Predictive Analysis of Future Liquor Store Sales

Booze “R” Us

OCTOBER 2023

ISSUED BY
CALIFORNIA POLYTECHNIC STATE UNIVERSITY SAN LUIS OBISPO

CONSULTANTS
BRENDAN CALLENDER
JADYN ELLIS
KYLE LEW
SOREN PATEAU
ANAGHA SIKHA

Introduction

As your business partners, we were tasked with creating a way to help predict future sales for Booze "R" Us liquor stores. To accomplish this goal, we have created a machine learning model to predict future growth and decline in sales for individual liquor stores in Iowa. In this report, we will introduce the data used and data preparation steps taken to create this predictive model. Additionally, we will describe the model fitting process and validation metrics used to pick the optimal model. From there, we will provide a description of the final model as well as performance metrics assessing the effectiveness of the model. Finally, we will provide the major takeaways for how you can use our model to predict future sales for Booze 'R' Us liquor stores.

Background

The data for our model was collected from the Iowa Department of Revenue, alcoholic beverages subdivision. The data tracks the purchasing of Class "E" Liquor for individual liquor stores in Iowa. An individual observation represents a single transaction tracking the purchase of alcohol by a single liquor store. An observation contains identifiers for each unique store as well as information about a store's purchase, such as the type of alcohol purchased and the amount of alcohol purchased in dollars. We used only the data from 2018 to 2022 so our model was being trained on recent alcohol sale trends to better match today's market conditions.

One important consideration is that this dataset does not track the purchasing of alcohol by customers from an individual store. Instead it is tracking the purchasing of alcohol by stores to sell to their customers. However, the purchasing of alcohol by a store is likely tied to how much alcohol a store is selling. So we will be training our model to predict the amount of alcohol purchased by a store in dollars to predict growth and decline in sales. If our model is projecting a store to purchase more alcohol for the next year, we can also predict that store to sell more alcohol in the next year. Additionally, we can quantify the growth and decline in sales by comparing next year's prediction to the previous year and calculating the percent change.

Data Preparation

The first step in preparing the data was getting the data to be aggregated by year for each individual liquor store. So rather than a single observation containing one transaction for a liquor store, a single observation now contains the statistics relating to the yearly transactions for a liquor store. Some of the statistics we included were the total amount of alcohol purchased in dollars, the number of bottles purchased, and the mean price of bottles purchased for the store that year.

For each liquor store, we also wanted to include categorical factors such as the location of the store and the store name. To be able to better predict sales for liquor stores across cities in differently populated areas, we pulled Iowa city data[1] to gain access to the population numbers for cities in Iowa. We were then able to group cities into 1 of 6 population brackets. The population ranges are shown in *Table 1*. Additionally we created a variable to determine whether a liquor store was part of a large chain, small business or something in between. The breakdown of groups can be found in *Table 2*. We determined this by looking at the number of distinct stores for a business across Iowa and classifying the company into 1 of 3 groups.

Table 1. Population Bracket Classification

Bracket 1	Bracket 2	Bracket 3	Bracket 4	Bracket 5	Bracket 6
< 1k	1k - 10k	10k - 20k	20k - 50k	50k - 100k	> 100k

Table 2. Store Size Classification

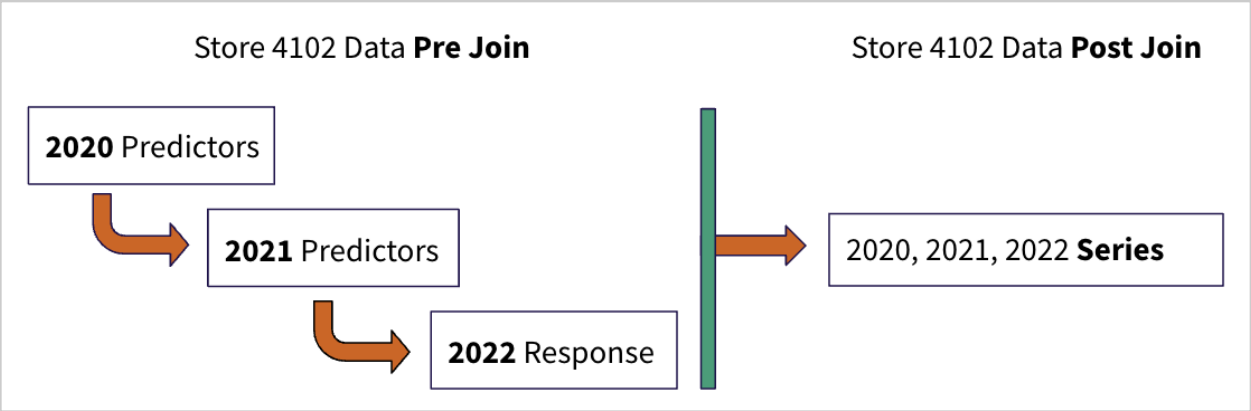
Small	Medium	Large
<10 Stores	10 - 30 Stores	30+ Stores

The next major step in preparing the data for the model was joining the data for different years across an individual store to create observations that were a series of data with sequential years. Because we wanted to use data from previous years to predict future

[1] Retrieved from https://www.iowa-demographics.com/cities_by_population

results, we needed to join the response for a single year with the data observations from the years before. A visual representation of the join we performed can be found below in *Figure 1*.

Figure 1. Data Joining Visualization



Model Selection

The model selection process was guided by one main motivation: Reduce the distance between what our model predicts, and what actually happened. The distance between a predicted value and the true value is called a prediction error. So throughout the model selection process, the model which minimized the collective prediction error was going to be the final model.

When calculating the prediction error for a model, we made sure to use cross validation. Cross validation is the process of training the model on one set of data and testing the effectiveness of the model on new data that the model has not seen before. This is very important for assessing the effectiveness of the model because it allows us to pick a model that can effectively predict on new data that it has not seen before like it will be doing for Booze 'R' Us liquor stores in the future. We used 5-fold cross validation for testing each model, meaning we trained the model on 80% of the data and tested it on the leftover 20%. Because there were 5 folds of data, we can pick 4 folds to train the model then use the last fold to test the model. This process can be repeated 5 times.

Multiple versions of models using different sets of predictors were considered in the search for the best predictive model. One of the first considerations was how many previous years of data to use in order to best predict the alcohol purchased for the next year. We considered

using the previous 3 years, 2 years, and 1 year of data. The next consideration was what variables to include for each year of data. For example, what data from the previous years is beneficial for predicting how much alcohol a store will purchase the next year? We also considered various categorical predictors relating to location and store size which were mentioned in the data preparation section.

We fit and tested these different models using ordinary least squares regression as well as ridge regression. Ridge regression is a form of penalized regression that prevents the model from being too geared toward fitting the training data. Ultimately all combinations for the different sets of variables and different types of regression models were judged the same with the goal of minimizing prediction error leading to a final model.

Final Model

The final model we came to for predicting the amount of alcohol purchased next year in Dollars for an individual liquor store has the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \epsilon$$

Where:

- Y is the predicted alcohol purchased, in USD, by an individual liquor store for the next year
- β_0, β_1, and β_2 are constants determined in the model fitting process
- X_1 is the alcohol purchased, in USD, by an individual liquor store 1 year prior to the predicted year
- X_2 is the alcohol purchased, in USD, by an individual liquor store 2 years prior to the predicted year
- ϵ is the left over prediction error for an individual observation. This is what we created our model to minimize

Fitting our model on training data produced the optimal estimates for β_0, β_1, and β_2 in order to best predict the alcohol purchased by a store over a year. These estimates were determined by fitting the data using ordinary least square regression.

Model Performance

$R^2 = 0.89$

The R^2 can be interpreted as follows: Our model is able to account for 89% of the variation in the amount of alcohol purchased by liquor stores for the next year. To put this in perspective, the standard deviation for alcohol purchased by individual liquor stores is $530,000. This means our model is extremely effective at cutting through the variation in alcohol purchased to determine whether or not a store will be purchasing more or less alcohol for a future year. Our model is able to do this by using the amount of alcohol purchased for the two previous years to better predict next year's result.

RMSE = $168,000

The root mean square error is a measure of prediction error that weighs larger errors more in its calculation. This relatively large RMSE is an indicator of our model tending to have large prediction errors for certain stores. Looking at the data, these large prediction errors tend to be associated with stores having major drops or jumps in alcohol spending. For example, one of the largest prediction errors in the data came from a liquor store which spent $3,000,000 on alcohol one year but then dropped to $800,000 for the year our model was predicting for. These types of observations are what is inflating the RMSE indicating our model is perhaps not effective for these types of unstable stores.

MAE = $65,000

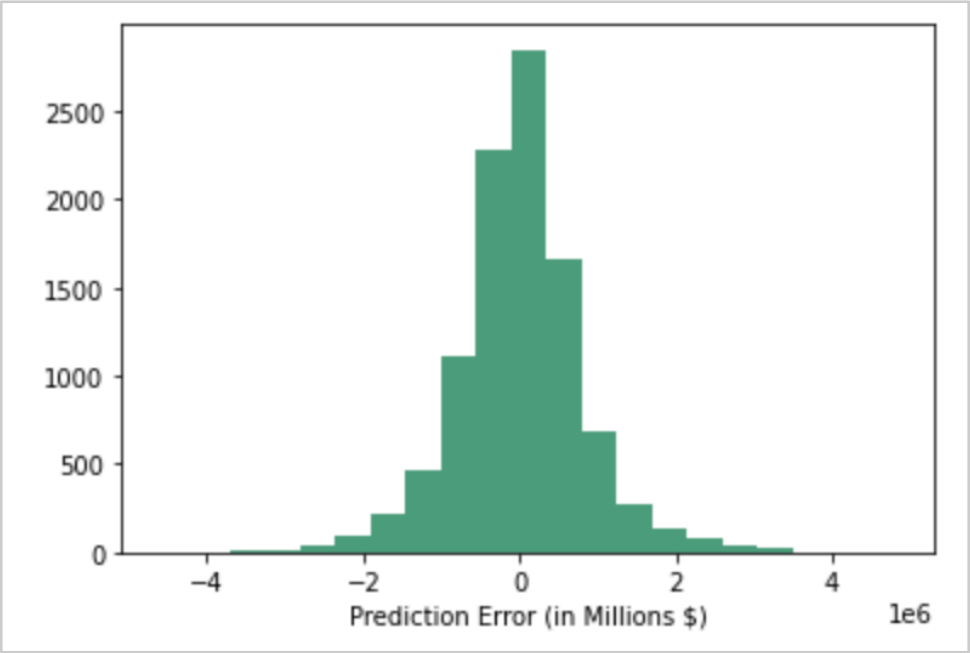
Mean absolute error is another measure of prediction error that equally weights all errors in its calculation. MAE can be interpreted as a typical prediction error for an individual prediction. So the predicted alcohol purchased for a liquor store for the next year is on average $65,000 dollars from the true value. Since MAE equally weighs all errors, it isn't influenced by outliers as much as RMSE. So a MAE of $65,000 is incredibly good considering the variation that exists within the amount of alcohol purchased by stores. (Recall Standard Deviation of $530,000)

Project Takeaways

The first takeaway is that our model is able to effectively project growth or decline in sales for individual liquor stores that follow more stable patterns for growth and decline. This is shown by MAE of $65,000 despite the presence of outliers in the data. Our model is able to cut through the variation in the amount of alcohol purchased by stores to deliver accurate and informed predictions. Recall final model R^2 of 0.89.

The last takeaway is that our model is also effective at predicting for groups of stores. Because our model is not prone to predicting values that are either too high or too low, our model can be used to project growth or decline for groupings of stores. *Figure 2* below shows this in practice. We selected 25 random liquor stores from the data. We were then able to sum our model predictions across these stores and compare this value to the true amount of alcohol purchased across the 25 stores. We then repeated this process many times and plotted each of the prediction errors. *Figure 2* shows that these predictions errors for groups of stores that are centered and most commonly around 0 meaning our predictions are accurate.

Figure 2. Distribution of Prediction Error Across Random Groups of 25 Stores



Ethical Considerations

We wanted to finish up by lastly addressing some ethical considerations around the use of our model.

Firstly, please refrain from using the model outside the scope of the data it was trained on. Our model was specifically catered for predicting alcohol sales in Iowa so please do not apply our model to liquor stores in different states which exist in very different markets. Also, consider the data used to create our model spanned from 2018 to 2022. So please be careful using our model far in the future when market conditions will likely be very different compared to the years in which our data spanned.

Additionally, please consider the ethics of using our model to make decisions on behalf of liquor stores without their knowledge. Because the data we used for our model did not come from individual stores, it would be possible to apply our model to the Booze “R” Us decision making process without the knowledge of the individual liquor stores it is about. Please inform stores of the use of our model and how the results of our model will affect them as a Booze “R” Us liquor store. Please guide the owners of your stores to this report for more information about the model.