University of Maryland, Baltimore County

Department of Data Science,

DATA 602: Introduction to Data Analysis and Machine Learning

Professor: Devin Fensterheim,

December 21st, 2022

Project Topic:

ANALYZE AND FORECAST IOWA LIQUOR SALES

By Team

JAMPANI AKHILTEJA (Campus ID: CK38174)

SOHAIL SHAIK (Campus ID: TH04570)

BUSINESS PROBLEM:

Iowa is an alcohol beverage control state, which means that the state has a monopoly on the wholesale distribution of alcohol throughout the state. Private retailers must buy alcohol from the state before selling it to individual customers. The state regulates the traffic of alcoholic beverages to protect the welfare, health, peace, morals, and safety of the state's citizens. Furthermore, since 2012, the state has provided monthly E license liquor sales data containing over 24 million transaction records. The Iowa state government wants to analyze its past data to reward its purchasers. Also, to forecast future outcomes.

BUSINESS OBJECTIVES:

Analyzing the past data:

- Examine and analyze the entire Iowa liquor sales data set.
- Find the best liquor category by volume sold, Scales (dollars), and profit (dollars)
- Find the monthly best vendor based on volume purchased, Scales (dollars), and profit(dollars). Iowa government would like to give some monthly rewards to its best vendors.
- Perform a time series analysis between essential features to identify the pattern.

Future forecast:

- Build a model to predict the total liquor volume sold by the Iowa government.
- Build a model to forecast total liquor sales (in dollars) from the Iowa government.
- Build a model to predict the total profit (in dollars) gained by vendors each time they purchased liquor from the Iowa government.
- Compare regression models and select the best model that fits the Iowa liquor dataset.
- Displaying the model's output and its results.

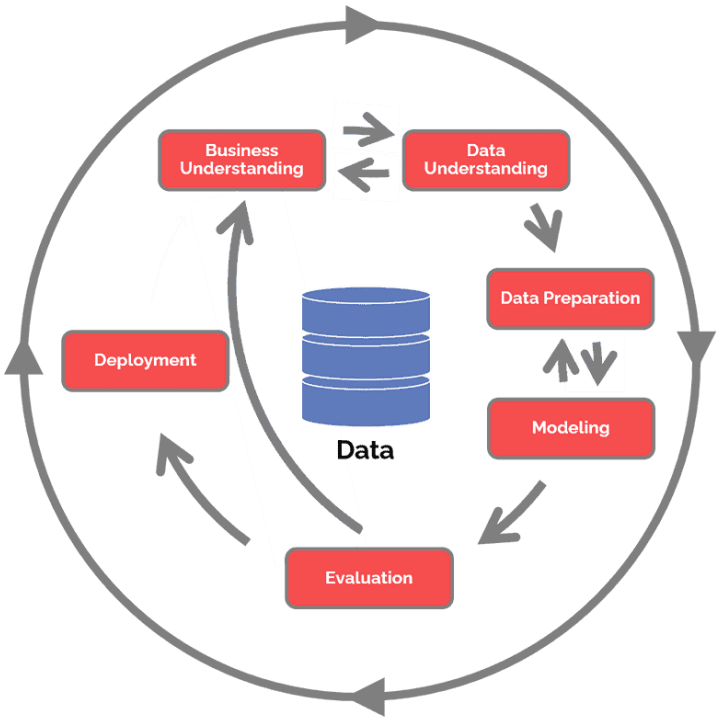
DATA SET:

The data is taken from Iowa's state-hosted open-source data portal. This portal contains over 24 million transactions, of which we are considering the last six months' data from 1st April 2022 to 1st October 2022, with 1.3 million transaction records. This data set contains 24 features.

URL: https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy/data

Approach:

For this project, we used the standard CRISP-DM approach.



Business understanding focuses on comprehending the project's goals and specifications.

Data understanding drives the focus to identify, collect, and analyze the data sets that can help us accomplish the project goals.

Data preparation is often referred to as “data munging,” which prepares the final data sets for modeling.

Modeling helps build and assess various models based on several different modeling techniques.

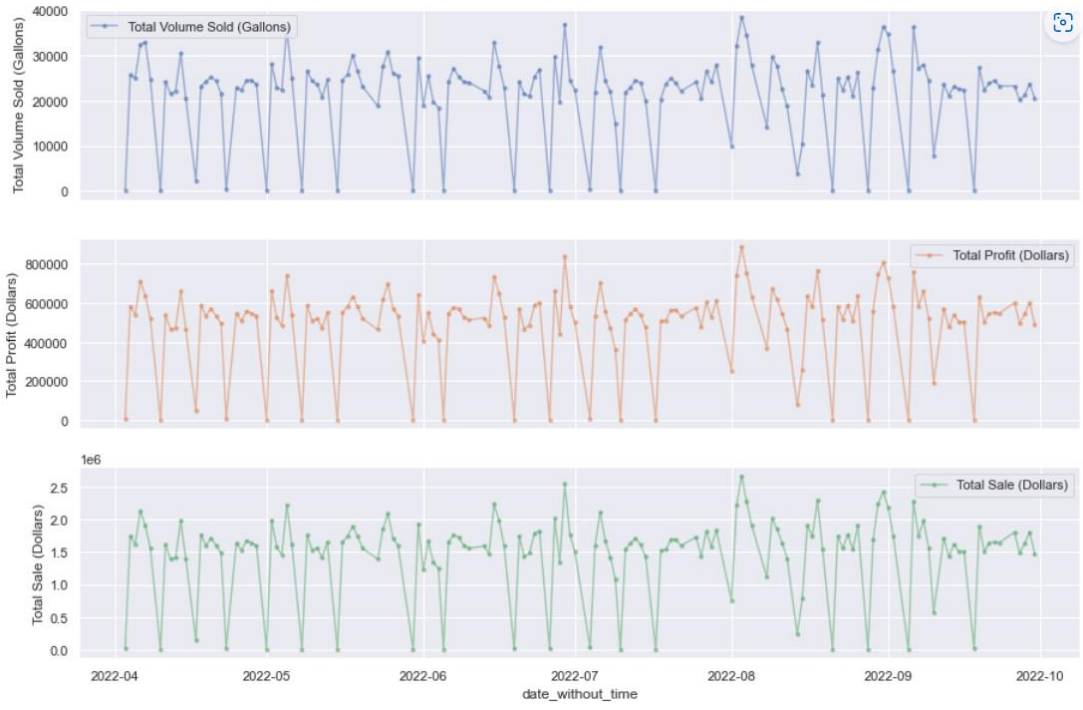
Evaluation helps us determine which model best meets the business and what to do next.

Deployment is organizing and presenting the results of data mining.

RESULTS:

Analyzing the past data:

- The total number of liquor categories in the Iowa government is 53.
- American vodkas are the best-sold liquor among all others, in 6 months, it sold 784091.48 gallons, a total scale of 33458563.66$ and a total profit by vendors is 11142611.56$
- Whiskey Liqueur 100 ml bottle is the top liquor bottle sold in the last six months.
- The total number of vendors for the Iowa government is 209.
- SAZERAC COMPANY INC has been the best vendor in the last six months based on the total volume sold.
- DIAGEO AMERICAS has been the best vendor in the last six months based on the total sales and profit the vendor gained by selling liquor.
- By doing a time series analysis, I observed that the Total Volume Sold (Gallons), Total Sales (Dollars), and Total Profit (Dollars) gained by vendors all these are following a similar pattern. And on Sundays, all of them reaching to its minima.



Future forecast:

Model 1 - Building a model to predict volume sold (Gallons).

- My target feature is Volume sold (Gallons).
- After cleaning and transforming the data in this model, I applied one hot encoding. As my categorical columns contain multiple unique values, this leads to multiple columns. So, in this model, I applied dimensionality reduction techniques (PCA).
- I used various models; below, I listed the r2 scores of those models.

Model Used	Mean R2 score of all cross-validations
Linear Regression	0.79029302
Lasso Regression	0.79026637
Ridge Regression	0.79029314
Elastic Net Regression	0.79029618
Decision Tree Regressor	0.97573807
Random Forest Regressor	0.96737079
Xgboost Regressor	0.97297516

- Among all model, Xgboost regressors perform very well. However, Random Forest Regressor and decision tree regressors are also performing well.
- After further analyzing Xgboost model, I achieved following results. This time I am checking predictive validity.
    - Train r2 score: 0.9984025391787233
    - Test r2 score: 0.9931746235641591
    - validation set r2 score: 0.9532716068879667
    - Mean absolute error (MAE) for test data set is 0.14480212302975223
    - Mean squared error (MSE) for test data set is 0.6170813013455062
    - Root means absolute error (RMSE) for test data set is 0.7855452255252439

Model 2 - Building a model to predict the price of sales(dollars) made by the Iowa government.

- My target feature is sales(dollars)
- After cleaning and transforming the data in this model, I applied one hot encoding. As my categorical columns contain multiple unique values, this leads to multiple columns. So, I am selecting the top 10 important features from each category column.
- I used various models; below, I listed the r2 scores of those models.

Model Used	Mean R2 score of all cross-validations
Linear Regression	0.6758622
Lasso Regression	0.6756529
Ridge Regression	0.6758711
Elastic Net Regression	0.6758717
Decision Tree Regressor	0.9480259
Random Forest Regressor	0.9700650
Xgboost Regressor	0.9737602

- Among all model, Xgboost regressors perform very well. However, Random Forest Regressor and decision tree regressors are also performing well.
- After further analyzing Xgboost model, I achieved following results. This time I am checking predictive validity.
    - Train r2 score: 0.9853017459545271
    - Test r2 score: 0.9970258563920585
    - validation set r2 score: 0.995787863911886
    - Mean absolute error (MAE) for test data set is 5.649905023195989
    - Mean squared error (MSE) for test data set is 767.295472845152
    - Root means absolute error (RMSE) for test data set is 27.70009878764247

Model 3 – Building a model to predict profit gain by the vendor(dollars).

- My target feature is profit gain (dollars).
- After cleaning and transforming the data in this model, I applied one hot encoding. As my categorical columns contain multiple unique values, this leads to multiple columns. So, in this model, I applied dimensionality reduction techniques (PCA).
- I used various models; below, I listed the r2 scores of those models.

Model Used	Mean R2 score of all cross-validations
Linear Regression	0.669890
Lasso Regression	0.669889
Ridge Regression	0.669810
Elastic Net Regression	0.669915
Decision Tree Regressor	0.909673
Random Forest Regressor	0.937095
Xgboost Regressor	0.910788

- Among all model, Random Forest regressors perform very well. However, Xgboost regressor and decision tree regressors are also performing well.
- After further analyzing Xgboost model, I achieved following results. This time I am checking predictive validity.
    - Train r2 score: 0.9896036179370348
    - Test r2 score: 0.9584824312901256
    - validation set r2 score: 0.9519112966729898
    - Mean absolute error (MAE) for test data set is 0.897515419685845
    - Mean squared error (MSE) for test data set is 1186.4073139754835
    - Root means absolute error (RMSE) for test data set is 34.44426387623175

CONCLUSION:

In this project, we achieved all our business objective. Now Iowa government can identify the best liquor category at the end of each month and reward the best vendors based on total volume purchased, total scales (dollars), and total profit (dollars). The government is also able to predict future outcomes of total liquor sales (in dollars), liquor volume sold, and total profit (in dollars).