Predicting Liquor Sales Using Iowa Liquor Data

Jenna Chan

October 2024

Contents

Introduction

This report presents the development of a predictive model aimed at forecasting monthly liquor sales, specifically the total amount of bottles sold for a type of liquor for a specific county within a given month. The model is designed to generalize across any state in the U.S. by incorporating external, non Iowa-specific factors such as unemployment, age demographics, personal income, and population data. The model provides a robust methodology for businesses to predict the total amount of bottles sold within a particular county for a given month for any given alcohol family type.

Data Description

The primary dataset used in this project is the Iowa liquor sales dataset[1], which contains spirits purchase information of Iowa Class "E" liquor licensees by product and date of purchase from January 1, 2012, to the present. Class "E" liquor licenses allow commercial establishments, such as grocery stores, liquor stores, and convenience stores, to sell liquor for off-premises consumption in original unopened containers. This dataset can be used to analyze total spirits sales in Iowa at the store level for individual products, aggregated by county and month.

In addition to this dataset, external socioeconomic datasets were integrated to better capture trends, including:

- Unemployment rates by county[2]
- Age demographics (median ages by gender)[3]
- Personal income per capita[4]
- County population data[5]

To ensure the model captures broader trends applicable to other states, we intentionally excluded Iowa-specific data, such as county-level information, as it would not accurately reflect the underlying factors driving variations in alcohol sales. Instead, we incorporated external datasets to identify fundamental patterns that extend beyond Iowa and are relevant across different markets in the U.S.

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

[2] https://data.iowa.gov/Economic-Statistics/Iowa-Local-Area-Unemployment-Statistics/tjnj-ed6z/about_data

[3] https://data.iowa.gov/Community-Demographics/Iowa-Median-Age-by-Sex-ACS-5-Year-Estimates-/jneb-h4gx/about_data

[4] https://data.iowa.gov/Economic-Statistics/Annual-Personal-Income-for-State-of-Iowa-by-County/st2k-2ti2/about_data

[5] https://data.iowa.gov/Community-Demographics/City-Population-in-Iowa-by-County-and-Year/y8va-rhk9/about_data

Data Preparation

The data was aggregated by county, month, and year to align with external datasets. Alcohol types with limited sales were grouped into broader categories (e.g., Whiskies, Vodkas, Rums) to improve prediction accuracy. Each observation in the final dataset represented the total sales for an alcohol category within a specific county and time period. After aggregation, the datasets were merged into one comprehensive dataset for further preparation and modeling.

To prepare the data for modeling, several steps were taken:

- Removal of irrelevant columns.
- Creation of interaction terms between alcohol types and seasons.
- Log transformations for skewed variables such as previous year sales and rolling averages.
- Polynomial features for key predictors (e.g., squared and cubic terms).
- Lag and rolling window features to capture time-dependent trends in sales.

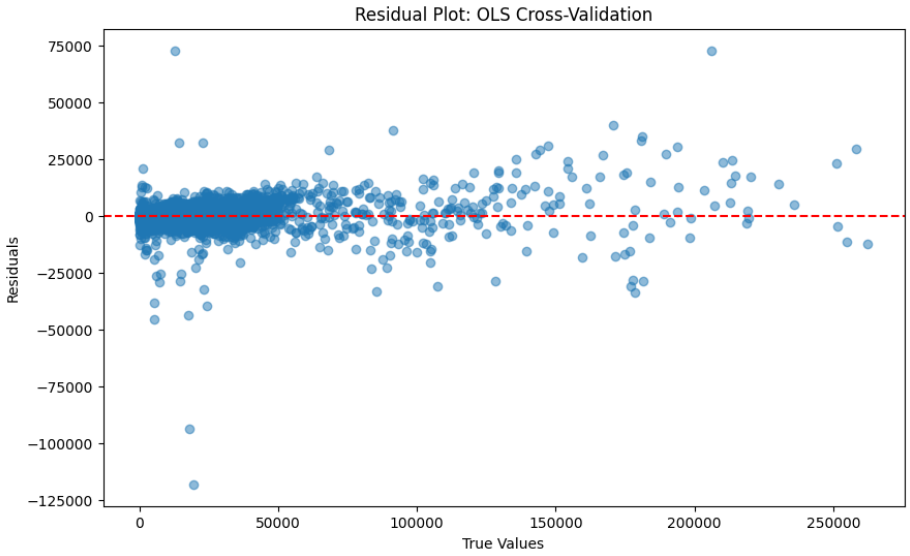

Figure 1: Heteroskedasticity in the Response Variable

Heteroskedasticity was addressed by applying a square root transformation to total bottle sales, as larger counties showed greater variability. Further solutions are suggested in the technical recommendations section.

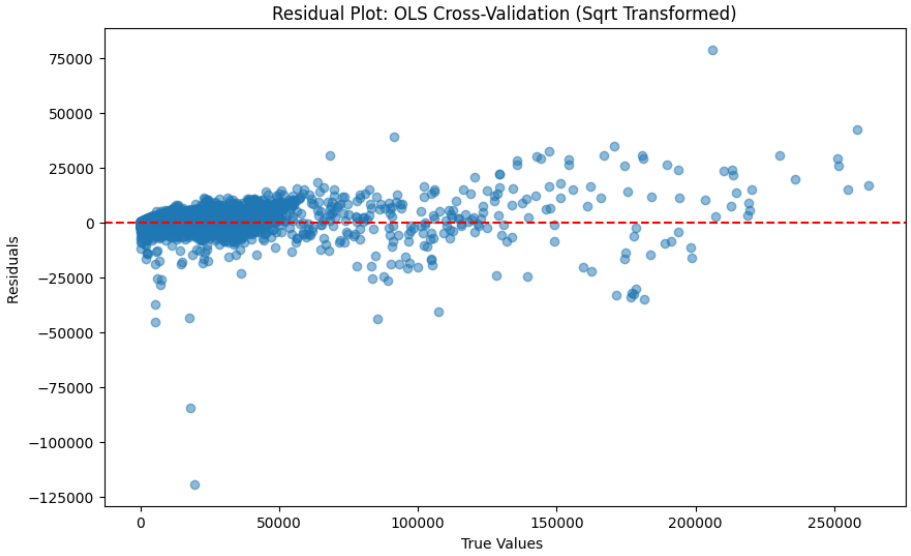

Figure 2: Heteroskedasticity in the transformed Response Variable

The square root transformation improved heteroskedasticity but did not fully resolve it. Alternatives like Box-Cox or gradient-boosted trees could be explored. Despite limitations, the transformation was retained, with further details provided in the technical suggestions section.

Model Selection & Validation Process

The model selection process focuses on developing a robust predictive model using multiple linear regression. Given that the goal is prediction, interpretability and concerns such as multicollinearity were deprioritized in favor of maximizing performance.

Evaluation Metric

Mean Squared Error (MSE) was chosen for its sensitivity to large errors, promoting better predictive performance. To ensure generalizability and avoid overfitting, 5-fold cross-validation was applied. Model selection involved comparing average MSE across identical splits using various regression techniques.

OLS Regression

The baseline model used Ordinary Least Squares (OLS) regression with all transformed features, evaluated via K-fold cross-validation. Three OLS models were compared: one with a non-transformed response, one square-root transformed, and one log-transformed.

Table 1: Comparison of Baseline MSE and Model MSEs

Model	**Baseline MSE**	**Model MSE**	**% Improvement**
OLS	12907263.1413	115576.0939	99.1%
OLS (Square-root)	11517904.3249	101956.0608	99.2%
OLS (Log)	12156091.974	142084.7072	98.8%

The square-root transformed OLS model was selected as the final model for its robust performance across identical data splits in all five folds. MSE was used over R-squared for its suitability in predictive models.

Model Summary and Final Model Justification

The final OLS model incorporated key features to capture trends in liquor sales, including previous year sales data, a 3-month rolling average, and their log transformations and polynomial terms to account for non-linear relationships. Time-dependent trends were addressed using 1-month and 3-month lagged sales, as well as rolling 3-month sales. County-level Personal Income was included to reflect economic factors influencing sales.

Dummy variables were used to represent different alcohol families (e.g., Gins, Vodkas, Whiskies) and seasonal effects (Spring, Summer, Winter), allowing the model to account for product type and seasonality. These features enabled the model to generalize effectively across counties and seasons.

Table 2: OLS Cross-Validation Results

Fold	RMSE	Mean Y	MedAE	Median Y
1	322.89	1198.38	59.26	284.00
2	323.59	1150.50	56.80	287.00
3	319.01	1148.84	57.57	289.00
4	313.96	1134.73	57.09	280.00
5	316.99	1147.16	55.78	279.00
Average	319.29	1155.92	57.30	283.80

The OLS cross-validation results demonstrate consistent performance across five folds, yielding an average RMSE of 319.29 bottles, indicating that predictions typically deviate by this amount. The model effectively predicts a wide range of sales despite county variability, with an average Median Absolute Error (MedAE) of 57.30, indicating many predictions are close to actual values. The lower median compared to mean sales suggests right-skewed data due to a few counties with significantly higher sales. Although the square-root transformation reduced some heteroskedasticity, the consistent RMSE highlights the need for further improvement, particularly for high-sales counties.

Ethical Concerns

Using county-level socioeconomic data in alcohol sales modeling raises ethical concerns about reinforcing stereotypes and unfairly targeting low-income communities. These models may inaccurately link lower income to higher alcohol consumption, ignoring broader social factors. Moreover, businesses could exploit this data to market alcohol aggressively to disadvantaged groups, increasing public health risks, similar to past practices by tobacco companies. Addressing these ethical considerations is essential to prevent predictive models from perpetuating inequality or stigmatizing vulnerable populations.

Recommendations

Current Model Usage

The current alcohol sales prediction model serves as a valuable tool for businesses aiming to forecast demand and optimize production strategies. By incorporating variables such as previous year sales, rolling averages, personal income, and seasonal indicators, businesses can accurately predict total bottles sold in specific counties. For instance, the model could project that 64,734 bottles of whiskey will be sold in a winter month for Armstrong County, Pennsylvania, allowing companies to anticipate demand shifts and adjust production schedules accordingly.

(a) Armstrong County Seal

(b) Armstrong County, Pennsylvania

Figure 3: Armstrong County Picture & Seal

Additionally, the model can inform marketing strategies by identifying peak sales seasons. With accurate forecasts, businesses can enhance their advertising efforts during periods of anticipated demand, optimizing market penetration based on the demand patterns identified. This approach enables companies to estimate their market share in various regions and adjust production volumes to align with predicted sales, ensuring they are prepared to meet demand.

For consumers looking to expand into Illinois for whiskey sales, the model can help identify counties that match their current market demographics and sales forecasts. By analyzing forecasted sales in counties like Cook, DuPage, and Lake, they can prioritize regions with similar income levels or consumption trends to their existing markets. However, a limitation exists in predicting large sales volumes in highly populated areas, such as Cook County, which may require further investigation and will be addressed in the technical recommendations section.

Further Technical Recommendations

Enhancing Feature Representation

The lack of predictive power for large alcohol sales in counties may stem from insufficient feature representation in the model. The current subset of features and external datasets may not capture the underlying trends necessary for accurate predictions. To address this, expanding the feature set could provide more insights into sales variability. Additional data sources, such as university enrollment, tourism and events, and public transportation, may help explain the dynamics in larger cities. With more time, sourcing these datasets could enhance our understanding of alcohol sales trends across various counties.

Further Transformations

Given more time, further transformations could enhance the model's performance by modifying both features and response variables. Adding more polynomial or interaction terms could help capture the non-linear relationships

present. Additionally, while square root and log transformations have been applied, there is potential for more effective transformations, such as a Box-Cox transformation, to better address the existing heteroskedasticity in the data.

Exploring Non-Linear Models

If earlier steps do not improve predictive performance, exploring non-linear machine learning models may be beneficial. These models can handle complex relationships that linear models struggle with.

Random Forests are a strong option, using multiple decision trees from various data subsets to capture intricate interactions while resisting overfitting. Gradient Boosting Machines (GBM) also excel at identifying subtle patterns by building models that focus on correcting prior errors. While Neural Networks offer flexibility for large datasets influenced by non-linear factors, they are generally a last resort due to their computational cost and often lower performance compared to Random Forests and XGBoost on tabular data.

Each non-linear model requires careful tuning to avoid overfitting. Given the uncertainty about whether the model's limitations stem from insufficient features or model choice, it may be more practical to implement these alternatives with existing data. This strategy supports a parsimonious model, which is advantageous, especially when comprehensive data sourcing is challenging across various counties.

Conclusion and Takeaways

This report presents a predictive model for forecasting monthly liquor sales across Iowa counties, using internal sales data and external socioeconomic factors. Utilizing Ordinary Least Squares (OLS) regression with a square root transformed response variable, the model effectively addresses heteroskedasticity and incorporates features like previous year sales, rolling averages, lagged values, and personal income.

While it performs well in most counties, challenges arise in predicting sales for larger counties with high variability, potentially due to insufficient features or linear regression limitations. Future enhancements could include expanding the feature set to incorporate tourism and public transportation data, as well as exploring non-linear models like Random Forests and Gradient Boosting Machines for improved accuracy.

In summary, the developed model is a robust tool for predicting liquor sales, providing valuable insights for businesses. Future iterations should refine the model to better address larger, high-variance counties and explore non-linear models for enhanced accuracy. Additionally, ethical considerations are crucial to prevent misuse of socioeconomic data that could exploit vulnerable populations or reinforce harmful stereotypes.

Appendix

List of Figures

List of Tables