Forecasting Daily Beer Sales in Mexico

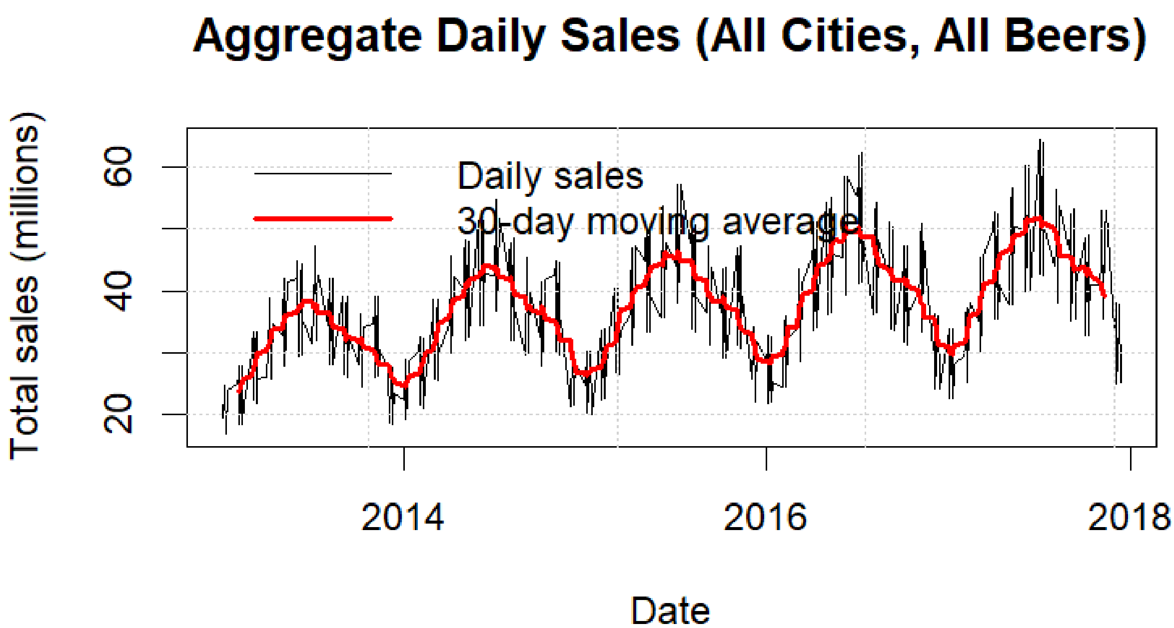
Evan Gattoni

Abstract

This project focuses on analyzing daily beer sales across 10 major Mexican cities. After aggregating all sales and products sold across the locations, several time-series models were used and evaluated to assess the best way to forecast sales patterns. The data exhibits high frequency variation, clear seasonal cycles, and high sales numbers, so to stabilize variance and make comparing these models easier, the series was log transformed for modeling. Five forecasting models were compared using rolling prediction errors, ARIMA (1,0,1), simple exponent smoothing, local level state-space Kalman filter model, seasonal ARIMA, and DLM were chosen. The metrics chosen to evaluate these models were mean squared error (MSE) and mean absolute error (MAE), over 355 forecasted points. The basic results were that simpler models, like ARIMA and exponential smoothing, tended to perform with the lowest errors, with similar performance. Seasonal differencing did not help to improve forecast accuracy, due to a large seasonal period and heavy noise in the daily observations. The results of testing these models would suggest that forecasting this dataset with recent behavior is more effective than deep seasonal patterns, as seen from the effectiveness of low parameter, fast updating models in this project.

Data Description

The objective of this project is to predict daily beer sales in Mexico for a certain brand. Time series models are used to forecast sales, and we can judge the performance of each model using back testing with the operational data. The dataset covers four years of sales from 2013 to 2017, and it contains 720 daily sales observations, representing the sales of 50 different products over 10 Mexican cities. Because of the large amount of noise present with all of the different products and cities, the first step taken for data cleaning was to aggregate all sales across city and product, to get a total measure of beer sales across the country. To stabilize the large variance and make models more easily comparable, the data was analyzed through a log scale transformation. The aggregate beer sales now had a daily total sales number, this is the statistic that is of interest for forecasting, and there are certain characteristics that can be observed. First, there is strong seasonality present in the data, with a peak in the summer of each year, where beer sales greatly exceeded the winter months. There was also significant volatility when looking at the day-by-day data, since one day is a short period, sales fluctuate greatly from one day to the next. Finally, there is an overall increasing trend in sales over the years, reflecting national demand for beer rising over the four years, this could be explained by population growth in Mexico. These characteristics are drawn from the plot below, and a 30-day moving average was added to highlight the overall trends in the data and remove some of the volatility so that we can make generalizations for the dataset.



Model Descriptions

The five models were evaluated using a rolling one step forecasting design, designed to simulate how forecasts are generated. The model was trained based on the first 365 days of sales, and the following days are forecasted with only the information prior to that day. Errors were computed on the log scale using two metrics, MSE and MAE. MSE is useful for evaluating overall model accuracy and emphasizes large errors by squaring the values. MAE is essentially average absolute deviation and is not quite as sensitive to larger errors as MSE. There were five models used in this project to forecast sales:

- ARIMA model is a basic moving average model that focuses on recent autocorrelation patterns, due to the nature of the log transformed series, it was fitted without differencing.

- Simple exponential smoothing model produces forecasts by weighting recent observations greater than older data points, this is useful for this dataset due to the large amount of noise present, and it provides a simpler estimate.
- Local level state-space model uses a Kalman filter based estimate, where observations come from a latent mean that is continuously updated over time. Due to the nature of the running mean, the model will naturally produce smoother estimates of demand the more time passes.
- Seasonal ARIMA is just an ARIMA model with a difference period of 365 days to capture trends from year to year, comparing the sales of the day from a current year to the same day one year prior.
- Discounted dynamic linear model (DLM) is a Bayesian filter model where the discount factor controls how the latent level adapts to new data. There were several discount factors tested to see how model flexibility changes accuracy of the prediction.

Results

Model	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
ARIMA	.02309	.1223
Exponential Smoothing	.02320	.1211
Local Level Kalman	.02325	.1220
SARIMA	.04241	.1473
DLM (δ = .8)	.02349	.1236
DLM (δ = .9)	.02694	.1321
DLM (δ = .95)	.03500	.1523
DLM (δ = .99)	.05055	.1846

The key findings from the results are that the simplest models produced the lowest errors. The ARIMA, exponential smoothing, and local level models all performed very similarly, with an MSE around .023, these errors were the lowest of any model. The seasonal ARIMA model performed much worse than expected, due to a lot of daily noise, so the long season differencing was very volatile and did not give the model a good picture to predict. I think that if the sales were grouped by month rather than by day, this model would have performed much better. DLM shows that as the discount factor approaches 1, the model becomes less accurate. The DLM with δ = .8 performed almost as good as the simple models, which shows that this data needs a model that can quickly adapt to new information.

Conclusions

The strongest takeaway from the model evaluation is that for this dataset, short term forecasting models were more effective at reducing error. Daily sales are very volatile, and due to this, longer term patterns were harder for the model to recognize and predict. The highly adaptive models had better predictions, and these results show that simplicity is often a good choice when forecasting. The presence of a lot of noise negatively affects the accuracy of the longer, more cumulative type of predictive models. Decisions that would be made off these findings would be the supplying of beer across Mexico based on the time of year. For example, breweries should be aware of high demand volume days and make sure that there is enough supply of beer to satisfy demand. While this project deals with aggregate sales across Mexico, suppliers could deviate the models by city, to get a more detailed picture of what cities will need supply of what products. This would allow companies to anticipate where they will need to produce and deliver beer. These models allow for both short term and long term planning for logistics and production for a beer producer.