Projection of Future Sales for Booze 'R' Us

Lana Huynh: lmhuynh@calpolyledu

Jamie Luna: jluna28@calpoly.edu

Rachel Roggenkemper: rroggenk@calpoly.edu

James Rounthwaite: jrounthw@calpoly.edu

Sophia Chung: spchung@calpoly.edu

Objective

The problem proposed to us was to project sales in the coming year for Booze 'R' Us given data from all of the liquor stores in the state of Iowa, so that Booze 'R' Us can decide whether it will have enough income to expand operations. With this in mind, our goal was to create a model that would project sales in the coming year given a list of your stores' data, ultimately so you can decide whether you will have enough income to expand your operations. Due to the importance of reliability, we prioritized a model that would make accurate predictions. Additionally, a limitation of this model was finding a metric to represent liquor store sales, since that data is not available to us. In place, we are utilizing the total cost of a liquor order, which is the number of bottles multiplied by the state bottle retail.

Data Overview

The dataset we used was data on alcohol sales in Iowa, which was collected by the state government. This dataset contains the spirits purchase information of Iowa Class "E" liquor licensees by product and date of purchase from January 1, 2012 to current. A Class E liquor license, for grocery stores, liquor stores, convenience stores, etc., allows commercial establishments to sell liquor for off-premises consumption in original unopened containers. The dataset was acquired from the following website:
https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy.

The original dataset included twenty-four columns, such as the date of the order, the store number and name, information about the location of the store, the category of liquor ordered, the vendor of the brand of liquor ordered, and the cost and volume of the individual liquor products ordered. Since the original dataset contains over twenty-seven millions rows of individual product purchases, we only analyzed a subset of the original data.

The subset of data we utilized contained all the individual product purchases from Central City Liquor, Inc. in Des Moines, Iowa. We chose to only look at one store's data since this is ultimately how Booze 'R' Us would use this model as well. After we had this subset of data, we

began feature engineering by using the date column to create three new columns: month, day, year. Feature engineering is the process of building new columns that provide more insight into data. Therefore, we could use these as predictors for sales. To get an initial understanding of the total sales per month over time, in Figure 1 shown below, we have plotted the total sales to get an idea of the trend of sales over time. As we can see, the year of 2012 had a drastically greater number of sales per month compared to any other month and year. An additional point of interest from this figure is in the middle of 2016, there seem to be little to no sales. Thus, now we have an initial idea of the patterns in previous years before we start trying to predict future sales.

Figure 1: Sales by Month Over Time for Central City Liquor, Inc.
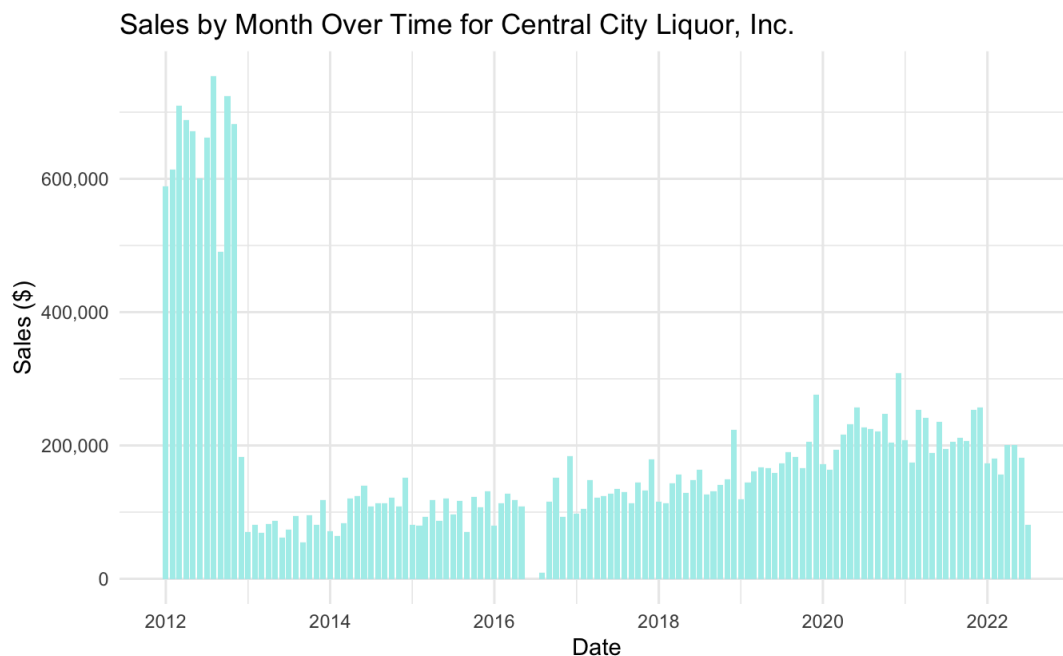
Additionally, since we are looking at the trend of total sales for month and year together in Figure 1, we were also interested in exploring the trends of month and year separately as well. To do this, we have plotted the total sales per month in Figure 2 and total sales per year in Figure 3, both shown below. As we can see in Figure 2, we see that April is the month with the highest sales. Additionally, Figure 2 displays that September, January, and February have the lowest sales of the year. A potential reason why January and February have low sales could be because of New Years' Resolutions. Many people could be making resolutions to not drink or drink less, which decreases the amount of sales, which ultimately means the liquor stores are buying less alcohol.

Figure 2: Sales by Month for Central City Liquor, Inc.

Sales by Month for Central City Liquor, Inc.

To move on to Figure 3, we see a similar trend as witnessed in Figure 1. The year of 2012 had a drastically greater number of sales by year compared to any other year. Ultimately, now we have an initial idea of the patterns by month before we start trying to predict future sales if breaking down sales by month and by year is something we want to consider in our model.

Figure 3: Sales by Year for Central City Liquor, Inc.

Sales by Year for Central City Liquor, Inc.

Next, we collapsed the categories of alcohol type into broader categories (amaretto, brandy, gin, liqueur, rum, schnapps, tequila, vodka, and whiskey) to have more accurate predictions, thus sacrificing a small amount of precision for higher accuracy. Having broader categories allows us to make more robust predictions. After creating these broader alcohol categories, we wanted to get an initial understanding of the sales for each alcohol category. To do this, in Figure 4 shown below, we have plotted the sales by each alcohol category. As we can see, whiskey by far has the greatest number of sales, with vodka and brandy coming in second and third, respectively. Thus, now we have an initial idea of the patterns for each alcohol category before we start trying to predict future sales if the alcohol category is something we want to consider in our model.

Figure 4: Sales by Alcohol Category for Central City Liquor, Inc.

Then, we filtered out columns that we deemed would not be relevant to our analysis, only keeping the cost of a liquor order, year, month, day, and our broader alcohol type. We then aggregated the purchases by day, counting how many times each category of alcohol was purchased each day and the total cost of alcohol for that day. Thus, each row represented a day of alcohol purchases by Central City Liquor, Inc. Lastly, we standardized the columns so we could achieve consistent model interpretations, especially since we used penalization such as Ridge when fitting our models.

Table 1: Sample of our Final Dataset

Total cost of liquor order	Month	Year	Amaretto	Brandy	Gin	Liqueur	Other	Rum	Schnapps	Tequila	Vodka	Whiskey
15922.40	-1.608	-1.744	-0.302	-0.819	0.678	0.869	-0.376	1.223	0.183	0.165	0.725	0.207
21295.84	-1.608	-1.744	5.682	0.822	2.883	3.797	1.253	4.071	6.128	2.802	4.663	3.489
74051.71	0.733	1.506	-0.302	-0.272	-0.425	0.524	1.525	-0.117	-0.383	1.484	-0.450	0.207
5684.83	0.733	1.506	-0.302	-0.090	-0.425	-0.682	1.796	-0.452	-0.100	-0.055	-0.104	0.718

To visualize what our final dataset looks like, Table 1 displays a sample of four rows of our data, including our column titles. Since we standardized the columns, the numbers don't make intuitive sense for interpretation. However, what the rows represent are on a particular day, for the alcohol category columns, it signifies how many (standardized) purchases of that specific alcohol category were made, and the total (standardized) cost of alcohol purchased for that day. Although each row represents a day of purchases, we decided to only use the standardized month and year in our analysis because we wanted to be able to predict monthly sales and did not think the day of purchase was relevant. During our model iteration we determined that the standardization was not entirely needed.

Model Selection Process

There are several models we considered using, but in the end, we tested out three different models to predict sales: ordinary least squares (OLS), ridge, and log-cosh. Additionally, we utilized scikit-learn as a baseline model to compare our results.

OLS regression is an optimization strategy that helps you find a straight line as close as possible to your data points in a linear regression model. In other words, OLS regression tries to match the data as closely as possible.

Ridge regression is used in cases where there are many parameters or predictors that affect the outcome. However, the main drawback of ridge regression is its high complexity.

Log-cosh also tries to match our data, but it is much more resilient to outliers than the other methods described.

We also needed to consider how to validate our data. We utilized k-fold cross validation and time-series cross validation.

In k-fold cross validation, the training data used in the model is split into k numbers of smaller sets, in which the model is then trained on k-1 folds of the training set. The remaining fold is then used as a validation set to evaluate the model.

In time-series cross validation, we pick a time point in our series and use everything that comes before it as training data and everything that comes after it as validation. This type of cross validation is more useful because we are using past data to predict future data. It does not make as much sense to use future data to predict past data.

In the end, we ended up choosing the OLS regression model, due to its simplicity in tandem with leveraging time-series cross validation to build the most optimal model.

Final Model Summary

After using OLS regression to build our model, we ended up with the model as the following:

Sales = 0.131 + 0.037*Year + 0.0776*Month - 0.036*Amaretto - 0.087*Brandy + 0.168*Gin + 0.058*Liqueur + 0.135*Rum + 0.238*Schnapps + 0.071*Tequila + 0.432*Vodka + 0.016*Whiskey

With this model, we ended up with an R^2 of 0.788 and a mean squared error (MSE) of 0.177. This means that our model explains about 78.8% of the variability in the data. Our MSE value of 0.177 is relatively low compared to other models that we fit.

Table 2. Coefficients of predictor variables
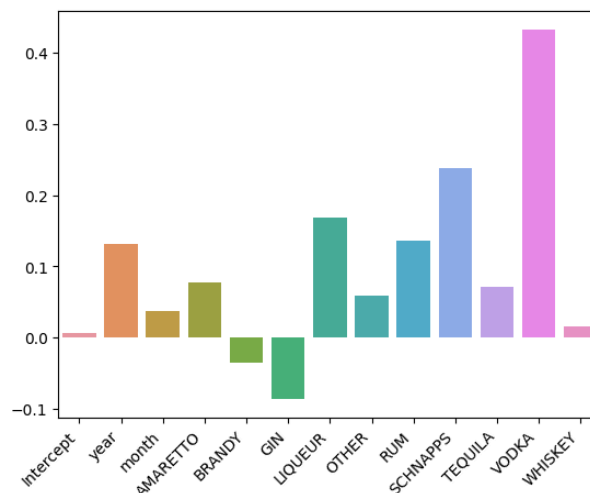
As we can see from Table 2, vodka has the largest impact on the cost of alcohol purchased for that day. As mentioned before, since we standardized these predictors, it is a little more difficult to get an intuitive interpretation with numbers; however, we can say that the more vodka the store buys compared to the other alcohols, the larger the sales will be.

Ethical Concerns and Recommendations

There are several ethical considerations that should be taken into account in this analysis. Firstly, it is important to ensure that the data used does not contain any personally identifiable information about any individual. All personal information should be anonymized or redacted.

In our model, the predictions we make are based on historical trends and data. It is important to note that these are forecasts, not guarantees. This model can only make educated guesses based on the past information it has been given. It is built on the assumption that past trends will continue into the future. As a result, there will inherently be a certain degree of uncertainty in our model. There are many factors that can influence sales, some which are beyond the scope of the data used (i.e. economic changes or natural disasters). Another limitation of our data that is critical to acknowledge is that we used the number of bottles multiplied by the state bottle retail as a proxy for sales which may not be entirely accurate. Overall, we made numerous assumptions about our predictors and response, so it is of utmost importance to ensure the interpretations are correct.

This model is a decision support tool. It is a valuable resource that can inform decisions, but those decisions should also be informed by expert judgment and real world business insights.

Takeaways and Conclusions

Our model will be a vital asset to your company, Booze 'R' Us as you look to expand your operations to new locations. With the structure behind our model, you will be able to plug in data for any store, or any number of stores, and see the projected sales for the next year. This will allow you to find your highest and lowest performing stores and to determine the viability of opening new locations. Our model also allows your company to see yearly trends. Some months have higher rates of selling alcohol than other months, and we wanted to ensure you would be able to see those trends so your stores can prepare accordingly. This gives a more granular view at a store's performance and can allow you to determine whether it is performing in a satisfactory manner on a monthly basis rather than only a yearly basis. Furthermore, our model also breaks sales into alcohol categories. By using our model, your company will be able to see which alcohol categories are the highest selling and when they are the highest selling so your stores can be prepared to capitalize on the opportunity to maximize their profits.

You will be able to know how accurate our model is and how well it is performing based on the R^2 value for each calculation. This value will be between 0 and 1. R^2 tells us how much of the

variation in total projected sales is explained by the model we provide. In other words, the closer the R^2 value is to 1, the more accurate it is at predicting the sales.