Project Proposal:
Booze ‘R’ Us

October 13, 2023

Submitted By:
Andrew Kerr - adkerr@calpoly.edu
Bella M - imccarty@calpoly.edu
Erik Luu - eeluu@calpoly.edu
Martin Hsu - mshsu@calpoly.edu
Matteo S - mshafe01@calpoly.edu

Booze ‘R’ Us requests a method by which to predict future sales for growth purposes. We propose a machine learning approach fitting a multiple linear regression model to historical monthly sales data from Booze ‘R’ Us’s storefronts. The model will fit historical data and accurately predict future monthly sales for a given storefront. Furthermore, the features included will reflect factors that we find tend to drive or have the greatest impact on monthly liquor sales.

Data Collection

To analyze patterns in sales to determine if Booze ‘R’ Us should expand its operations, we plan to take subsets of the data of liquor sales in Iowa, filtering for only a single franchise at a time, such as “Casey’s General Store.” This franchise will consist of multiple storefronts, closely matching the business model of Booze ‘R’ Us’s multiple franchise locations. The years 2017-2020 will be selected to predict 2020 sales using 2017-2019 sales. The data shall be pulled directly from the Iowa state government’s data website, data.iowa.gov[1]. As illustrated in Figure 1, the profit consistently exhibits an upward trajectory over the examined period, hitting a maximum profit of $634,245 in the latest month. We seek to explain these trends in the data and apply the same methodology to Booze ‘R’ Us’s data.

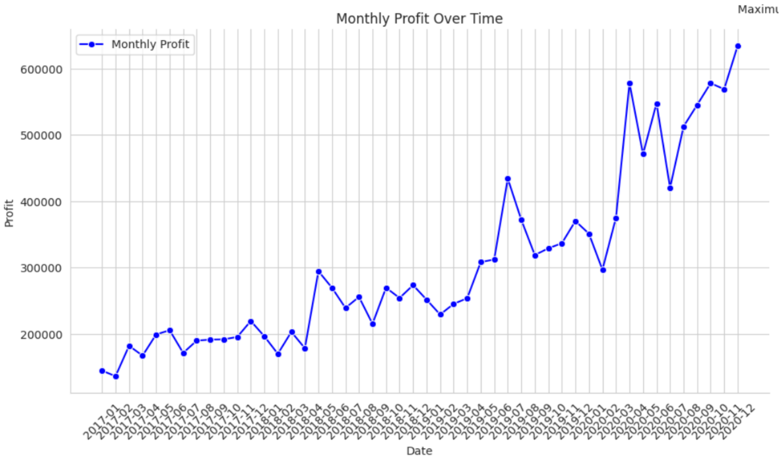

Figure 1: Monthly Profit Trend for Example Franchise (“Casey’s General Store”)

[1] https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy

We will categorize liquors into small and large sizes. Similarly, we will categorize the sale price per bottle into three categories: cheap, average price, and expensive. Using other existing columns, we plan to determine the year, month, number of full packs sold, and number of single bottles sold in excess of whole packs for each transaction. Additionally, we plan to clean the liquor's "Category Name" data to create easily accessible and generalizable categories of alcohol such as Gin, Rum, Tequila, Vodka, and more. This will allow an analysis of alcohol transactions for each type of alcohol at the storefront in the corresponding year/month. These features were chosen to provide more insights into what drives the most value in monthly sales to optimize a predictive model.

Modeling Process

The models we plan to fit are composed of combinations of the following variables:

Input Parameters
Size of liquors (<750ml; >=750ml)
Sale price (<$10; $10-$25; >$25)
Generalized categories of alcohol
Location of Store
Near Holiday (2 weeks)

Output Response
Monthly sales by storefront ($)

We will utilize K-fold cross-validation. This allows us to evaluate model performance more robustly by using different subsets of the data for training and validation. By testing the model across different folds (sets of data), we can better understand the variation in performance and sources of error in the model.

Linear regression is a simple yet powerful statistical model for understanding the relationship between variables. A key advantage of this model is interpretability. If stakeholders can see how the model directly uncovers relationships between sales and driving factors, they can make clear business decisions. By fitting a linear equation to the data, we can effectively describe how changes in one variable are associated with changes in another. They also make minimal assumptions about the underlying data and are computationally efficient, making them versatile.

Project Outcomes

In our final product, we aim to produce a custom system that continually develops a linear regression model to predict future liquor sales based on historical data. This model shall use a subset of the features specified in the feature engineering section. The linear regression modeling program shall allow the client to enter historical transaction data from the client's storefronts. It will estimate model parameters and validate results via a mathematically rigorous process and use them to project monthly sales predictions for storefronts, with a predetermined and reasonable standard of accuracy and precision.

The presentation of our findings will be characterized by visually informative, professionally formatted outputs, including charts, tables, and succinct summaries, ensuring accessibility to a wide range of stakeholders, from executives to the analytical team. To provide a broader context for our results, we will incorporate relevant references, enabling the client's organization to make decisions rooted in data-backed insights. This will allow the client to accurately prepare to gracefully meet the expectations of the greater public and boost efficiency.