameters for SVR might be in terms of its C-value, epsilon, and kernel.

Results

Linear Regression: After splitting our dataset into training and testing sets, we train an ordinary least squares regression on our training data. We get a test R-squared value of 0.2006 and RMSE value of 177593.6667. From the coefficient estimates obtained from this model, we see that the predictor variables that have a p-value of below 0.05 are ARP, Total Units, Contains CBD, Pax Filter, and Mood Effect, or our first 5 variables. The other 9 variables have p-values of 0.391 and do not seem to have a significant effect in predicting the response variable. When this model was run, a warning was produced that multicollinearity issues may exist in the data.

To attempt to solve this issue, a variance inflation factor (VIF) analysis was run on the columns of the dataframe of the predictor variables. It was found that Year and Under 20 had extremely high VIF values of over 40 and Total Units and Total Units Rolling Average had fairly

high VIF values of around 11 as well. Thus, the columns Year and Total Units Rolling Average were dropped from the data frame, and the OLS linear regression model was rerun.

This time, we get a test R-squared value of 0.1584 and RMSE value of 182221.6161. Looking again at the coefficient estimates of the model, we see that the predictor variables with a significant p-value are now ARP, Total Units, Contains CBD, and Pax Filter. The other 10 predictor variables have p-values of 0.511 and thus do not seem to be significant in predicting total sales. The warning describing potential multicollinearity issues persisted even with this model.

A 10-fold cross validation was run on the original linear regression model to ensure that no overfitting was occurring. Doing this produced a test R-squared value of 0.2118 and RMSE of 177690.6407.

Random Forest Regression: To create a more optimized prediction model, an ensemble method was implemented to predict total sales based on our predictor variables, specifically random forest regression. Using 100 estimators, a maximum depth of 5, and maximum features of 2, we obtained a test R-squared value of 0.4585 and RMSE of 146165.3345.

Similarly to the linear regression model, we also run a 10-fold cross validation on this random forest regression model. This produced a test R-squared value of 0.4802 and RMSE of 144319.2389.

Finally, we run a grid search to optimize the parameters of the random forest regression. Running a 5-fold grid search with 10 iterations to find the optimal values of the number of estimators, maximum features, and maximum depth, we find that the optimal parameters for the model uses 294 estimators, ‘auto’ for the number of maximum features, and 55 as our maximum

depth. Running a random forest regression model with these parameters, we get a test R-squared value of 0.958 and RMSE of 40688.5701.

K-Nearest Neighbors: Due to the flexibility of the model, a K-Nearest Neighbors model was also run on our data. Fitting a KNN model with k = 5, the default number of nearest neighbors to take into account when calculating our predictions, we obtain a test R-squared value of 0.4591 and RMSE of 146075.549.

We also run an analysis to determine the optimal value of k for this model. Training KNN models with k values ranging from 1 to 75, we see that the highest test R-squared value is obtained when we have k as 3 or 5. Values below or above this range result in lower test accuracies.

Support Vector Regression: Finally, we also try a support vector regression model on our data. Using a C-value of 1.0, epsilon value of 0.2, and default kernel of rbf, we get a test R-squared value of -0.1439 and RMSE of 212440.5349. Tweaking these parameters and trying a C-value of 0.5, epsilon value of 0.8, and sigmoid, polynomial, and linear kernels do not seem to significantly improve or decrease our results at all. All of the 6 SVR model attempts resulted in an R-squared value below -0.1 and RMSE of over 200000.

Discussion

From the results obtained in the previous section, it seems that SVR and linear regression were not very successful in predicting total sales based on our predictor variables. KNN and random forest regression showed comparable performances, with the exception of our optimal random forest regression which had the highest R-squared value of all the models by far. This may

indicate that our data is not linearly separable, as both of the regression techniques that involve linear models performed very poorly on our dataset.

Both our linear regression models had low R-squared values of around 0.2, even when variables with high multicollinearity were removed. One notable finding is that after the first 5 predictor variables, the rest of the variables had identical and high p-values. This may be because there continues to be some collinearity between some of the brand details features, as they are identical for each row of the same brand. The features that were found to be significant predictors in both linear regression models were ARP, Total Units, Contains CBD, and Pax Filter.

All of the SVR model trials obtained extremely low R-squared values, all of which were negative. This indicates that the models perform more poorly than it would have if it had simply predicted the average total sales value every time. Even with different kernels, the model performed poorly, possibly indicating that there is too much noise in the data or data is too large or imbalanced.

Both our initial KNN model with k = 5 and our initial random forest regression model obtained R-squared values slightly below 0.5. This performance was much better than the results obtained from the linear regression and indicates that these two models are more suited for our data than linear regression. A more flexible, non-linear model seems to be more effective in aiding our predictions.

Our optimal random forest regression model using the optimal parameters found through the grid search obtained the highest R-squared value seen of over 0.95. This is likely due to the fact that the parameters chosen greatly increased the complexity of the model. With 294 trees in the forest, no limit on the number of maximum features to consider when looking for the best split, and 55 as our maximum tree depth, it is very likely that this model is severely overfitting to

our data. As we have data from all brands in both our training and testing sets, the model is likely using the brand details to determine what brand an instance belongs to and then looking at its previous financial history to predict future total sales. If we want to continue analyzing the same brands in the future, this may be an effective approach, but if we want to be able to incorporate new brands into our modeling, this approach will likely not work very well.

From these findings, it seems to be the case that Munchies should use a brand's previous financial history in order to most effectively predict its future sales and success. In addition to a brand's previous total sales data, its ARP and total units sold are also useful variables to take into account. The details about the products a brand sells is also significant in predicting its success, such as the types of products it sells and the effects of these products on the consumer. Features of brands such as whether or not they sell products that contain CBD or a pax filter have been shown to be significant in predicting the brand's success. Looking at features of products of brands that have been successful in the past would be useful in determining what products a new company should produce.

Some next steps for further analytic work could include trying to improve the poor performance of the linear regression model by incorporating interaction and polynomial terms. Additionally, it could be beneficial to switch the order of predictors between the various brand details features to see if the set of predictors that are significant changes when the order that they are used changes. After determining with more certainty that these predictors are indeed the ones that are significant in predicting total sales, reattempting SVR, KNN, and random forest regression with only these significant predictors may decrease noise in the data and result in a better fitting model. A Principal Component Analysis (PCA) could also be used to determine these significant features. Finally, running a k-fold cross validation on our high performing

optimal decision tree would help us determine whether or not overfitting occurred in this model, and the severity to which it did. Further modeling techniques should always be considered when continuing further work, such as bagging and boosting, ridge and lasso regression, and neural networks.