Iowa Liquor Retail Sales

Stefan Popov, B. Sc.

Abstract. This work presents the capstone project work for the course "Applied Data Science Capstone" as part of the IBM Data Science Professional Certificate on the Coursera platform. In it, I present an analysis of a publicly available data set on liquor retails sales in the state of Iowa. Please refer to the notebook for more technical information on what has been done.

1 Introduction / Business problem

I will perform an analysis of a publicly available data set on liquor retail sales in the state of Iowa, U.S., to try and gain some insights on some of the most common retail problems. Retailers often want to forecast their demand in order to predict when and how much goods should be stocked. Analysis of this data set can provide insights into the habits of alcohol consumption in Iowa. They might provide liquor stores with more accurate information about their demand and help them better control their stocks; and they might help distributors optimize their delivery routes by predicting the demand days ahead and thus preventing over- or under-stocking of different types of alcohols (supply-chain problems).

For this project, I will assume the role of owner of a liquor store in Iowa, and it is time to reorder inventory. Before I do so, I would like to get better idea of which products are most popular so that I order the right amounts. Using the visualisation tool on the Iowa government page, I had generated a plot of sales made by store name, and have decided to investigate the sales of the store that reported the most sales for the period for which the data is available, and that is the store '"Hy-Vee #3 / BDI / Des Moines"' in the 'Polk' country, address: '"3221 SE 14TH ST"' (geographical point: 'POINT(41.554101 -93.596754)') which reported over '160K' sales between '2012-2020'. The plot of Figure 1 illustrates the top X stores by total sales in 2012-2020.

2 Data

The data is publicly available both on the Iowa's government web-page at *Iowa Liquor Retail Sales | data.iowa.gov* (2020) and is included in Google's publicly released data sets on the Google Cloud Platform at *Iowa Liquor Retail Sales | Marketplace* (2020). On the former page it is freely available to download in various formats (csv, Excel, etc.) while on the latter it is integrated in BigQuery - Google's analytic warehouse tool where users can process up to 1 TB of data free of charge.

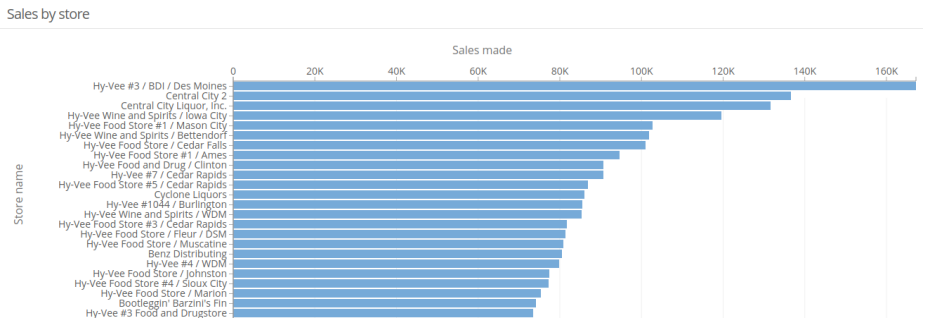

Fig. 1: Total sales by store

This data set contains every wholesale purchase of liquor in the State of Iowa by retailers for sale to individuals since January 1, 2012. The State of Iowa controls the wholesale distribution of liquor intended for retail sale, which means this data set offers a complete view of retail liquor sales in the entire state. The data set contains every wholesale order of liquor by all grocery stores, liquor stores, convenience stores, etc., with details about the store and location, the exact liquor brand and size, and the number of bottles ordered.

The data set is updated monthly, usually on the first day of the month, therefore it is up to date. Attributes include information about:

- Vendors: name and number
- Items: number, description, bottle volume, category
- Order: store cost, retail cost, number of bottles sold, volume sold (litters and gallons),

The original data set contains 24 attributes much of whom are of no interest to us, therefore I will project the initial set to only three columns: ‘Date‘ - day the sale was made, ‘Item Description‘ - which liquor brand was bought, and ‘Bottles Sold‘ - the target variable we are trying to predict.

To enhance our predictive power, we are going to incorporate weather data into our analysis. NOAA’s Global Historical Climatology Network (GHCN) contains complex weather data that might help us. It provides weather observation from more than 11K stations around the world, and for some of its stations the data goes back to year 1763! In our scenario, we would like to obtain some weather parameters from the store’s nearest weather stations. First, we will obtain a list of weather stations that are in 50 km radius from the store. Then, we will compare that list with the weather data for the period 2012-2020, so that we are sure that the stations are still operating. There are 97 weather stations within 50km radius of the store. Next, we will filter the weather stations active in period 2012-2020 who have recorded the:

- Maximum temperature for the day (TMAX flag)
- Minimum temperature for the day (TMIN flag)

– Amount of snow fall (SNOW flag)
– Amount of precipitation (PRCP flag)

This information is contained in the ‘inventory‘ data set. For each station, it records the ‘first‘ and ‘last‘ year of measuring each ‘element‘. We would require that each nearby station has records for all of these four elements and that their ‘first‘ year is less than ‘2012‘, while the latest year is ‘2020‘. There are seven weather stations in the store's vicinity that record *all* four weather attributes for the period 2012-2020.

Et, voilà! We finally have our closest weather station. It is located only 5.2 km from the store, at ‘POINT(41.5339, -93.6531)‘ with code ‘USW00014933‘.

3 Methodology

In this section we will prepare the data for training a machine learning model and perform some feature engineering to increase the model's predictive performance. Forecasting models use data from the past to make predictions for the future. The data in our model will be historical sales and weather statistics, and we will use it to predict sales in the future.

More specifically, we are trying to predict the ‘Bottles Sold‘ variable for each brand of liquor - ‘Item Description‘ for a given month (‘Year-Month‘ variable), on the first of that month. For example, we take the data as of January 2020 and use it to predict the sales for February 2020, data as of February 2020 to predict March 2020, etc.

The feature that we are going to use to predict sales are the following:

– ‘Month‘ - the month of the year we are trying to predict
– ‘Month Sales Lag X‘ - how many bottles of the brand of liquor were sold in the previous month (X = 1), the month before that (X = 2), the month before that (X = 3), and one year ago (X = 12)
– ‘PRCP Lag X‘ - the average monthly precipitation last month (X = 1) and one year ago (X = 12)
– ‘SNOW Lag X‘ - the average monthly snowfall last month (X = 1) and one year ago (X = 12)
– ‘TMIN Lag X‘ - the average monthly minimum temperature last month (X = 1) and one year ago (X = 12)
– ‘TMAX Lag X‘ - the average monthly maximum temperature last month (X = 1) and one year ago (X = 12)

A few things to consider here.

1. The features we are trying to construct require data up to 12 months prior to make prediction for a given month. Our data set starts in 2012, but since we need 12 months of data to start predictions, we can not train on 2012 data nor have it included into our training data set.

2. Some of liquor brands are rarely stocked, probably only when the inventory gets too low. These are unpopular brands and we will not model them. Therefore, they too will be removed from the data set. We will set up a threshold parameter ‘popularity‘, and all liquor brands must have at least ‘popularity‘ bottles sold by our store since 2012.

We will use this data set to predict our monthly sales. We will build three regression models: linear, ridge and lasso. We will use the R-squared value as an indicator of how good our model is.

In the example laid out here, we are trying to predict the sales for each month in 2020. For that reason, we will train our model on data up to January 2020. In general, we want to use as much data from the past and up to the period for which we are trying to predict. Table 1 summarizes the information about the R-squared values.

Month	**Linear**	**Lasso**	**Ridge**
2020-01	0.8731	0.8737	0.8731
2020-02	0.8461	0.8462	0.8461
2020-03	0.7181	0.7180	0.7181
2020-04	-0.6544	-0.6518	-0.6544
2020-05	-1.3093	-1.3200	1.3093
2020-06	0.6268	0.6269	0.6268
2020-07	0.7392	0.7392	0.7392
2020-08	0.7848	0.7848	0.7848

Table 1: R-squared values for each model for each month

We can observe that Linear and Ridge regression perform exactly the same, the differences between them are less than 10e-4.

In general, we would have expected the R-squared value to start high and then continuously decrease as the months progress. This is because the sales are highly correlated with their immediate previous values, and because our model would not be trained on data immediately preceding it, therefore, it should perform worse.

That is true for our case here, but for April and May 2020 the R-squared values deviate substantially. This is of course because of the COVID-19 pandemic. The plot in Figure 2 below plots the monthly sales for the entire data set.

We clearly see the steep drop down at the right end of the plot corresponding to the COVID-19 period. This kind of deviation could not have been predicted with any machine learning model, due to the fact that it is cause by an unpredictable ”higher-power”. Literally no one in the world expected something like that, so we would not hold it against simple regression models for not being able to predict it.

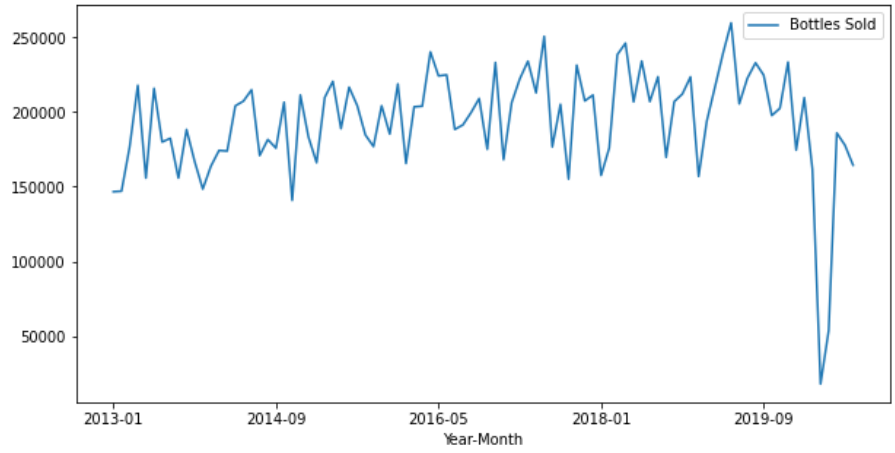

Fig. 2: Monthly sales

4 Results and discussion

Our analysis show how relatively easy we can build a machine learning model to help us with our decisions for restocking our store inventory.

First, we identified our business problem. We tried to predict our monthly sales by using historical sales and weather data. The historical sales data is our own, and there would be no need to query an external source for it; a store has records of its sales. The weather data was freely available for download from the NOAA's Global Historical Climatology Network (GHCN).

GHCN maintains a vastly complex database of weather information, and we needed quite some time to be able find the relevant data for us. We had to find the weather stations that are in the vicinity of the store, which are still operating in 2020 (some of stations had data going back to the 1700s!) and are recording the weather information that we need, that is temperature, snowfall and precipitation.

Once we found the station closest to us, we queried its data and merged it with our historical sales to obtain our final data set that is ready for modeling.

Then, we apply three regression models: linear, Ridge and Lasso. All models exhibit similar predictive performance, and we recommend using the linear regression for its simplicity.

The models perform rather good, with their R-squared values ranging around 0.88.

As everything in the world, the model's performance suffered substantially when trying to predict sales for the period of the COVID-19 pandemic in the U.S. (April and May 2020). The sales for this period drop very low, something that has not been observed in our data set and could not have been modeled by any other machine model.

5 Conclusion

To conclude this work, I will say a few words about the application of this project.

Predict store sales based on historical data is very common and known retail problem and there are already some very known solutions to it. In general, a store owner could just export his data to a machine learning expert, and have him return his predictions. He would then use this data to make decision on restocking his inventory.

This notebook served more as an education tool for me, than as a real-world project that can be applied. I am not saying that this does not have to potential or that it can not be implemented in an actual case study, but I am having a hard time believing that a store owner from Iowa would trust my work.

Maybe I will show it to some local stores in my country (Macedonia), and see if there is any interest to see this work being applied here.

Bibliography

Iowa Liquor Retail Sales | data.iowa.gov (2020). Last accessed 06 June 2020.
URL: *https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy*

Iowa Liquor Retail Sales | Marketplace (2020). Last accessed 06 June 2020.
URL: *https://console.cloud.google.com/marketplace/details/iowa-department-of-commerce/iowa-liquor-sales?filter=solution-type%3Adataset—&filter=category%3Aanalytics—&id=18f0a495-8e20-4124-a349-0c4c167b60ab*