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, because failing to predict the future demand can result in product surplus or shortage. In an era where social media dominates people's lives, data analytics has become a powerful tool to generate actionable business insights. Scholle IPN tasked us to test if social media brings any predictive power to forecast the future Coca-Cola bag sales. In this project, we examined whether people's discussion and engagement on social media have any impact on the consumption of Coca-Cola products in real life. In addition to our team, two other research groups analyze different types of independent variables to predict Coca-Cola bag sales.

In order to achieve that, we collected external data from social media sources and combined it with internal data provided by Scholle IPN. After preliminary exploratory analysis, we built traditional and advaned machine learning models, including regression with ARIMA errors model, XGBoost model and Long Short-Term Memory (Recurrent Neural Network) model. The XGBoost model using the differenced social media variables returned the best predictions while still maintaining a decent level of interpretability. This model combined with other challenger models were stacked together to create three variations of ensemble models - mean average, stacked linear regression, stacked random forest.

Key findings from this project include the following:

1. Pepsi and Taco Bell lagged variables were observed to have the most importance among predictors, according to the XGBoost model feature importance.
2. Ensemble modeling proved to be a great approach to bring individual models together. Among all the social media models, the stacked linear regression model had the best combination of accuracy and stability. Furthermore, ensembling the models from all Scholle IPN research teams produced the best overall model.
3. Identifying cross-correlated variables and conducting principal component analysis were vital steps for modeling. These procedures to reduce the dimensionality of social media variables led to an improved accuracy.

We concluded that Scholle IPN should utilize social media data sources together with their internal sales data to forecast future Coca-Cola bag demand. For future work, we recommended an examination of 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 focusing primarily in bag-in-box packaging. The company implements 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 provided a forecasting solution to Scholle IPN using social media data. This project primarily focused 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 examined whether we can substantially improve Scholle IPN's Coca-Cola demand forecasts by using user mentions and company posts in social media channels as primary variables.

Research Purpose

The purpose of this research is to forecast Coca-Cola bag orders by utilizing social media data. When customers express their opinions in social media, businesses like Scholle IPN can gain valuable insights that can inform business decisions. For example, if a McDonald'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 took into account additional considerations. First, since social media posts are usually in the form of text, we explored methods to convert text to numerical data. Second, we explored different ways to reduce the dimension of social media variables to account for its high dimensionality. And finally, we used 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 were aggregated or averaged at the monthly level. The forecasting window for this project was 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 focused our project on the following variables: tweet text, number of likes, number of retweets, and number of replies. For Google Trends, we extracted monthly trend values for selected topics at the monthly level. Additional data retrieval rules were applied to ensure that Twitter and Google Trends data are from the United States and Canada (see appendix).

Background

Social Media Variables

For this forecasting project, we focused on collecting social media data on selected topics. These selected topics are Coca-Cola, Pepsi, McDonald's, Taco Bell, and "jobs". Coca-Cola was selected because it is their product's demand that Scholle IPN is interested in forecasting. Pepsi was selected due to its position as the main competitor for Coca-Cola. Meanwhile, McDonald's and Taco Bell were chosen because they are the top quick service restaurant partners for Coca-Cola and Pepsi based on 2017 rankings (QSR Magazine, 2018). The topic "jobs" was 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 bags 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 tweet data were Twitter posts from regular online consumers tweeting about Coca-Cola, Pepsi, McDonald's, Taco Bell, and "jobs." Company-level tweet data were Twitter posts by the official Twitter accounts of Coca-Cola, Pepsi, McDonald's, and Taco Bell. Meanwhile, Google Trends data was the trend value for Coca-Cola and the other selected topics on a given month. T

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 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". 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

For this project, we only calculated sentiment scores for user-generated tweets because we assumed that tweets generated by the official company accounts are positively biased to-

wards their own brand. The R package sentimentR was used to calculate sentiment scores for each tweet. This package took into account additional information such as valence shifters and de-amplifiers resulting in a more accurate sentiment score (see appendix). The sentiment scores for each selected term's collective tweets per month was averaged at the monthly level to generate the monthly average sentiment score. The monthly average sentiment score was then added as a variable to forecasting Coca-Cola bag orders.

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, it displays a stable mean and stable variance over time (Hyndman, 2018). This means that it is less likely to produce spurious relationships and misleading results. The KPSS test (Kwiatkowski, 1992) was conducted on each social media variable to determine its level of stationarity (Hyndman, 1992). It was determined that the dependent variable, monthly quantity of Coca-Cola bag orders, is stationary. This means that this variable requires no further transformations. However, the independent variables diplayed varied results and required 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. Figure 1 demonstrated 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.

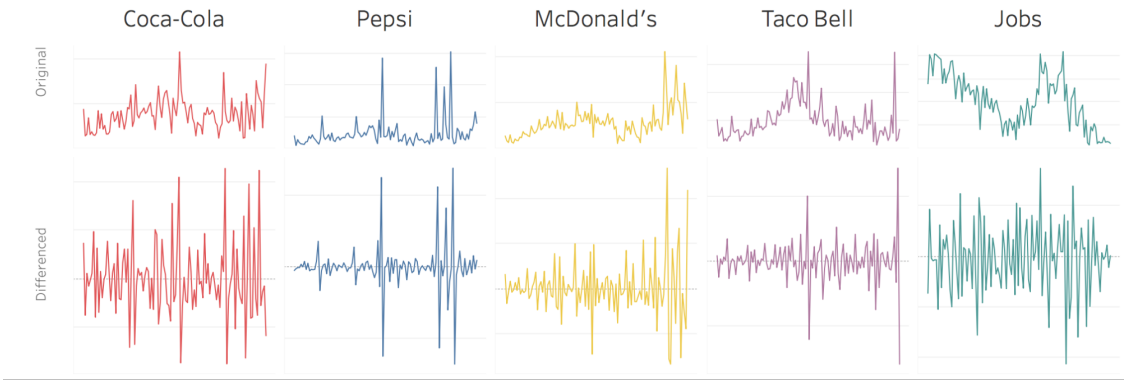

Figure 1: Differenced social media variables

Dimension Reduction

When building forecasting models, it is important to be aware of their level of complexity. In this project, we collected Twitter data with over a hundred features for a single tweet (tweet text, user profile data, etc.). Using all of these features would have made our

forecasting models highly complex and likely result in poor predictions. Fortunately, the scope of this project limited the tweet information required for modeling 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 left us with 46 total independent variables per observation (see Data section). We used two main approaches in this project to further reduce our social media variable's dimensions.

Method A - Principal Component Analysis

The first method we used 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 (Kambhatla, 1997). The advantage of PCA is that a limited number of principal components will explain the majority of the variance in the data. This allowed us to discard the principal components that provide little additional information. However, the major limitation for PCA is that it is generally less interpretable than using the original variables. According to our PCA model of the social media variables, our team identified seven principal components (Figure 2) as the optimal number of components. These 7 components explained 97% of the variable's total variance, successfully reducing the overall social media variables of 46.

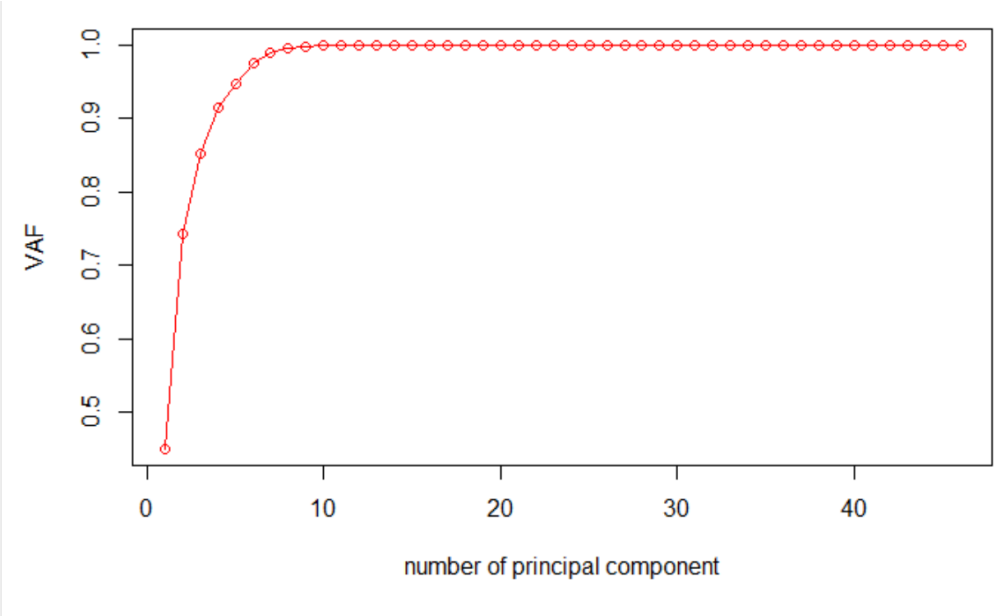

Figure 2: Variable Explained vs. Principal Components Plot

Method B - Cross Correlation

The second method we employed to reduce our total features is by testing independent variables for cross correlation with the dependent variable (Brockwell, 1991). This identified how many months in advance a social media variable could 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 cross correlation value for the current month, then an increase in Coca-Cola bag orders can be expected the following month. For this method, we limit our analysis to only the previous six lags of an independent variable due to the dynamic nature of social media discussions. What is popular this month, may no longer be popular in subsequent months. We checked for cross-correlation on the original variables as well as the principal component variables. The cross correlation approach identified eight lags from the social media variables that were cross correlated with Coca-Cola bag orders (Table 1).

Table 1: Social Media Variables with Specified Lags

Social Media Variable	Significant Lag
Coca-Cola Account tweets	t-1
Taco Bell Account tweets	t-1
Job Google Trend	t-2
McDonald's Google Trend	t-2
McDonald's Account replies	t-5
Taco Bell Google Trend	t-5
McDonald's Google Trend	t-5
Pepsi Account tweets	t-6

In addition, we identified seven lags from principal components that are cross correlated with Coca-Cola bag orders (Table 2).

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 is used to forecast future Coca-Cola bag orders (see Modeling Framework). In addition to these machine learning models, this project demonstrated the strength of the ensemble model approach (Pavlyshenko, 2019). The main assumption to ensemble modeling is that combining all lower-level models results 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 came from three distinct sources: Scholle IPN, Twitter, and Google Trends. The aggregation of all relevant data was at the monthly level with a date range from October 2009 to October 2018. This ensured that the training data had enough observations for our forecasting models. The dependent variable for this project is the monthly total bag orders from Coca-Cola, which is calculated by using Scholle IPN's internal sales data. The independent variables were 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. The main advantage of Twitter is that it allows us to get a feeling about Twitter users and their opinions on Coca-Cola, Pepsi, McDonald's, Taco Bell, and 'jobs'. Most importantly, Twitter gave us access to data elements about a tweet that are 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 made 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 limited the amount of tweets we need to collect and allowed us to prepare for any data storage and computational needs.

Google Trends gave 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 allowed us to select the aggregate level (monthly, annual), location, and 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 made it challenging to validate unusual trends in a search query.

Additional information on how search queries were configured are available in the appendix. Table 3 is a summary table of the monthly social media variables:

Table 3: 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 was important to conduct exploratory analysis on these variables. Analyzing all the variables prior to modeling allowed us to better understand trends in our variables, which aided in the interpretation of our forecasting results. In this section, we provide a brief analysis of each variable and highlight trends that were valuable when predicting the demand for Coca-Cola bag orders.

Monthly Coca-Cola Bag Orders

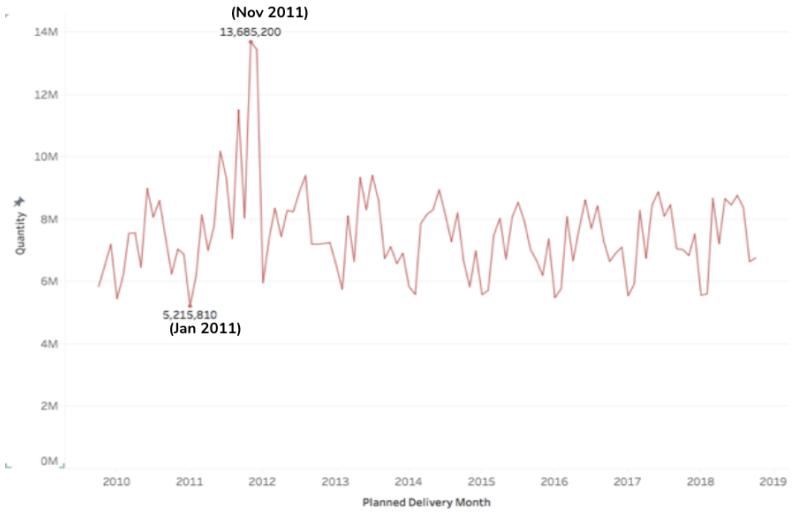

Figure 3: Monthly Coca-Cola Bag Orders Time-series Plot

Figure 3 is a time plot of monthly Coca-Cola syrup bag orders from October 2009 to October 2019. We observed an annual seasonal pattern with no consistent trend over time. Overall, the bag quantity orders showed a mean of 7,521,158 per month. Expectedly, Scholle IPN experienced the highest volume of Coca-Cola bag orders during summer months (June-August) with averages of over eight million ordered bags (Figure 4). The only other month with an average of over eight million Coca-Cola bags ordered is during the month of March.

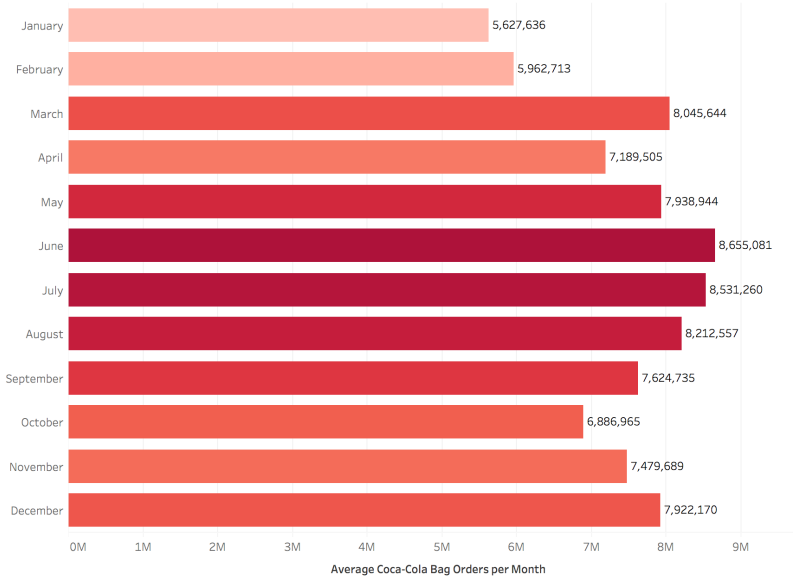

Figure 4: Average Coca-Cola Bag Orders by Month

User-generated Tweets

User-generated tweets provided our forecasting models with information regarding how frequently Twitter users talked 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 represented 42.7% of all the user-generated tweets collected. Figure 5 demonstrate how "jobs" represented the majority of the collected user-level data.

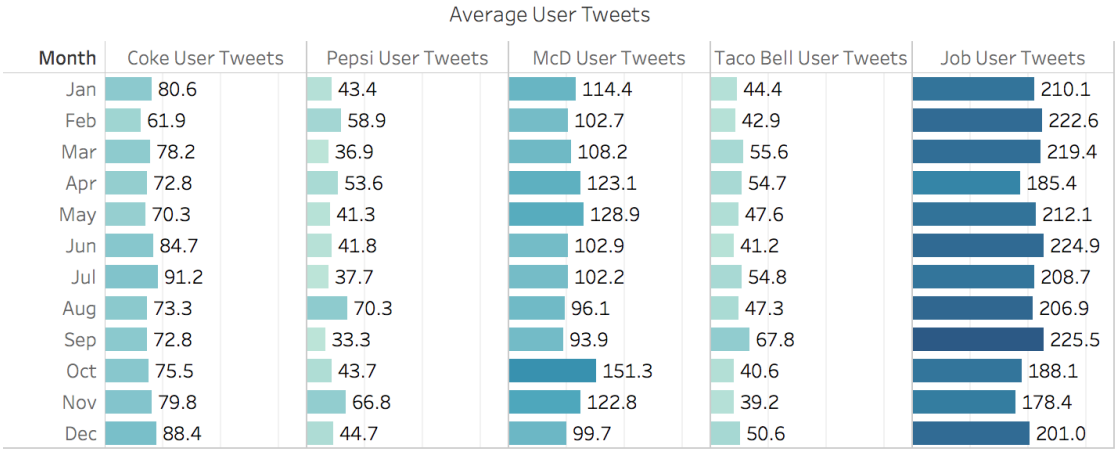

Figure 5: Average User tweets by month

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 provided value to our analysis as the term "jobs" has a large sample size. This gave us better indication on how economic activity in North America impacted Coca-Cola bag orders. However, a lower total of user-generated tweets for all other relevant terms resulted. Generally, limited sample size for variables could adversely impact its ability to predict.

Company-generated Tweets

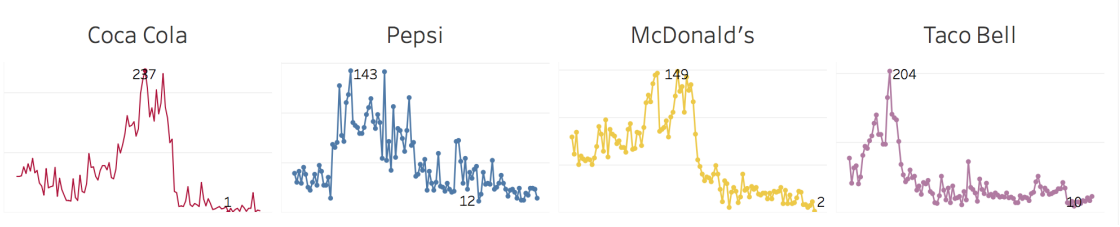

Figure 6: Company Tweet Volume Time-series Plot

To continue the exploration of covariates, we explored company-generated tweets. This set of social media variables provided 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 in Figure 6 (not built to scale) clearly showed a general trend among all four companies.

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. Figure 7 are time plots (not built to scale) of total consumer reactions (replies, likes, retweets) to company Twitter posts over time.

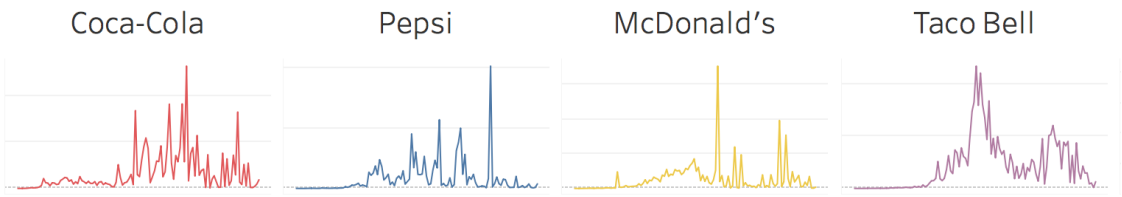

Figure 7: Company Tweet Reaction (Likes, Replies, Retweets) Time-series Plot

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 provided an easily attainable source of data regarding general interest in Coca-Cola, Pepsi, McDonald's, Taco Bell, and 'jobs'. This provided useful information when forecasting Coca-Cola bag orders. The time series plots in Figure 8 are each topic's 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.

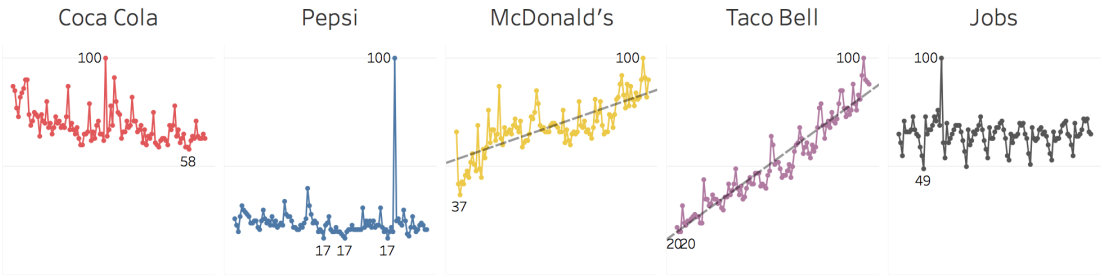

Figure 8: Google Trends (by topic) Time-series Plot

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 was observed in the time series plot for 'jobs' in October 2011. After conducting additional research, this spike was attributed to a sudden interest in Steve Jobs passing away in that same month. 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 were imputed with a value that fit the distribution of the rest of the dataset.

Modeling Framework

Metrics

The following metrics were used to measure the forecasting models built using social media variables.

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

$$\sum_{i=1}^{N}(1-|\frac{F_t-A_t}{A_t}|)(100)$$

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

$$\sum_{i=1}^{N}\sqrt{(F_t-A_t)^2}$$

- Mean Accuracy - Measures the forecast accuracy.

$$\frac{100}{n}\sum_{i=1}^{N}\frac{F_t-A_t}{(|F_t|+|A_t|)^2}$$

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

$$\frac{P_O}{P_T}$$

where, F = forecast, A = actual, t = time-step, P = Count of predictions, O = Overestimated, and T = Total.

Selection of Forecasting Models

In this section, we describe the forecasting models used to determine if social media variables have any impact on Coca-Cola bag order forecasts. We built three types of models - baseline model, challenger models, and ensemble models. Each forecasting model was set to have a forecasting window of 18 months. The baseline model is a forecasting model using only the monthly Coca-Cola bag orders. The challenger models are three different forecasting models that utilized the social media variables. The ensemble models are three models that combined all three challenger models.

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

The challenger models selected for this project were: regression with Arima errors, XGBoost, and Long Short-Term Memory (LSTM). The regression with Arima errors assumes

that Coca-Cola bag sales does not only depend on the dependent variable, but also depended on the external variables selected (Hyndman, 2018). 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 (Chen, 2016). 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 (Hochreiter, 1997). This model selectively chooses what it "remembers", and what to "forget."

Each challenger model was trained using 48 months of training observations (see Baseline Model Results). In addition, each challenger model was trained on two sets of social media variables. The first set of social media variables were the differenced variables identified to be cross-correlated with Coca-Cola bag orders (see Dimension Reduction section). The second set of social media variables were the principal components identified to be cross-correlated with Coca-Cola bag orders (see Dimension Reduction section). The set of social media variables that produced the better challenger model results were used for the ensemble models (Figure 9).

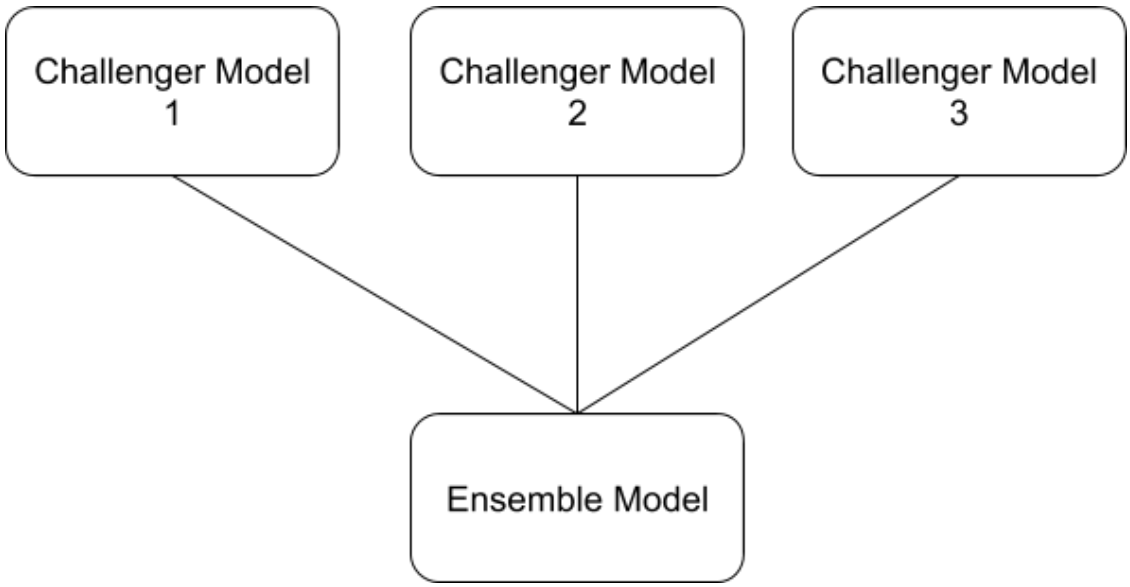

Figure 9: Ensemble Framework

The next type of models built were ensemble models of all three challenger models. For ensemble modeling, the predictions of the three challenger models were used as training observations. The ensemble models built were the 'average of all models', linear regression model, and random forest model. The average of all predictions ensemble model is the most interpretable of all the ensemble models as it simply took the average of all predicted values from all the challenger models. Meanwhile, the linear model ensemble model assigned a weight to each challenger model's results. The more a challenger model contributed to the predicted results, the more weight the linear ensemble model assigned it. Finally, the

random forest ensemble model is the most complex and least interpretable among the three ensemble models. The results of each ensemble model were evaluated using the same metrics as the baseline and challenger models.

Coca-Cola Ensemble Model

Scholle IPN hosted two other research teams from the University of Chicago to predict Coca-Cola bag orders using other external variables. The two teams use related products data and economic/demographic data. Since each team focused on different types of indicators, it is possible that a model is weak in forecasting certain aspects of Coca-Cola bag orders. In order to account for these potential weaknesses, we employ an ensemble model to combine the three best models from each research team and the Scholle IPN Coca-Cola baseline model. The predictions of each model is blended using three methods: (1) average of all models, (2) linear regression, and (3) random forest. The resulting model is evaluated using the established metrics. The best performing ensemble model is selected as the final Scholle IPN Coca-Cola bag order forecasting model.

Coca-Cola System Integrated Workflow

There are three major stages in the workflow for Coca-Cola research teams: data processing, data modeling and prediction (Figure 10). In the data processing procedure, all teams were provided with Scholle IPN's internal sales data. In addition, each team acquired additional data sources externally for individual analysis. After initial data cleaning and engineering, each team applied predictive models to find the best performing model for their external data. For the Coca-Cola team ensemble model, a linear regression model was selected, which was used to generate the final Coca-Cola bag forecasts.

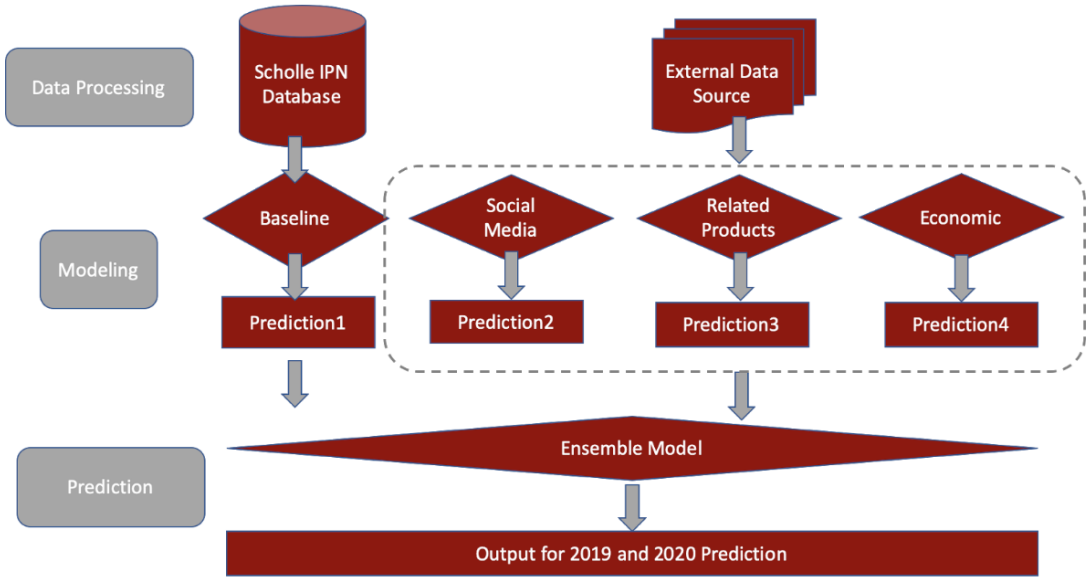

Figure 10: Coca-Cola Integrated Workflow

Scholle Forecasting Model Monitoring

The purpose of model monitoring is to ensure that the model is predictive during the deployment stage. The monitoring system will provide a systematic approach to check whether or not new predictions are out of reasonable range. The following two methods were used for monitoring:

- sMAPE/RMSE Monitoring

After a given month's Coca-Cola bag order quantity is finalized, this value will be used to calculate sMAPE and RMSE values from the forecasts in each of the previous six months. The averages and standard deviations for this six month window will be calculated. If the actual sMAPE and RMSE values for that given month is outside of two standard deviations, an alert will be sent to Scholle IPN.

Forecasting Month	F_0 (actual)	F_1	F_2	F_3	F_4	F_5	F_6
April 2019	*2019-04*	*2019-05*	*2019-06*	*2019-07*	*2019-08*	*2019-09*	***2018-08***
May 2019	*2019-05*	*2019-06*	*2019-07*	*2019-08*	*2019-09*	***2019-10***	
June 2019	*2019-06*	*2019-07*	*2019-08*	*2019-09*	***2019-10***		
July 2019	*2019-07*	*2019-08*	*2019-09*	***2019-10***			
August 2019	*2019-08*	*2019-09*	***2019-10***				
September 2019	*2019-09*	***2019-10***					
October 2019	***2019-10***						

Figure 11: sMAPE and RMSE Monitoring Framework

- Residual Monitoring

The residuals for each month's prediction is calculated by subtracting the predicted value from actual Coca-Cola bag orders. It is either a positive or negative value but overall residuals for the 18-month forecast window should follow a normal distribution around zero mean, which indicates the model is not biased.

Calculate the mean and sigma of residuals on a 18-month basis using the rolling window method as new data is processed each month. Check if the mean stays within plus/minus two standard deviations of zero. If it meets the criteria, we conclude the model is performing well, otherwise we conclude the model is biased and Scholle IPN will get notified.

Findings

Baseline Model Results

Before we tested for the impact of social media variables in forecasting Coca-Cola bag orders, we first established baseline metrics. We used a seasonal Arima model incorporating only the Scholle IPN's Coca-Cola bag order data 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]. Figure 12 is a plot of the forecast horizon against each model's sMAPE values.

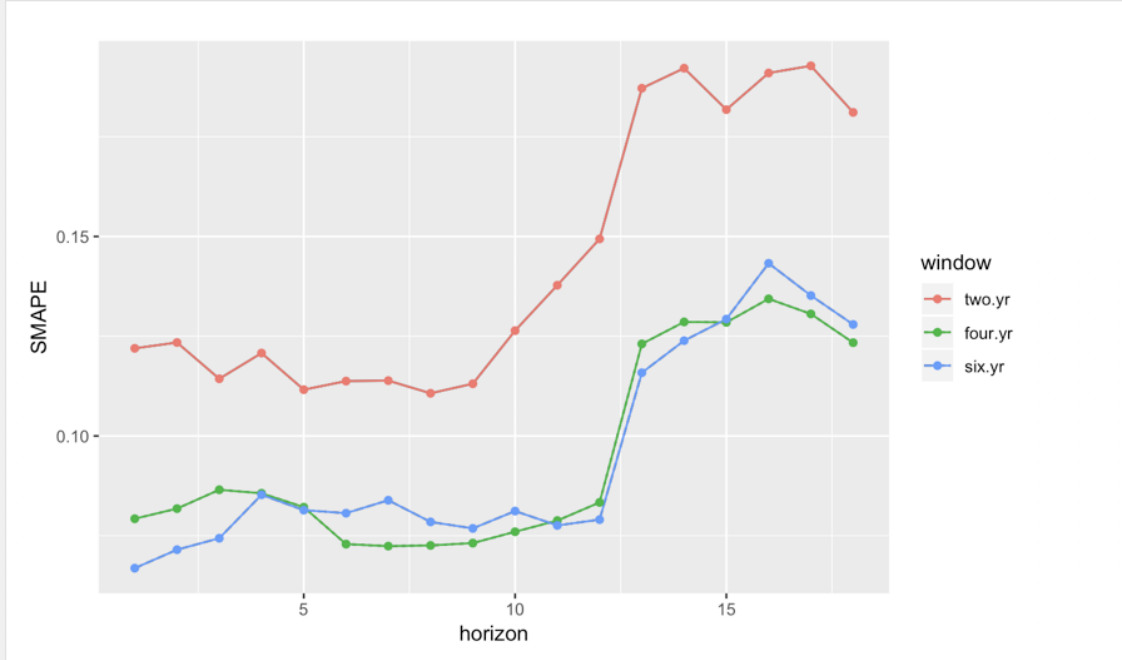

Figure 12: sMAPE vs. Horizon Plot

Figure 12 clearly showed how prediction results flatten out after a certain point in time for each sliding window model (2, 4 and 6 years). For this baseline model, the prediction results flattened out after twelve months. In other words, this baseline model provided the most value with a forecast horizon capped at twelve months. In addition, figure 12 clearly showed that the two-year window model is considerably worse than either the four-year and six-year sliding window models. Overall, there was no considerable improvement in

sMAPE values from the four-year window model to the six-year window model. Since there was not much improvement observed when using six years worth of data over four years, the baseline model selected was the seasonal Arima model trained on four years worth of observations.

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, we found that regression with Arima errors produced the best results using the lagged principal components. With a sMAPE value of 9.76% and RMSE value of 888,544, which was an improvement compared to the Scholle Coca-Cola baseline model.

Social Media Variable	sMAPE	RMSE	Mean Accuracy	Bias (%)
Differenced Variables	30.84%	2,793,003	73.24%	55.00%
Principal Components	9.76%	888,544	90.20%	33.00%

XGBoost Results

For the XGBoost model, we introduced a month of year variable to allow the model to learn seasonality. The XGBoost performed better using the lagged social media variables. This model produced a sMAPE value of 6.21% and RMSE value of 597,983, which is the best performing challenger model.

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

After running the model, we are able to rank the features based on importance (Figure 13).

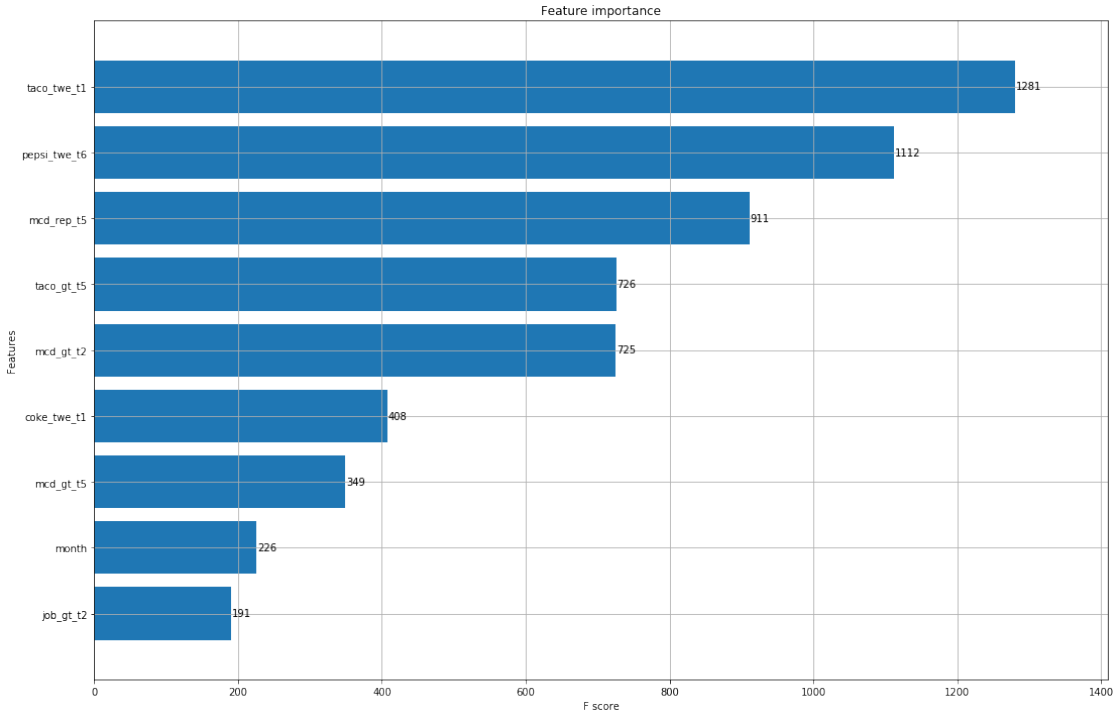

Figure 13: XGBoost Feature Importance

Furthermore, we used the better performing XGBoost model (differenced variables) and conducted a cross validation with a sliding window of fourty-eight months from May 2010 to April 2013. This exercise displayed how predictive power increased as the the sliding window moved forward (Figure 14).

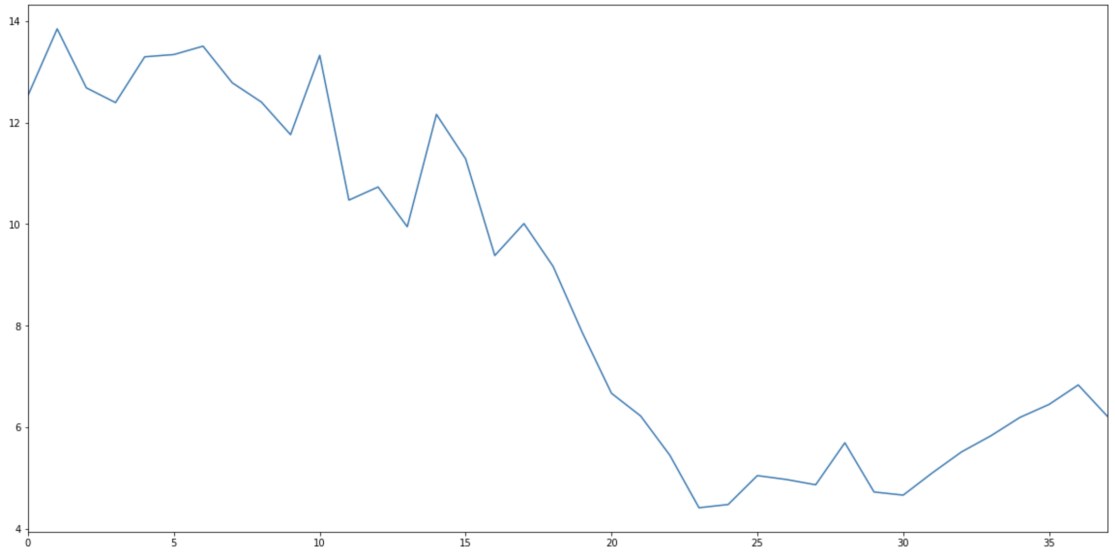

Figure 14: XGBoost sMAPE vs. Sliding Window Iteration

In other words, training the XGBoost model on the early parts of the decade was not very predictive. However, as the decade continues, and presumably when user engagement on Twitter increased, the XGBoost model began to perform better.

Long Short-Term Memory Results

For the LSTM model, we also introduced a month of year variable to allow the model to learn seasonality. Similar to XGBoost, this model produced better performance using the lagged social media variables. The LSTM model using lagged social media variabled produced a sMAPE value of 12.3% and RMSE value of 1,023,904. Among all the challenger models, the LSTM model had the worst performing metrics.

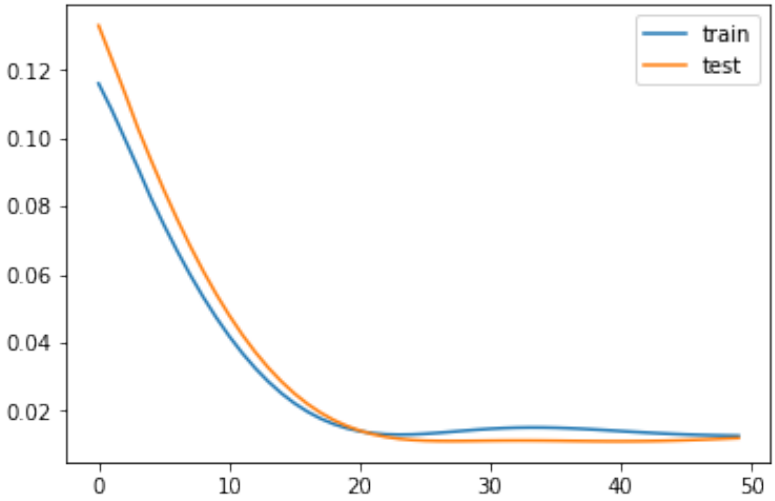

Figure 15: LSTM Epoch vs. Loss Function

Social Media Variable	sMAPE	RMSE	Mean Accuracy	Bias (%)
Differenced Variables	12.30%	1,023,904	87.56%	38.00%
Principal Components	13.46%	1,145,561	86.82%	38.89%

Ensemble Model Results

After building challenger models that utilized social media variables, we constructed ensemble models.

Model	sMAPE	RMSE	Mean Accuracy	Bias (%)
Scholle Baseline Model	7.43%	667,832	92.84%	27.78%

Model	sMAPE	RMSE	Mean Accuracy	Bias (%)
Average of Challenger Models	6.99%	676,394	93.11%	44.00%
Linear Regression	5.69%	542,203	94.25%	50.00%
Random Forest	2.80%	271,128	97.15%	66.00%

Ensembling the three challenger models we built produced more accurate Coca-Cola bag order predictions. The results showed that the random forest ensemble model performs the best according to the sMAPE and RMSE metrics. However, since the linear model produced the most stable results, this will be selected as the champion model. Figure 16 is a visualization of our challenger model forecasts against the actual quantity of Coca-Cola bag orders.

Figure 16: Model Predictions vs. Forecast Window

Residual Analysis

The linear regression ensemble model is selected as the best social media forecasting model to predict Coca-Cola bag orders. We further evaluated the results of this model through residual analysis. The main assumption for a stable model is that the residuals is centered around a mean of zero. The mean of the residuals of the champion model is 35,118. The skewness of the residuals is -0.56, which means that it is left-skewed, which we can see by looking at the residual density plot (Figure 17).

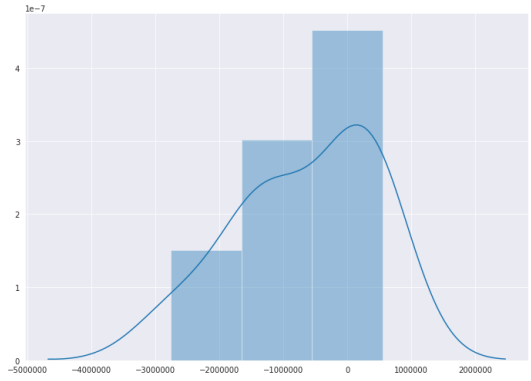

Figure 17: Social Media Ensemble Model Residual Density Function Plot

When looking at an autocorrelation plot of the residuals (Figure 18), we observed that it observed white noise meaning that is uncorrelated, mean of zero, and has a constant variance. This means that there is little additional information that we can draw based from the residuals.

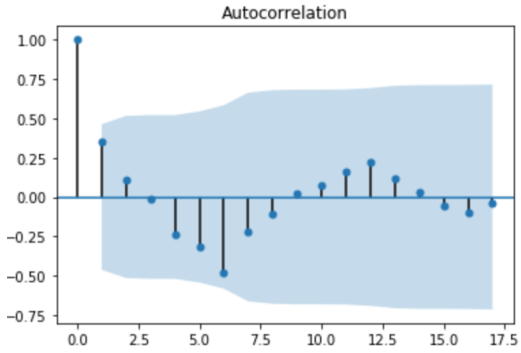

Figure 18: Social Media Ensemble Model Residual ACF Plot

Coca-Cola Ensemble Model Results

To further demonstrate the strength of ensemble modeling, this method was also used to combine models generated by other Scholle IPN Coca-Cola research teams. Other teams were solving the same business problem by 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 best performing ensemble model is the linear regression model with a sMAPE of 1.05% and RMSE of 119,437. A summary of the model's evaluation metrics are outlined below.

Model	sMAPE	RMSE	Mean Accuracy	Bias (%)
Scholle Baseline Model	7.43%	667,832	92.84%	27.78%
Social Media Ensemble	5.76%	545,211	94.22%	50.00%

Model	sMAPE	RMSE	Mean Accuracy	Bias (%)
Linear Regression	1.05%	93,662	98.95%	44.44%
Random Forest	1.34%	119,437	98.64%	50.00%

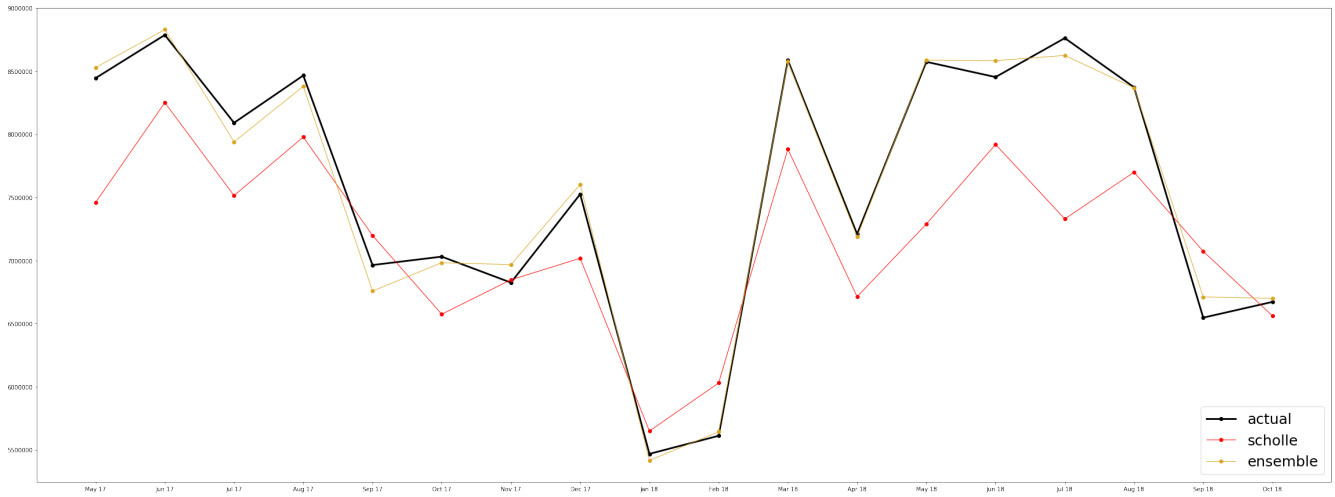

Figure 19: Coke Ensemble Model Predictions vs. Forecast Window

Conclusion

In summary, we learned that social media variables improved Scholle IPN's forecasts of Coca-Cola bag orders. Even though sentiment analysis is an effective way of converting textual social media data into numerical data, we found that sentiment scores are largely dependent upon the text data available. Our user tweet data was not evenly distributed, which led to biased sentiment scores. This rendered these variables to be inconclusive for predicting Coca-Cola bag orders. Additionally, social media variables identified to be cross-correlated with Coca-Cola bag orders improved the XGBoost and LSTM models. According to XGBoost, the top three features were variables derived from company-generated tweets, which were Taco Bell account tweets (t-1), Pepsi account tweets (t-6), and McDonald account replies (t-5). Meanwhile, the regression with Arima errors model experienced improved accuracy when cross-correlated principal component variables were used. Furthermore, using an ensemble approach provided better overall forecasting results. The results from ensembling the social media challenger models (Regression with Arima errors, XGBoost, and LSTM) were considerably better than any of the individual model results. In addition, the ensemble model of the best performing models for each Scholle IPN research teams produced the best overall Coca-Cola forecasting model.

Recommendations

Based on this research, our team recommends the following actions for Scholle IPN:

- Leverage Google Trends and Twitter as primary tools to forecast Coca-Cola bag orders. These are data sources that are easily accessible and provide high returns in predictive power.
- Emphasis should be placed on the collection of data from social media accounts of Coca-Cola's competitors and their partnered quick service restaurants.
- Ensure that social media models are trained using more recent data. Our research indicated that social media data from 2009 to 2013 were less accurate in predicting Coca-Cola bag orders.

Future Work

Our analysis found that the top variables were related to Pepsi, Taco Bell, and McDonald's. Collecting social media data from additional Coca-Cola competitors could present an opportunity to improve the existing forecasting model. These variables could provide incremental gains in predicting more accurate Coca-Cola bag sales. However, caution must be used when adding additional features to model. It is important for C-suite executives to thoroughly understand the factors that impact predictions in order to make sound business decisions. Shapley Values for better interpretability, as it gives the average marginal contribution of a feature value across all possible coalitions (Molnar, 2019).

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.

Cross Correlation Plots

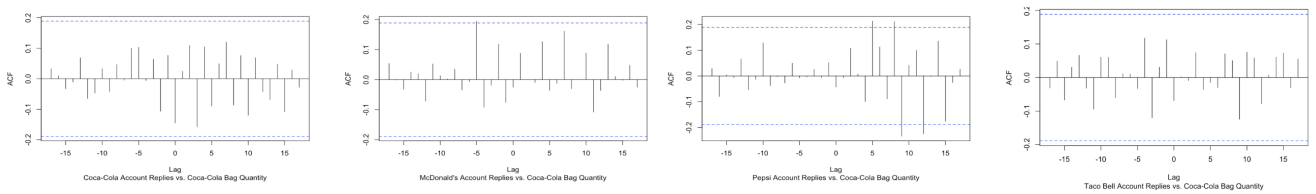

Figure 20: Cross-Correlation Plots A

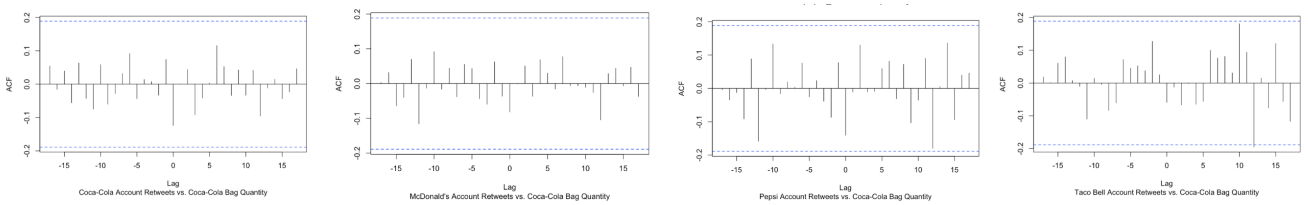

Figure 21: Cross-Correlation Plots B

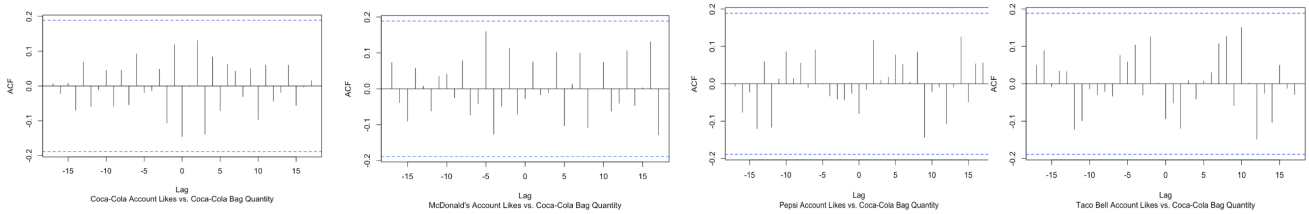

Figure 22: Cross-Correlation Plots C

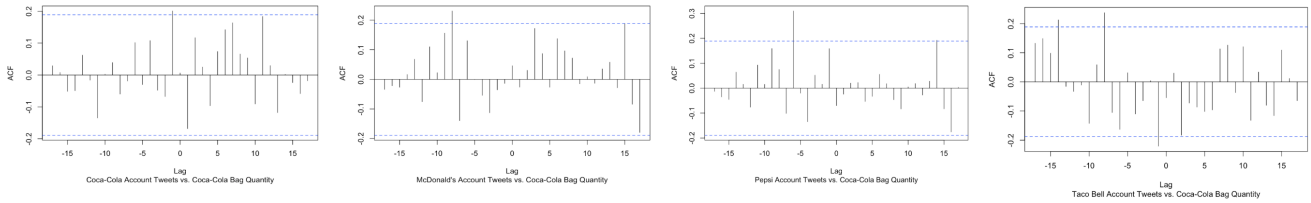

Figure 23: Cross-Correlation Plots D

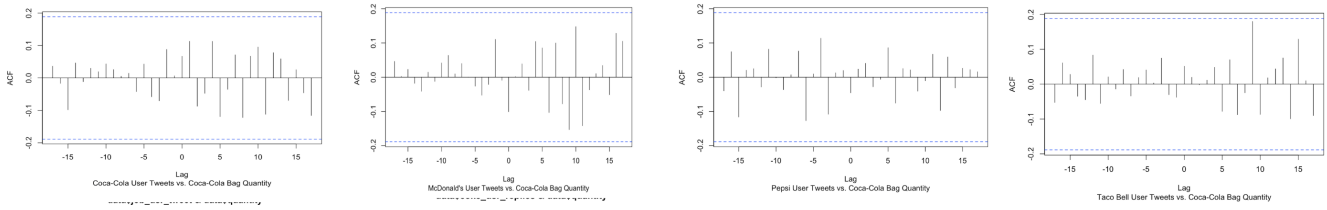

Figure 24: Cross-Correlation Plots E

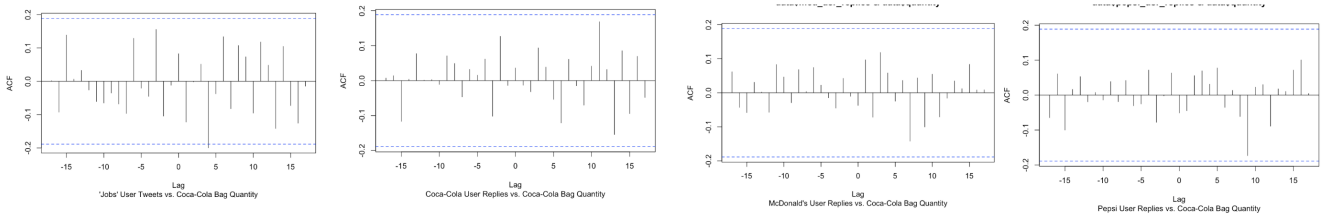

Figure 25: Cross-Correlation Plots F

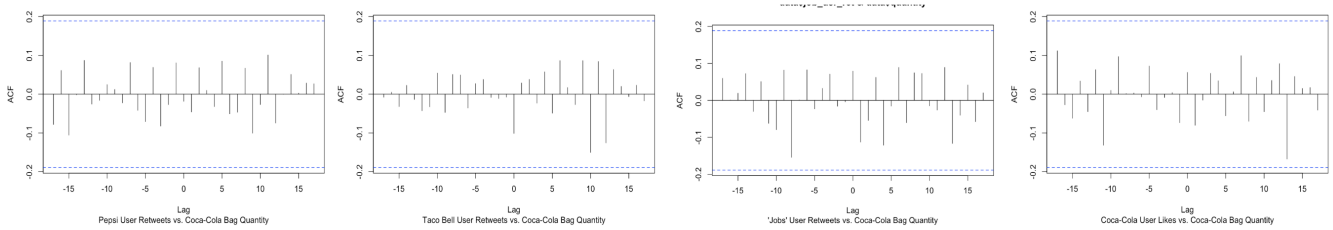

Figure 26: Cross-Correlation Plots G

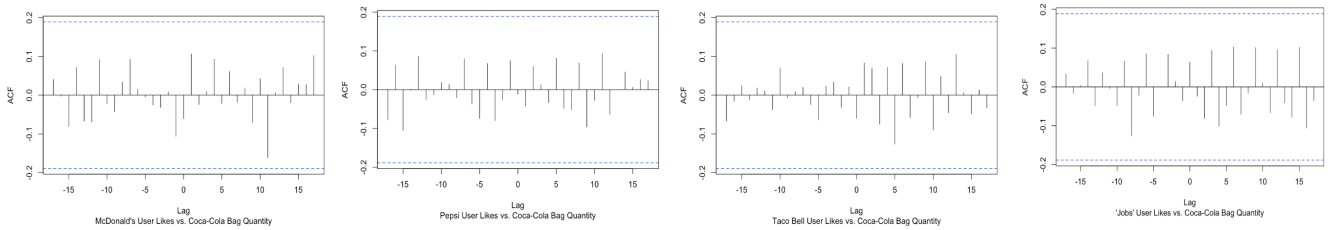

Figure 27: Cross-Correlation Plots H

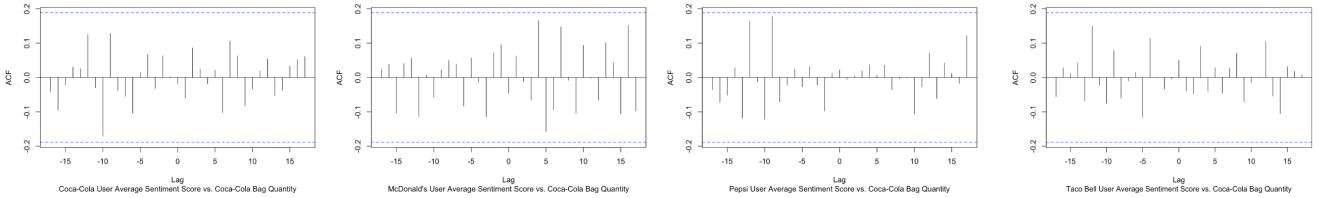

Figure 28: Cross-Correlation Plots I

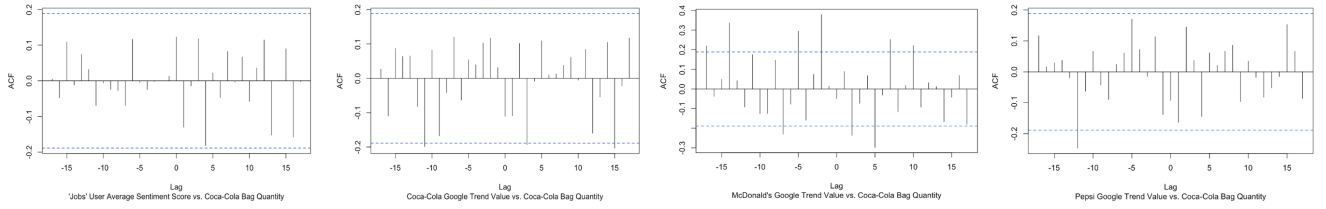

Figure 29: Cross-Correlation Plots J

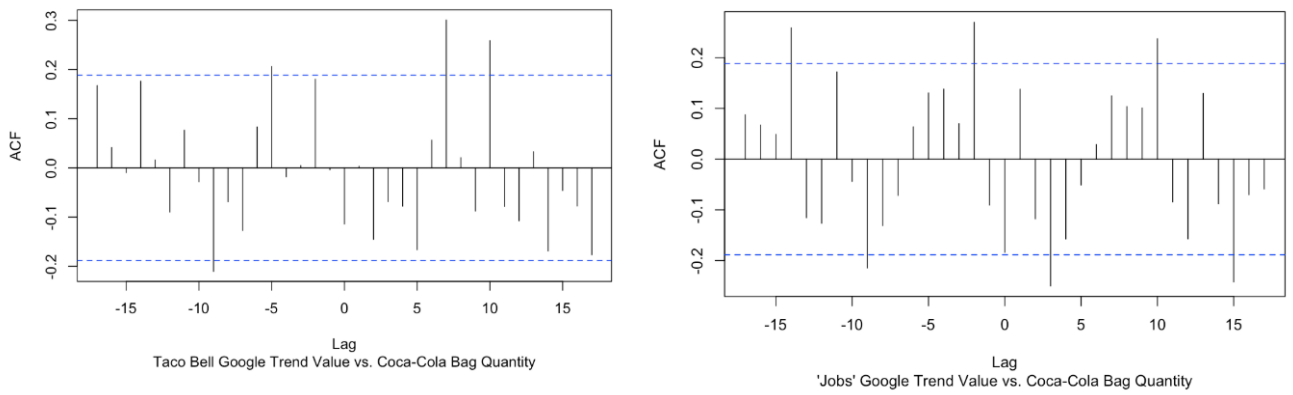

Figure 30: Cross-Correlation Plots K

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 con-

tent, 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.

Brockwell, P. J., & Davis, R. A. (2018). *Time series: Theory and methods, 2nd edition* (pp. 373–375). Springer-Verlag, New York.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Connor, J., Martin, R., & Atlas, L. (1994). Recurrent neural networks and robust time series prediction. *IEEE Transactions on Neural Networks*, *5*(2), 240–254.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (p. Section 9.2). Melbourne, Australia: OTexts.

Kambhatla, N., & Leen, T. K. (1997). Dimension reduction by local principal component analysis. *Neural Computation*, 1493–1516. Retrieved from https://doi.org/10.1162/neco.1997.9.7.1493

Kwiatkowski, D., Peter Schmidt, P. C. P. snd, & YongcheolShin. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 159–178. Retrieved from `https://doi.org/10.1016/0304-4076(92)90104-Y`

Magazine, Q. (2018). The qsr 50. Retrieved from `https://www.qsrmagazine.com/content/qsr50-2018-top-50-chart`

Molnar, C. (2019). *Interpretable machine learning - a guide for making black box models explainable* (pp. Chapter 5, Section8). Springer-Verlag, New York.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *MDPI*. Retrieved from `https://www.mdpi.com/2306-5729/4/1/15/pdf`

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting.