Forecasting Coca-Cola Bag Orders Using Social Media

By

Elijah Ampo, Ruohan Zhou, and Yingkun Zhu

Supervisor: Arnab Bose

A Capstone Project

Submitted to the University of Chicago in partial fulfillment
of the requirements for the degree of

Master of Science in Analytics

Graham School of Continuing Liberal and Professional Studies

May, 2019

The Capstone Project committee for Elijah Ampo, Ruohan Zhou, and Yingkun Zhu
Certifies that this is the approved version of the following capstone project report:

Forecasting Coca-Cola Bag Orders Using Social Media

Approved by Supervising Committee:

_____   _____

Arnab Bose   Dr. Sema Barlas

Abstract

Scholle IPN is a global manufacturing company that has experienced variability in their sales forecasts. In this project, we demonstrate how Scholle IPN can leverage social media data to predict orders from their clients. We introduce dimension reduction methods to account for the high dimensional nature of social media data. In addition, an ensemble model approach to sales forecasting is used to generate the best sales forecasting model.

Keywords: Time Series, Machine Learning, sARIMA, Regression with ARIMA Errors, XGBoost, Long Short-Term Memory (LSTM), Ensemble, Social Media, Linear Regression.

Executive Summary

Manufacturing companies need to forecast the demand of their products ahead of time, so that raw material preparation, production, and shipping schedule can be arranged accordingly. Failing to predict the future demand can result in waste due to over-producing, business opportunities lose because of under-stocking or unable to deliver in a timely manner, etc. In an era where social media dominates people's life, and data analytics are growing more and more powerful when it comes to drawing insights to enhance business actions, Scholle IPN tasked us to test if Social Media brings any predictive power to forecast the future Coca-Cola bag sales. With this problem stated, we took this task on as a trial to examine whether or not people's discussion and engagement on Social Media has any impact on the consumption of the Coke product in real life.

In order to achieve that, we started with gathering external data from Social Media, and combined with the internal data provided by Scholle IPN. After preliminary exploratory on the data - both internal and external, we built traditional Regression with ARIMA errors models, then more advanced Machine Learning models, namely, XGBoost and LSTM RNN. XGBoost using the differenced social media variables gave the best predictions (lowest sMAPE score) among all the individual models while still has a decent level of interpretability for us to explain the model. Since we obtained the various predicting results from the single predicting models, we also stacked these individual models together and created three ensemble models - mean average, Linear Regression, and Random Forest. Random Forest ensemble model achieved the best outcomes, with the lowest error score. Key findings from this project including the following:

- Social Media contents, specifically, online discussions and engagement level of certain brands and ad campaigns have helped Machine Learning models to forecast the future bag sales.
- Different Social Media variables have different importances when it comes to modeling. For instance, the jobs related conversation online demonstrated higher level of importance than other brands-related conversations in our XGBoost model feature importance.
- Reducing the dimensionality when handling a large number of features improves the predictive accuracy; Cross Correlation check, Principal Component Analysis are also vital before modeling in our case.
- Ensemble modeling is a great way to bring the individual models together as inputs, and establish a better model with stronger predictive capacity. The output from the ensemble model has been proven to be the best results out of all. We recommend

Scholle IPN to continue monitoring the Social Media conversations, and follow our practice of adding the social media variables to the internal historical sales data. For future work, we also recommend to examine the year-to-year comparison between the forecasting bag sales with the actual bag sales.

Table of Contents

List of Figures

List of Tables

Introduction

Scholle IPN is a global manufacturing company based in Northlake, IL, with products focused primarily in bag-in-box packaging. The company is a pioneer in its industry by implementing a combination of qualitative observations and quantitative analyses in forecasting their products' sales. However, variability in these sales forecasts present challenges for Scholle IPN in raw material preparation, operational efficiency, and asset management.

Problem Statement

In this project, we provide a forecasting solution to Scholle IPN using social media data. This project primarily focuses on one of Scholle IPN's main clients, Coca-Cola. Coca-Cola uses Scholle IPN's state-of-the art bags to store beverage products at quick service restaurant (QSR) partners worldwide. Since 2014, Coca-Cola has accounted for 95.68% of Scholle's syrup-related bag order shipments, so inaccurate forecasts of future orders could result in operational inefficiencies. Minimizing these operational inefficiencies is important to maintain Scholle's partnership with Coca-Cola. In order to solve this business problem, we examine whether we can substantially improve Scholle IPN's Coca-Cola demand forecasts by using social media as the primary variable.

Research Purpose

The purpose of this research is to forecast Coca-Cola bag orders by utilizing social media data. When customer express their opinions in social media, businesses like Scholle IPN can gain valuable insights that can inform business decisions. For example, if a Mc-Donald's promotion is generating discussion posts online, then Scholle IPN can potentially expect an increase in bag orders from Coca-Cola. In order to make these online discussions actionable, we must first take into account additional considerations. First, since social media posts are usually in the form of text, we explore methods to convert text to numerical data. Second, we explore different ways to reduce the dimension of social media variables to account for its high dimensionality. And finally, we use these social media variables to forecast Coca-Cola bag orders using an ensemble approach. Below is a list of research objectives for this project:

- Convert text-based social media data to numerical data using natural language processing.
- Reduce the dimension of social media variables using different methods.
- Forecast Coca-Cola bag orders using an ensemble approach.

Variables and Scope

The scope of this forecasting project is limited to predicting the future monthly bag orders for Coca-Cola in the United States and Canada. All variables will be aggregated or averaged at the monthly level. The forecasting window for this project will be 18 months for all models to accommodate Scholle's business needs. The main social media variables used to predict Coca-Cola bag orders were collected from Twitter and Google Trends. For Twitter, we focus our project on the following variables: tweet text, number of likes, number of retweets, and number of replies. For Google Trends, monthly trend values for selected topics are extracted at the monthly level. Additional data retrieval rules are applied to ensure that Twitter and Google Trends data are from the United States and Canada (see Appendix A).

Background

Social Media Variables

For this forecasting project, we focus on collecting social media data on relevant topics. These relevant topics are Coca-Cola, Pepsi, McDonald's, Taco Bell, and "jobs". Coca-Cola is selected because it is their product's demand that Scholle IPN is interested in forecasting. Pepsi is selected due to its position as the main competitor for Coca-Cola. Meanwhile, McDonald's and Taco Bell are chosen because they are the top quick service restaurant partners for Coca-Cola and Pepsi based on 2018 annual sales revenue. The topic 'jobs' is selected to gather job-related tweets intended to capture economic activity in the United States and Canada. Relationships between social media activity on these topics and the quantity of Coca-Cola bag ordered can be useful information for our forecasting models. Twitter data is a combination of user-level tweets and company-level tweets, while Google Trend data is monthly trend value for our selected terms. User-level tweets are Twitter posts from regular online consumers tweeting about Coca-Cola, Pepsi, McDonald's, Taco Bell, and "jobs." Company-level tweets are Twitter posts by the official Twitter accounts of Coca-Cola, Pepsi, McDonald's, and Taco Bell. Meanwhile, Google Trends is a popularity measure for Coca-Cola and other relevant terms based on their search frequency over time.

Sentiment Analysis

One method of quantifying text-based social media data for our forecasting models is by implementing sentiment analysis. This is a technique in natural language processing that we will use for each tweet to generate a numerical value signifying whether consumers have a positive or negative outlook on Coca-Cola, Pepsi, McDonald's, Taco Bell, or jobs. We can then average sentiment scores by month for each relevant term, and use this as an additional predictor to forecast Coca-Cola bag orders. For this project, we will only calculate sentiment scores for user-generated tweets because we assume that tweets generated by the official company accounts are all positive. Prior to conducting sentiment analysis, text processing steps must be conducted on tweets. The following steps were conducted on the tweets: *Remove stop words on tweets* Tokenize tweets *Lemmatize tweets

The R package sentimentR was used to calculate sentiment scores for each tweet. This

package takes into account additional information such as valence shifters and de-amplifiers resulting in a more accurate sentiment score (see Appendix A). The sentiment scores for each selected term's collective tweets per month will be averaged at the monthly level to generate the monthly average sentiment variable.

Stationarity and Differencing

An important step to consider when forecasting is to remove trends and seasonality from variables in order to make it stationary. When time series data is stationary displaying a stable mean and stable variance over time, it is less likely to produce spurious relationships and misleading results. Statistical tests (KPSS test) were conducted on each variable to determine its stationarity (see Appendix A). After testing, it was determined that the dependent variable, monthly quantity of Coca-Cola bag orders, was stationary. This means that this variable requires no further transformations. However, the independent variables showed varying results and require additional processing. One way of transforming time series data to become stationary is the method of differencing. Differencing is the method of subtracting the value of the current time step from the value of previous time step(s). This method was applied to all social media variables to ensure that they were all stationary. The visual below demonstrates how the method of differencing is able to remove trends from the monthly total user tweets. The top half of the visual are time plots of the pre-differenced variables, while the bottom are time plots of the differenced variables.

```
include_graphics(path = "figure/differencing.jpg")
```
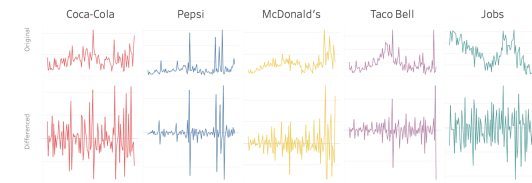

Figure 1: differenced social media data

Dimension Reduction

When building forecasting models, it is important to be aware of the level of complexity of these models. In this project, we will be collecting Twitter data that allow users to collect over a hundred features for a single tweet (tweet text, user profile data, etc.). Using all of these variables will make our forecasting models highly complex and likely result in poor predictions. Fortunately, the scope of this project limits tweet information to only a tweet's text, number of likes, number of retweets, and number of replies. However, the complexity of this project (relevant topics, social media data type, social media source) still leaves us with 46 total independent variables per observation (see Data section). We will use two main approaches in this project to further reduce our social media variable's dimensions.

Method A - Principal Component Analysis

The first method we will use to reduce the dimensionality of our social media variables is principal component analysis (PCA). PCA uses an orthogonal transformation of our variables into linearly uncorrelated variables called principal components. The main idea is that the majority of the variance explained will be concentrated on a limited number of principal components. This allows us to discard the principal component variables that provide little additional information. This approach further reduces the overall dimension of our original set of variables. When performing this technique with our social media variables, we are able to observe the "elbow feature" at number of principal component (n) = 7 . This tells us that only the first seven principal components is required to explain most of the total variance (97%) of our variables. By using PCA, we were able to reduce our overall social media variables from 46 to 7.
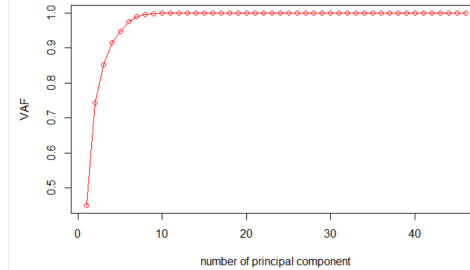
```
include_graphics(path = "figure/pca.jpg")
```


Figure 2: Principal Component Analysis

Method B - Cross Correlation

The second method we will employ to reduce our total features is by testing our independent variables for cross correlation with the dependent variable. Mainly, this approach will inform us on how many months in advance a social media variable can lead to an increase or decrease in Coca-Cola bag orders. For example, consider when a social media variable was found to be significantly positively correlated with Coca-Cola bag orders at lag t-1. If this social media variable has a positive value for the current month, then an increase in Coca-Cola bag orders can be expected the following month. In this project, we will check for cross correlation on each independent variable up to six months prior (t-6). We will do this on the original variables as well as the principal component variables. Using the cross correlation approach, we were able to identify eight lags from the social media variables that were cross correlated with Coca-Cola bag orders. Below is a table of these social media variables with their specified lags listed.

Table 1: Social Media Variables with Specified Lags

Social Media Variable	Significant Lag
Coca-Cola Account tweet	s t-1

Social Media Variable	Significant Lag
Taco Bell Account tweet	s t-1
Job Google Trend	t-2
McDonald's Google Trend	t-2
McDonald's Account repli	es t-5
Taco Bell Google Trend	t-5
McDonald's Google Trend	t-5
Pepsi Account tweets	t-6

In addition, we were able to identify seven lags from principal components that were cross correlated with Coca-Cola bag orders. Below is a table of these principal components with their specified lags listed.

Table 2: Principal Components with Specified Lags

Principal Component	Significant Lag
PC7	t-1
PC6	t-2
PC4	t-2
PC6	t-3
PC4	t-3
PC3	t-5
PC3	t-6

Ensemble Modeling

A variety of different machine learning models will be used to forecast future Coca-Cola bag orders (see Modeling Framework). In addition to these machine learning models, this project will demonstrate the strength of the ensemble model approach. The main assumption to ensemble modeling is that combining all lower-level models will result in a more accurate, overall model. An ensemble model is able to highlight the strength of each individual model and account for each model's weaknesses. The ensemble model approach will be used in this project to produce the best model.

Methodology

Data

The data used to predict Coca-Cola bag orders will come from three distinct sources: Scholle IPN, Twitter, and Google Trends. The aggregation of all relevant data will be at the monthly level with a date range from October 2009 to October 2018. This will ensure that the data will have enough observations for our forecasting models. The dependent variable for this project will be the monthly total bag orders from Coca-Cola, which will be calculated using Scholle IPN's internal sales data. The independent variables are social media variables collected from Twitter and Google Trends. Twitter is a widely used social media platform in the United States and Canada that was founded in 2006. This platform will allow us to get a feeling about users and their opinions on Coca-Cola, Pepsi, McDonald's, Taco Bell, and 'jobs'. Most importantly, Twitter gives users access to data elements about a tweet necessary for this project, including the date a tweet was posted and the reactions a tweet received (i.e. number of likes, replies, and retweets). Twitter's years of existence, popularity, and data features make it an ideal social media data source for this project. However, Twitter does have a number of limitations. One of the main disadvantages of using Twitter data is its high volume and high dimensionality. Twitter receives millions of tweets a day that has information about the actual tweet (number of likes, replies, etc.), the user who posted the tweet (username, location, etc.) and the users who interact with the tweet (replied to tweet, username, etc.). A well defined scope will limit the amount of tweets to be collected and will allow us to prepare for any data storage and computational needs.

Google Trends will give us an opportunity to see how frequently internet users search for Coca-Cola, Pepsi, McDonald's, Pepsi, and 'jobs' in Google. The main advantage of Google Trends is the ease in which we are able to collect this data. Google Trends allow users to select aggregate level (monthly, annual), location, and a date range for each query. The main disadvantage of Google Trends is that the raw data used to generate the trend value is not available. This makes it challenging to validate unusual trends in a search query. For more information on the specific variables, please refer to the Social Media Variables section. For more information on how the social media data were collected, please refer to Appendix A. Below is a summary table of the monthly social media variables:

Social Media Variable	Coca-Cola	Pepsi	McDonald's	Taco Bell	Jobs
Total Account Tweets	x	x	x	x	n/a
Total Account Replies	x	x	x	x	n/a
Total Account Likes	x	x	x	x	n/a
Total Account Retweets	x	x	x	x	n/a
Total User Tweets	x	x	x	x	x
Total User Replies	x	x	x	x	x
Total User Likes	x	x	x	x	x
Total User Retweets	x	x	x	x	x
Average Tweets Sentiment	x	x	x	x	x
Google Trend	x	x	x	x	x

Exploratory Data Analysis

In order to ensure maximum utility of the collected social media data and produce accurate Coca-Cola bag order forecasts, it is important to conduct exploratory analysis on these variables. Analyzing all the variables prior to modeling will allow us to better understand trends in our variables. Identifying these trends can aid in the interpretation of our forecasting results. In this section, we will provide a brief analysis of each variable and highlight trends that could be of value when predicting the demand for Coca-Cola bag orders.

- Monthly Coca-Cola Bag Orders

[INSERT FIGURE]

Above is a time plot of monthly Coca-Cola syrup bag orders from October 2009 to October 2019. At a glance, we observe an annual seasonal pattern with no consistent trend over time. Overall, the bag quantity orders show a mean of 7,521,158 per month. Expectedly, Scholle IPN experience the highest volume of Coca-Cola bag orders during summer months (June-August) with averages over eight million ordered bags. The only other month with an average of over eight million Coca-Cola bags is during the month of March.

- User-generated Tweets User-generated tweets provides our forecasting models with information regarding how frequently Twitter users talk about Coca-Cola, Pepsi, McDonald's, Taco Bell, and "jobs". When looking at the average monthly tweet mentions for each term, the term "jobs" surprisingly appeared the most. Jobs-related tweets in the dataset represent 42.7% of all the user-generated tweets collected. The

visual below of the average monthly tweet volume for each selected term demonstrates how "jobs" dominated our collected data.

[INSERT FIGURE]

The disparity on tweet volume across topics can be explained by the limitations imposed by the Twitter Search API (please refer to the Appendix for more information). Originally, we did not anticipate the general public to engage in job-related conversation on a social media platform compared to the other relevant terms. This finding provides value to our analysis as the term "jobs" now has a large sample size and can potentially give us better indication on how economic activity in North America impact Coca-Cola bag orders. However, this resulted in a lower total of user-generated tweets for all other relevant terms. Limited sample size for these variables could adversely impact their ability to forecast Coca-Cola bag orders.

- Company-generated Tweets To continue the exploration of covariates, we will explore company-generated tweets. This set of social media variables will provide us information about promotional behavior for these selected companies (monthly tweets) and the subsequent consumer reaction to these promotions (monthly likes, retweets and replies). A total of 24,442 such tweets were collected between October 2009 to October 2018. Among the four companies, the official Twitter accounts of Coca-Cola and McDonald's were the most active in terms of posting Twitter content. The time series plots below (not built to scale) clearly show a general trend among all four companies.

[INSERT FIGURE]

For each company, the total number of tweets per month were low in the beginning portion of the time plot. This was followed by a surge in tweet volume in the middle part of the decade suggesting how all four companies started to heavily utilize Twitter as a promotional tool. This change in behavior from companies could be attributed to the increased use of Twitter by social media consumers around the same time. The time plots below (not built to scale) of total consumer reactions (replies, likes, retweets) is a representation of how Twitter users who engaged with each company reacted to their tweets over time.

Noticeably, Twitter users became much more engaged with these Twitter accounts towards the middle portion of the time plots. This could be a sign of how effective these company's promotions are during the middle of the decade, or it could be an indicator of when social media started becoming more popular in mainstream society. The limited Twitter activity by the Twitter accounts and users during the earlier portion of the time plots should be considered when modeling.

- Google Trends Although Google Trends is not a social media platform, it does provide an easily attainable source of data regarding general interest in Coca-Cola, Pepsi, McDonald's, Taco Bell, and 'jobs'. This could provide useful information when forecasting Coca-Cola bag orders. Above are time series plots for each term and their monthly Google Trend value from October 2009 to October 2018. Generally, there seems to be a seasonal trend that appear for each search term. It is

interesting to note that the two quick service restaurants, McDonald's and Taco Bell, experienced an increasing trend over time.

[INSERT FIGURE]

When using Google Trends, it is important to consider that trend values are calculated strictly based on the search term entered. This strict rule could fail in certain instances where a single term could have different meanings. For example, an unusual spike can be observed in the time series plot for 'jobs' in October 2011. After conducting additional research, this spike can be attributed to a sudden interest in Steve Jobs when he passed away. For Pepsi, a spike in April 2017 was traced back due to its connection to a controversial advertisement involving the celebrity Kendall Jenner. These outliers will be imputed with a value that fit the distribution of the rest of the dataset.

Modeling Framework

Metrics

The following metrics are going to be used to measure the forecasting models we will build using social media variables.

- Symmetric Mean Absolute Percent Error (sMAPE) - A measure based on percentage (or relative) errors.

- Root Mean Square Error (RMSE) - Standard deviation of the mean squared error.

- Mean Accuracy - Measures the forecast accuracy.

- Percentage (%) Bias - Measures the percentage of times the dependent variable was over- or under-estimated.

Selection of Forecasting Models

In this section, we will describe the forecasting models we will use to determine if social media variables have any impact Coca-Cola bag order forecasts. We will build three types of models - baseline model, challenger models, and ensemble model. Each forecasting model will have a forecasting window of 18 months. The baseline model will be a forecasting model using only the monthly Coca-Cola bag orders. The challenger models will be a set of three different forecasting models that will utilize the social media variables. The ensemble model will be a model that combines all three challenger models.

The baseline model will be used to compare how challenger models that utilize social media data fare against a forecasting model that does not utilize this data. The baseline model will be built by using a seasonal Arima model. A sliding-window cross validation of two-years, four-years, and six-years will be used to determine the optimal number of observations used for training the rest of our models.

The challenger models selected for this project are: Regression with Arima errors, XG-Boost, and Long Short-Term Memory (LSTM). Each challenger model will be trained using 48 months of training observations (see Baseline Model Results). In addition, each challenger model will be trained on two sets of social media variables. The first set of social media variables are the difference variables that were identified to be cross-correlated with Coca-Cola bag orders (see Dimension Reduction section). The second set of of social media variables are the principal components that were identified to be cross-correlated with Coca-Cola bag orders (see Dimension Reduction section). The Regression with Arima errors assumes that Coca-Cola bag sales does not only depend on the dependent variable, but also depend on external variables selected. The XGBoost model is a more complex 'regularized boosting' technique that seeks a good bias-variance tradeoff to reduce overfitting. It allows cross-validation at each iteration of the boosting process and thus it is easy to get the exact optimum number of boosting iterations in a single run. The final challenger model, LSTM, is a special kind of recurrent neural network that is capable of learning long-term dependencies by using adaptive, non-linear gates to update cell state information. This model LSTM selectively chooses what it "remembers", and what to "forget." In addition,

LSTM selects how much of its memory it should output. The final model we will use is an ensemble model of all three challenger models. Please refer to the Ensemble Modeling section under Background for more information. We will compare how the challenger models and ensemble model perform against the baseline model using the metrics defined in the previous section.

Findings

Baseline Model Results

Before we test the impact of social media variables in forecasting Coca-Cola bag orders, we must first establish baseline metrics. We will use a seasonal Arima model that uses only the dependent variable to establish our baseline model. Our analysis found that the best seasonal Arima model had an order of (3,0,1)(1,0,0)[12]. Cross validation with a sliding window of two-years, four-years, and six-years was then conducted. The purpose of cross validation is to establish the metric that the challenger models will be compared against and to find the point in the forecast horizon in which the predictions "flatten out". The chart below is a plot of the forecast horizon against each model's sMAPE values.

[INSERT FIGURE]

It is clear that the two-year window model is considerably worse than either the four-year and six-year sliding window models. When you look at a plot of each model against the average sMAPE value for each forecast window, you can see that the sMAPE values flatten out at the four year mark. There is no considerable improvement in sMAPE values between the four-year window model and the six-year window mode. Therefore, the best baseline model is the four-year sliding window model.

Sliding Window	sMAPE	RMSE	Mean Accuracy	Bias (%)
2 Year	12.79%	1,066,599	86.56%	50.00%
4 Year	13.93%	1,215,263	86.64%	38.89%
6 Year	14.14%	1,066,599	86.49%	38.89%

Regression with ARIMA Errors Results

By evaluating the residuals and metrics above, we found that manually passing the ARIMA order (3,0,1)(1,0,0) [12] to the model gave us the best results. The residuals after the model demonstrate the white noise feature, meaning that the model has captured all the information from the time series data. When reviewing the forecast based on the regression with arima (3,0,1)(1,0,0)[12] model, the results are relatively conservative.

As shown above, the forecasted value is relatively stable across the test time period, compared with the actual values. The model produced an sMAPE of 9.04% and RMSE of 0.94M.

Social Media Var	sMAPE	RMSE	Mean Accuracy	Bias (%)
Differenced Social Media Variables	37.79%	3,107,383	65.05%	55.55%
Principal Components	19.11%	1,381,487	80.33%	38.89%

XGBoost Results

For the XGBoost model, we introduced a month of year variable to allow the model to learn seasonality. XGBoost was implemented in Python using the parameters below before using grid search to identify the best tuning parameter.

After running the model, we are able to rank the features based on importance (see below). [INSERT FIGURE]

Social Media Var	sMAPE	RMSE	Mean Accuracy	Bias (%)
Differenced Social Media Variables	6.21%	597,983	93.88%	27.78%
Principal Components	7.27%	701,980	93.01%	38.89%

Long Short-Term Memory Results

We implemented LSTM in Python using the same set of variables as the previous models. Below is a model summary of LSTM prior to implementation. [INSERT FIGURE]

Tuning the parameters is a vital part when deploying any Neural Network models, and LSTM is no exception. Listed below are a few key parameters we tweaked:

- Epochs - One Epoch is when an entire dataset is passed forward and backward through the neural network only once. Since one epoch is too big to feed to the computer at once we divide it in several smaller batches. (we need to quote this!)
- Optimizer - adam
- Loss function - mean squared error Activation function - since we scaled the train and test sets beforehand, we chose to leave the activation function default.
- Dropout - we had 4 dropout layers embedded in the model. Dropout is a regularization technique for neural network models proposed by Srivastava, et al. in their 2014 paper Dropout: A Simple Way to Prevent Neural Networks from Overfitting.

[INSERT FIGURE]

Social Media Var	sMAPE	RMSE	Mean Accuracy	Bias (%)
Differenced Social Media Variables	11.8%	994,364	88.47%	44.4%
Principal Components	38.89%	1,114,561	86.82%	38.89%

Ensemble Model Results

After building challenger models that utilize social media variables, we will now construct an ensemble model. Ensembling the three challenger models we built should produce more accurate Coca-Cola bag order predictions. We will use several ensemble models and compare these results against each individual challenger model. The ensemble models we will build are 'average of all models', linear regression, and random forest.

The average of all predictions ensemble model is the most interpretable of all the ensemble models as it simply takes the average of all predicted values among all challenger models. Meanwhile, the linear model ensemble model will assign a weight to each challenger model's results. The more that a model contributes to the predicted results, the more weight the linear model will assign it. Finally, the random forest ensemble model is the most complex among the three. This model will capture how each challenger model's specific weaknesses are in prediction. For the linear regression and random forest models, the predictions of the three challenger models will be used as training observations.

Model	sMAPE	RMSE	Mean Accuracy	Bias (%)
Scholle Baseline Model	7.43%	667,832	92.84%	27.78%
Average of all Models	8.41%	809,139	91.53%	33.33%
Linear Regression	5.76%	545,211	94.22%	50.00%
Random Forest	3.57%	356,563	96.37%	44.44%

The results show that the random forest ensemble model performs the best according to the sMAPE and RMSE metrics. However, since the linear model produced acceptable results and is the simpler ensemble model, this will be selected as the champion model. Below is a graph visualizing our individual model forecasts against the actual quantity (black).

[INSERT FIGURE]

Coca-Cola Ensemble Model Results

To further demonstrate the strength of ensemble modeling, this method was also used to combine the models generated by other capstone teams. Other teams were solving the same business problem, but using a different set of covariates (related products and economy).

The idea is that ensembling all the team's models will compensate for each model's weaknesses and produce a strong overall model. The final model generated incremental gains from our best ensemble model. The results are outlined in the next section.

Model	sMAPE	RMSE	Mean Accuracy	Bias (%)
Scholle Baseline Model	7.43%	667,832	92.84%	27.78%
Linear Regression (social media)	5.76%	545,211	94.22%	50.00%
Linear Regression	1.05%	93,662	98.95%	44.44%
Random Forest	1.34%	119,437	98.64%	50.00%

Conclusion

In summary, we learned that including social media variables is helpful in predicting Coca-Cola bag orders, and it also provided a comparable results to the Scholle Coca-Cola baseline model. For the individual models, the best results came when the social media variables were differenced or when the dimensions were reduced leading to increased predictive power. When comparing the two challenger models, the XGBoost model did a better job of capturing seasonality than the regression with ARIMA errors model, as well as the LSTM RNN. In addition, the XGBoost approach allows us to point out the most important variables. As for an ensemble approach to modeling, this technique proved to be a strong approach to produce better forecasts. For our project, the random forest ensemble returned the best RMSE and sMAPE values. For the ensemble models among the different teams, a linear regression model produced an output that was incrementally better than any individual model.

Recommendations

For this project, we focused our data collection on only a handful of quick service restaurants and competitors. Collecting social media data from Coca-Cola's competitors presents an opportunity to look at how people engage with and view these beverage companies at an industry level. Hence it may bring better accuracy in predicting the Coca-Cola bag sales for Scholle IPN.

Appendix A

KPSS Test `https://www.rdocumentation.org/packages/tseries/versions/` `0.10-46/topics/kpss.test` (Insert table of KPSS test results - before and after differencing)

Data Cleansing Scholle IPN Data The Scholle IPN data provides a variety of relevant data features that uninformed users can use to gain knowledge about Scholle IPN's product. Our team will use this data to find relevant trends that may be useful in forecasting. The table below describes the Scholle IPN metadata:

Prior to searching for any relevant patterns in the Scholle IPN data, it is best practice to conduct preliminary data cleansing. First, we notice that there are 8,204 total orders in this dataset ranging from a delivery date of 10/1/2009 to 11/2/2018. It is important to note that this dataset is a combination of orders from all of Scholle IPN clients in their syrup bag product line. Of all the orders, 5,118 of the records lists 'Coca-Cola' as the top business partner. Other issues include null values being present for the 'Ship_to_State' variable and negative values for 'quantity' variable.

The following data cleansing steps were conducted to prepare the Scholle IPN data for monthly aggregation: *Filter syrup bag orders for 'Coca-Cola' top business partner.* Drop records with quantities of less than one thousand.
Impute records with missing values in 'Ship_to_State' with 'Bayamon, PR'. Drop records outside of the United States and Canada market. *Drop records that occurred after October 2018.

Once the steps listed have been conducted, the primary data is left with only 4,875 records. The final step to data cleansing will be aggregating the records at the monthly level. The 'Planned Delivery Date' field will be used to aggregate quantities at the desired level.

Twitter Data Collection The selected terms (Coca-Cola, Pepsi, McDonald's, Taco Bell, Jobs) will be used to collect user-generated tweets. The collection of this data will be done by using Twitter's Search API. Twitter posts containing the terms 'coke', 'pepsi', 'mcdonalds', 'taco bell', or 'jobs' will be collected from the 15th of every month from July 2009 to December 2018. Twitter's Search API impose strict restrictions on query results depending on the account level purchased. For this project, a maximum of 500 user-generated tweets will be collected per month in the given date range. For example, for the month of December 2018, there will be 500 tweets collected related to Coca-Cola, Pepsi, McDonald's, Taco Bell, and jobs. This means that at any given month, the distribution of

tweets across different terms can differ depending on what search term is trending.

Since this project is concerned with the US and Canada markets, we need to add a series of Search API query constraints to ensure that collected tweets are mainly from users in these selected markets. The first constraint was to query results in the english language by filtering for user profiles with country codes of 'us' or 'ca'. The second constraint involves the time of day that the tweets were posted. Since Twitter Search API collects tweets in reverse chronological order given a date and time, a time constraint of 23:00 UTC (6 pm CST) was applied to the queries to ensure that collected tweets will fall in a time range when most people in the US and Canada are off normal working hours. This will provide a robust collection of tweets from a variety of different users across different regions of the US and Canada. The final output of these Twitter search queries will include the tweet content, tweet date and time, number of replies, number of retweets, and number of likes.

The second approach to collecting Twitter data is to gather historical Twitter posts generated by the official Twitter accounts of Coca-Cola, Pepsi, McDonald's, and Taco Bell from July 2009 to December 2018 using a separate web scraping tool. These selected Twitter accounts vary greatly in their online activity. While some accounts, like Taco Bell, reply frequently to their followers, other accounts tend to be less engaging and post infrequently. For the purpose of this project, only independent tweets posted by each Twitter account will be collected. This means that tweet replies to Twitter users will not be collected. The main benefit of this approach is that it allows us to focus on tweets that promote individual marketing campaigns. The final output of this data collection step will include tweet content, tweet date and time, number of replies, number of retweets, and number of likes.
The total output of the Twitter data collection results are 56,528 user-generated tweets and 24,442 company-generated tweets from July 2009 to December 2018. For both datasets, tweets outside of October 2009 to October 2018 will be filtered out to prepare for time series modeling.

The user-generated tweets will require additional pre-processing steps prior to monthly aggregation. The Twitter Search API collects tweets on a rule-based approach meaning it is probable that nonsensical tweets may be included in the raw data collected. For example, tweets referencing the drug 'cocaine' (or 'coke' in colloquial terms) were identified in the raw data. A series of natural language processing steps were conducted on the tweet's text to help screen out these "noisy" tweets. In addition, these steps will be important to prepare the tweet data for aggregation. Some of these processing steps include removing stopwords, creating n-grams, and tokenizing text. After conducting text processing steps, an R package was used to calculate the sentiment of a given tweet. This step will provide us with an additional feature that will inform business users about user sentiment on their product. After these pre-processing steps, the tweets will now be aggregated on a monthly level to generate the following independent variables for each search term (Coca-Cola, Pepsi, McDonald's, Taco Bell, "Jobs"): total tweets per month, total replies per month, total likes per month, total retweets per month, and average sentiment per month. There are a total of 25 variables generated from this step. The company-generated tweets from Coca-Cola, Pepsi, McDonald's, and Taco Bell will be aggregated on a monthly level to generate the following independent variables for each Twitter account (Coca-Cola, Pepsi,

McDonald's and Taco Bell): total tweets per month, total replies per month, total likes per month, total retweets per month. There are a total 16 variables generated from this step.

Google Trends Data Google Trends data will also be collected to better anticipate the demand for Coca-Cola products. Although Google Trends is not a social media platform, it does provide its users with quantified information about the relative popularity of a search term over time. The monthly Google Trend values for the following terms will be collected: Coca-Cola, Pepsi, McDonald's, Taco Bell, and 'Jobs'. Each topic will be queried individually with the date range October 2009 to October 2018. Google Trend values for each term will range between 0 to 100, with 100 signifying the month in which the term was most popular. No additional preprocessing is required for Google Trends data. There are a total of five variables generated from this step.

Exploratory Data Analysis - Scholle IPN The maximum quantity of bags ordered occurred in November 2011 with a total of 13,685,000 bag orders, while the minimum quantity of bags ordered occurred in January 2011 with a total of 5,215,810 bag orders. After exploring the Scholle IPN data, it was also observed that the Columbus (OH), Atlanta (GA), and Dallas (TX) are the top cities where Coca-Cola bag orders were being delivered. Given that information, it is less surprising to observe that the top three states by quantity of bag order deliveries were Ohio, Georgia, and Texas. In terms of Coca-Cola bag products, the stock-keeping unit (SKU) 200258 and 200144 were the two most ordered SKU.

Exploratory Data Analysis - Social Media Variables Similar to the Scholle IPN data, it is important to explore our covariates to identify any underlying trends that could be used to forecast Coca-Cola demand. For user-generated tweets, a total of 52,952 user-generated tweets were collected between October 2009 to October 2018. Among the four companies, McDonald's is the most talked about company over the given date range. When exploring which terms appear the most with each other, the term that appeared most commonly with Coca-Cola is Pepsi.

Sentiment Analysis `https://github.com/trinker/sentimentr`

References

Angel, E. (2000). *Interactive computer graphics : A top-down approach with opengl.* Boston, MA: Addison Wesley Longman.

Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with quicktime.* Boston, MA: Wesley Addison Longman.

Angel, E. (2001b). *Test second book by angel.* Boston, MA: Wesley Addison Longman.