20 – Apr – 2019

PROJECT DELIVERY

Predict Future Sales

Written by:

- Bilal Emad ELDin
- Karim Ashraf
- Ahmed Osman Mohamed
- Mohamed Mamdouh

Submitted to:

Eng. Hussein Fadl

Contact: bilalemadeldin@gmail.com

Table of Contents

Introduction

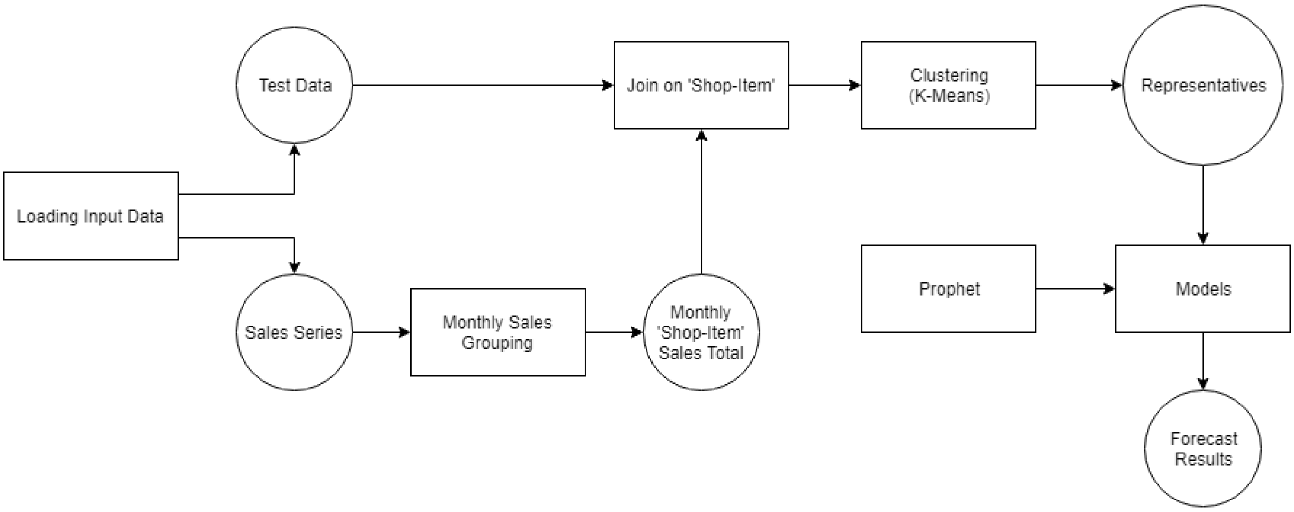
1C, one of the largest Russian software firms, would like to extract an approximate number for future sales depending on past sales data. Providing information needed to build and train a model, 1C expects to get an answer to the question stated in the problem description.

Problem Description

How many products will be sold next month per product and store?

1C has multiple stores all around Russia, using the predicted data it can put strategies to boost sales, restock before outages, and create marketing strategies targeting specific audience or softwares.

Project Pipeline

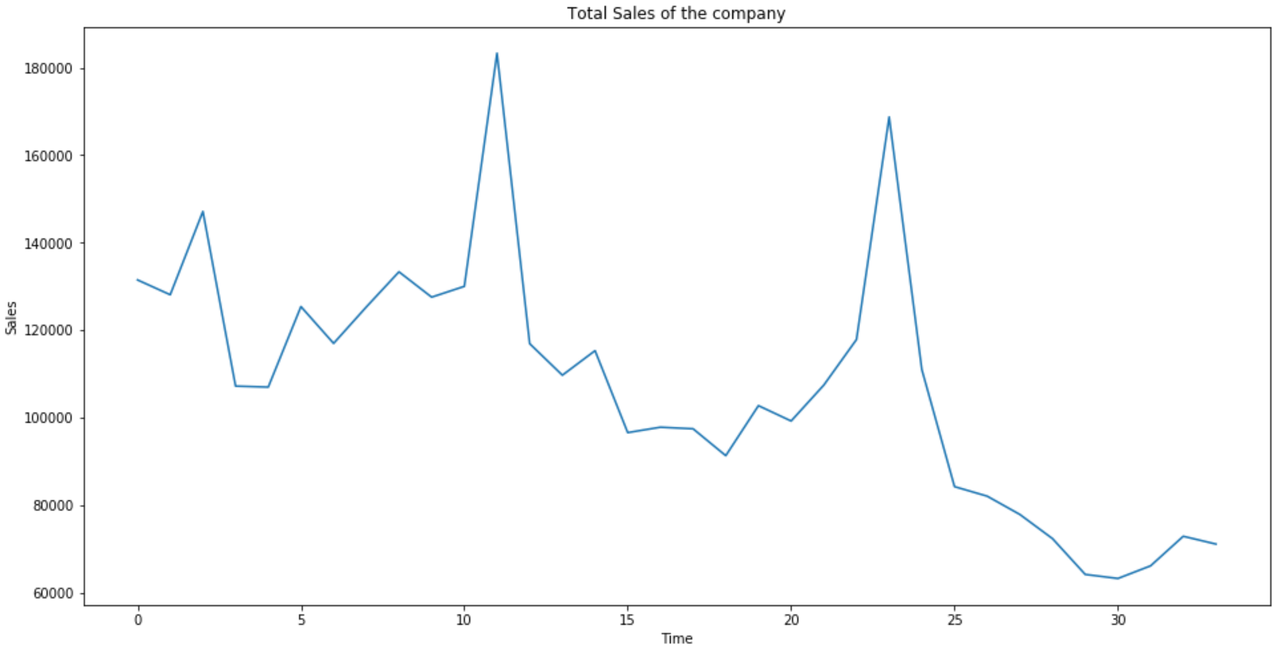


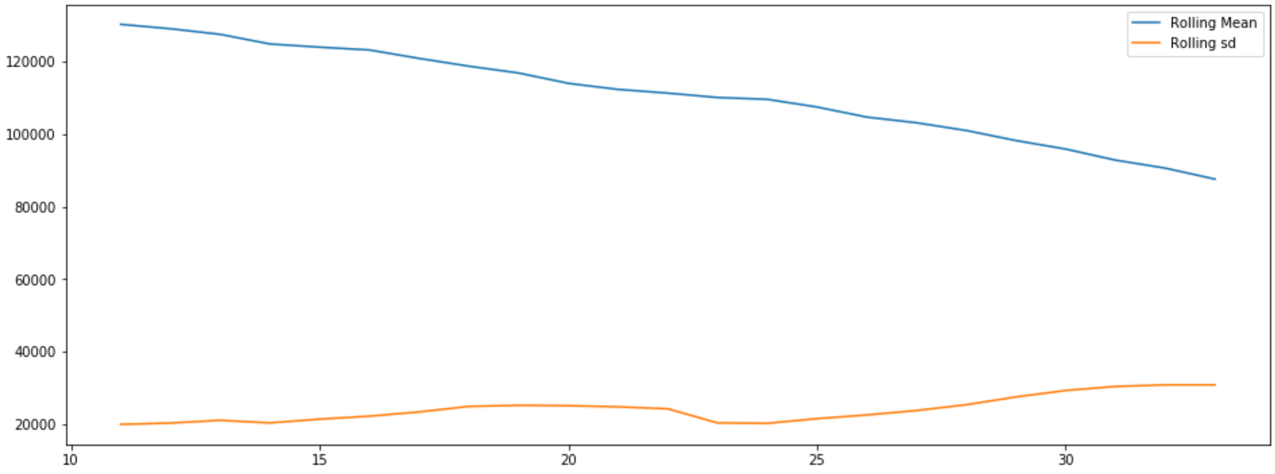
Analysis & Solution

Before we start, we had to analyze the data provided in our competition. We had to check the data to determine what would be useful for our prediction model vs what can be considered extra baggage that would hinder our movement to achieve our desired model. In the end we determined that our main components of data needed would be the dates, shop id, item id, and item count per day. Using this data we started to research ways to construct our model trying different learning algorithms in the process before we achieved our goal.

○ Data visualization

We initially wanted to visualise what the total sales of the whole company was. The result was the following graph.





The noted observations were that the data follows a clear seasonality and has a decreasing trend over time.

○ Data preprocessing

Before we even start working, we faced our first problem when importing our dataset, it was huge. Our training data had nearly 3 million records, reading and parsing this data would've normally taken quite a bit of time to finish. Thankfully we decided to utilize our knowledge of dynamic programming to create a function to help with parsing which eventually took seconds to finish.

○ Model building

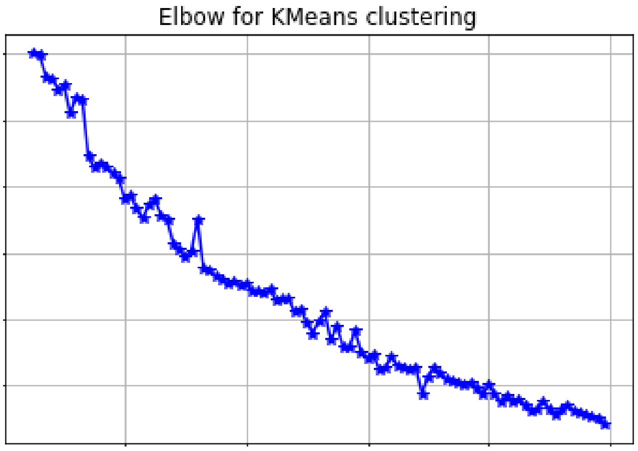
Since the data provided was already split into training data and test data, only splitting for a validation dataset was needed for our processing.

To allow shared work between all team members, we decided to use Colaboratory by google for a real-time shared working environment.

We grouped the data together into total monthly sales indexing them with a ‘Shop-Item’ composite key. This enabled us to get the per-month sales of every combination throughout the 34 months provided in the dataset; which in turn, left us with upwards of a million single time series to model and forecast. So we needed to find a good solution to this.

○ Model training

After finishing our model preprocessing and building, it was time to start our model training. Our first step in training our model was to cluster the data using k-means. We utilized the elbow method we learned in our lecture to determine the optimal number of clusters to use before fitting our grouped data and determining our centroids. This was achieved by running a clustering operation over a wide range of K’s and determining the best, most efficient solution. The result is shown in the graph below.



We chose K=300 to be the optimal number of cluster that provided a good accuracy while being time efficient. All that was left to do is to model a forecasting model for each cluster and assign their forecasts to the Shop-Item combination that fell in each cluster.

This was achieved via Prophet. Prophet is an open-source data analysis and forecasting library made by Facebook that makes forecasting single time series easy and accurate. Prophet follows sklearn syntax and works using 4 main components at its core:

- A piecewise linear or logistic growth curve trend. Prophet automatically detects changes in trends by selecting changepoints from the data
- A yearly seasonal component modeled using Fourier series
- A weekly seasonal component using dummy variables
- A user-provided list of important holidays

● Results and Evaluation

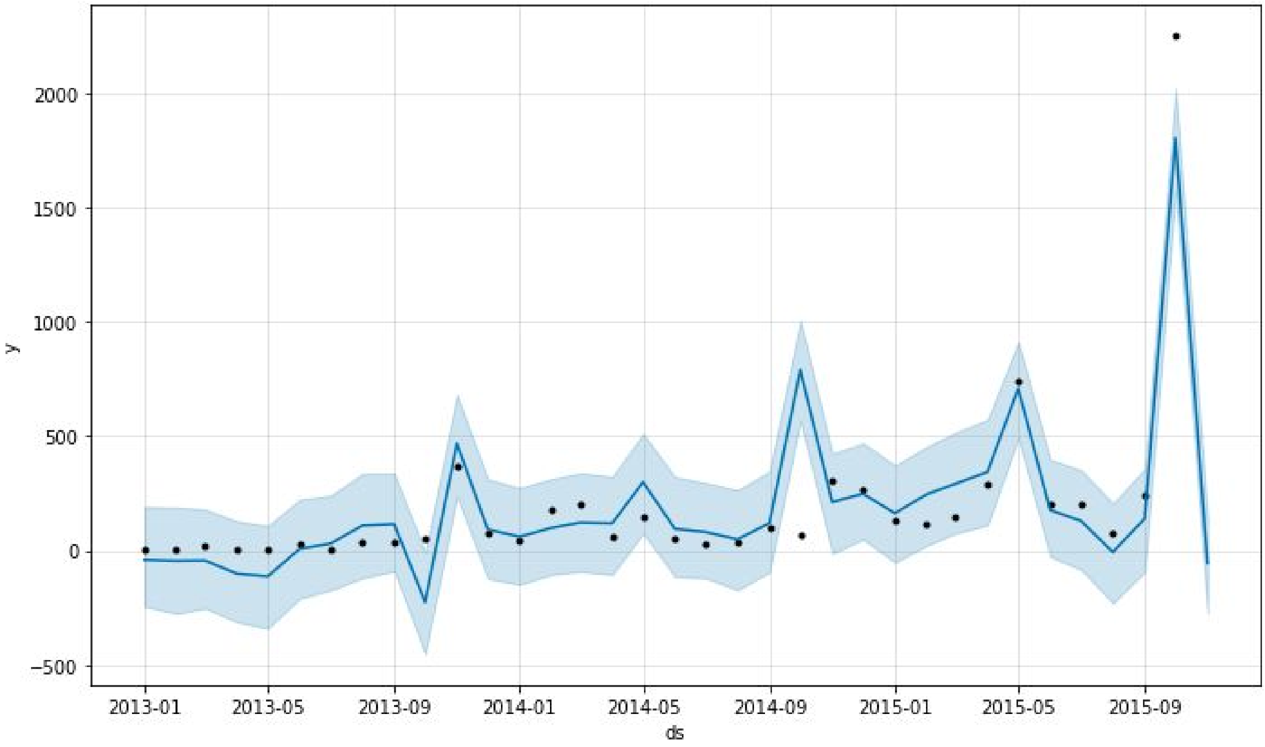
After we were done training our model, Testing and validation was the next step of utmost importance. To start we had to use some of our training data for validation as mentioned earlier, for this we researched and found that a variance of percentages of data can be used when creating a validation dataset. In our case, we decided to take 5% of our data randomly for validation purposes. This data then helped us measure the accuracy of our model which turned out to be 86%.

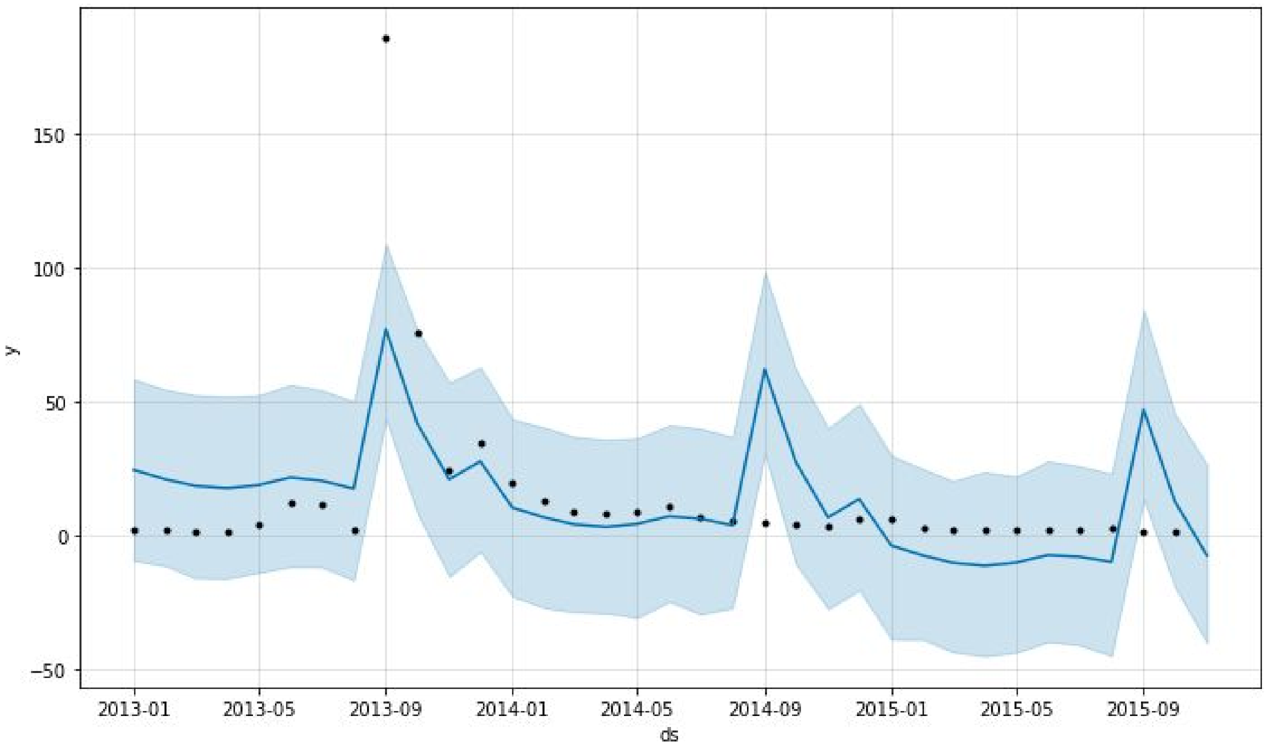
The results was 300 forecasts for 300 clusters. Since each ‘Shop-Item’ combination fell in one of these clusters, they each had one of the 300 forecasts as the next month’s predicted sales. A Sample of the results were as following:

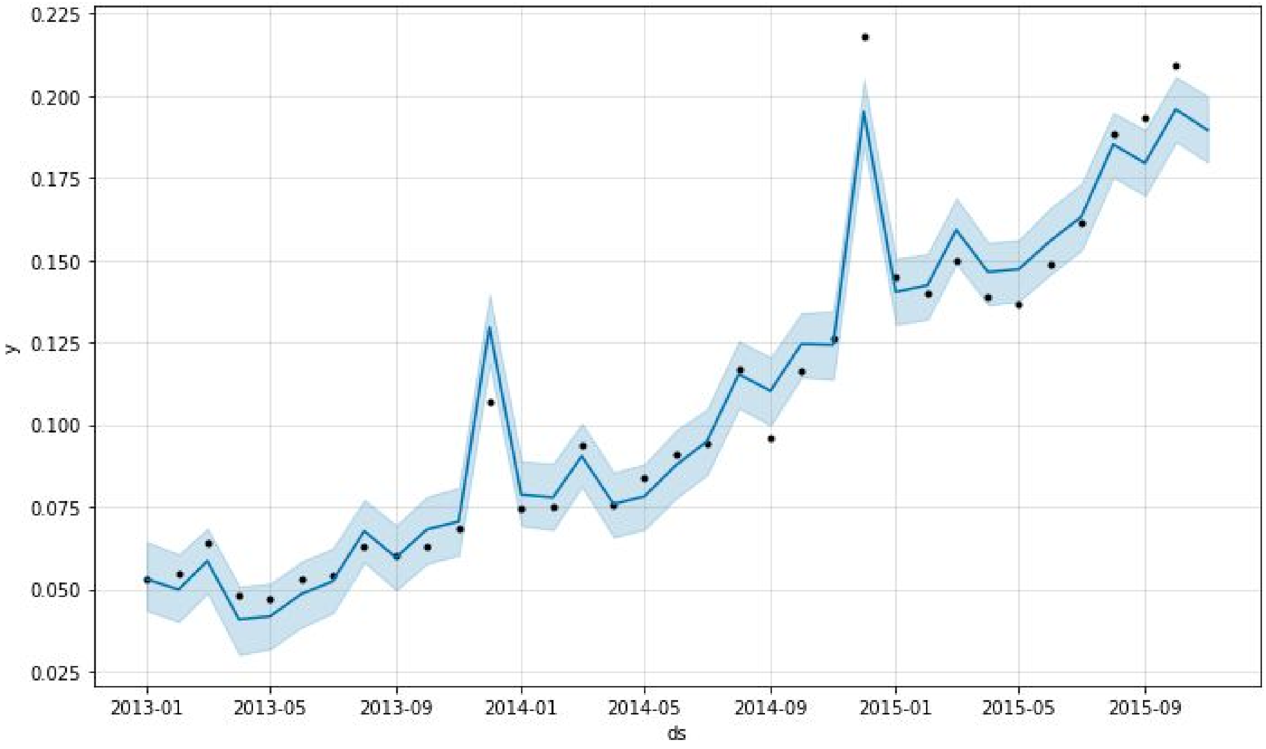
Cluster	Next Month’s Forecast
0	**-53.95**
1	**-7.33**
2	**0.19**
3	**3.59**

4	**12.09**
5	**14.99**
6	**95.96**
7	**114.25**
8	**122.39**
9	**162.15**
10	**498.82**
11	**601.30**
12	**691.51**
13	**877.15**

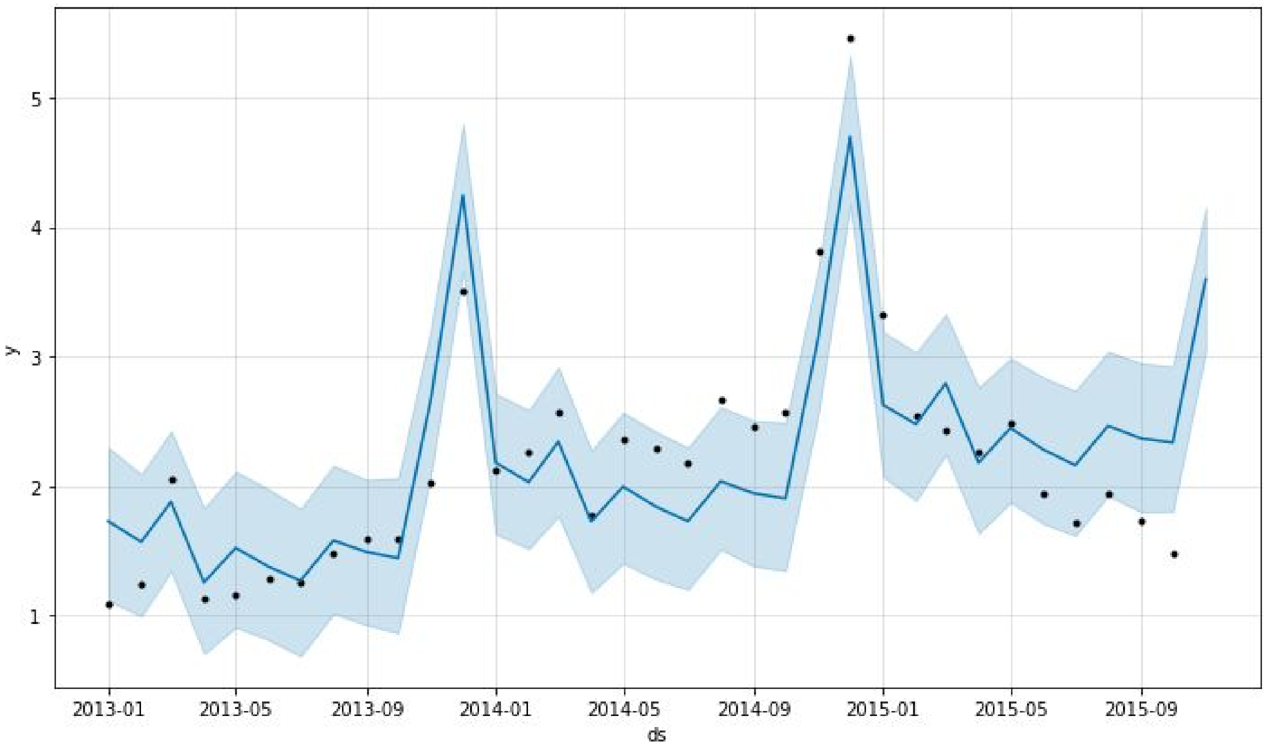
The detailed data is visualised in the following graphs. Graphs are shown respectively.

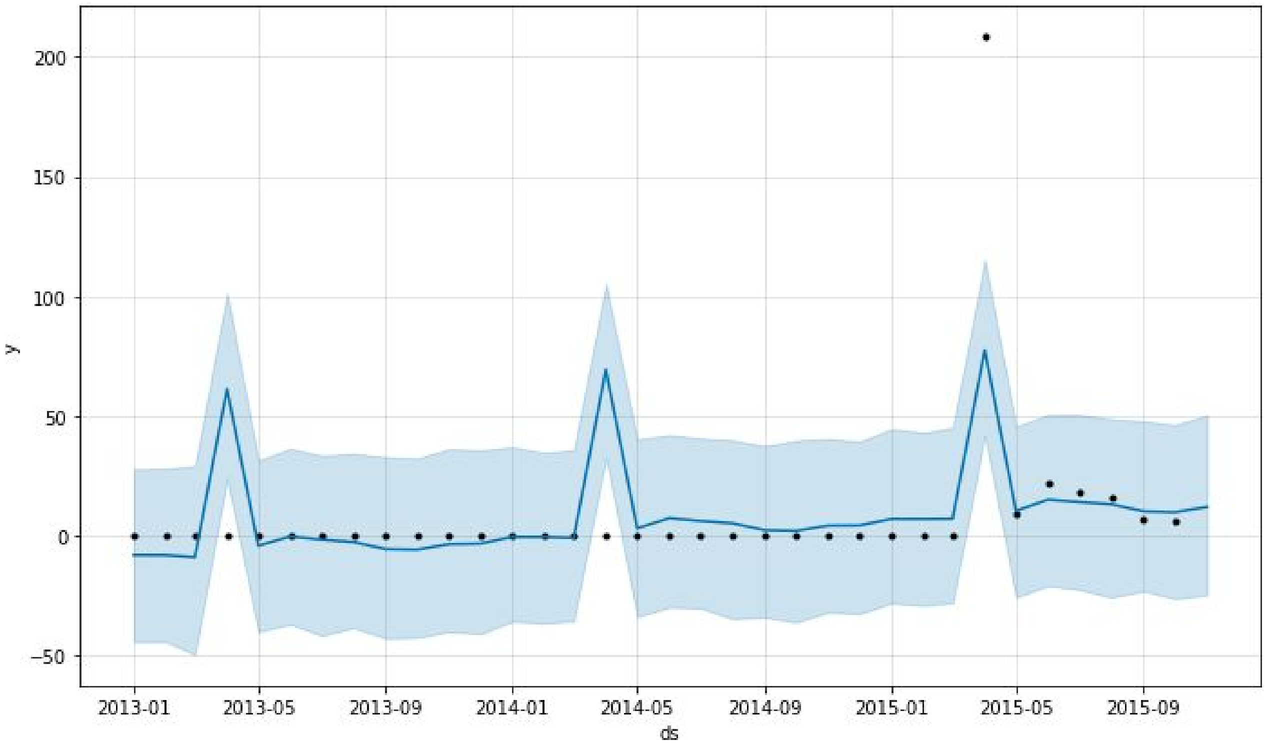




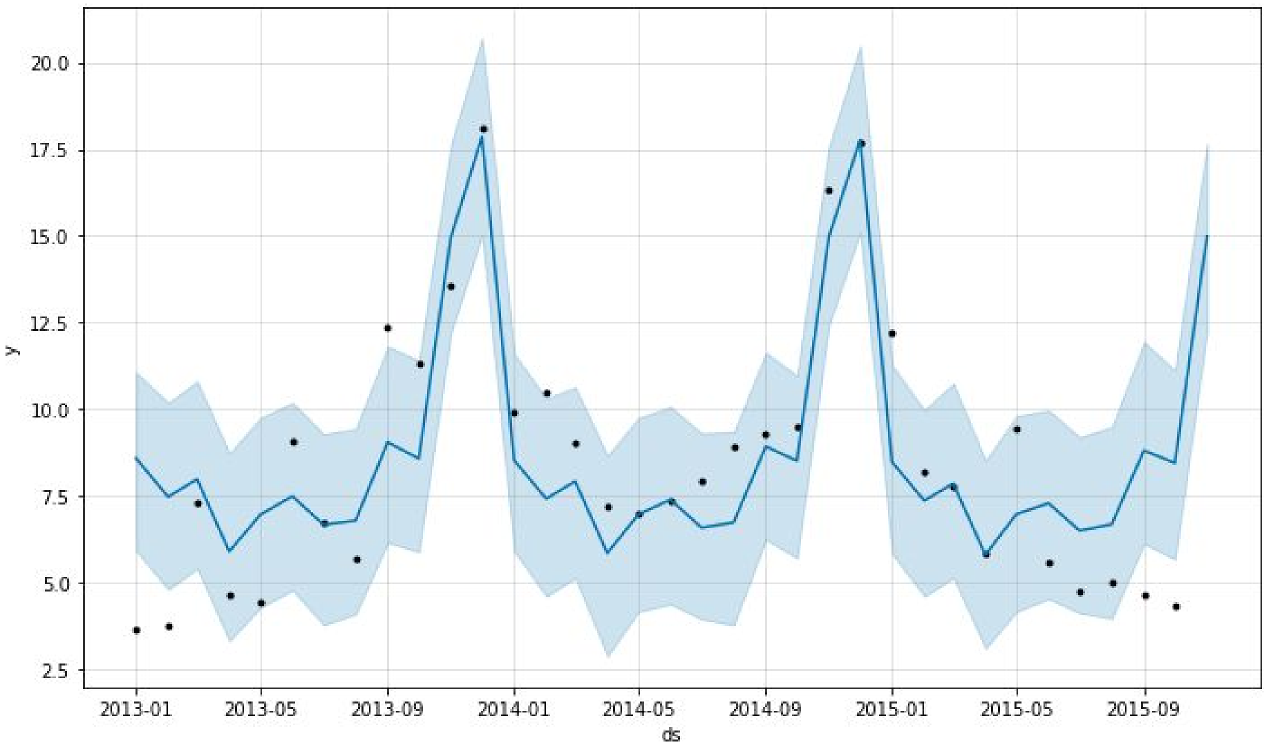


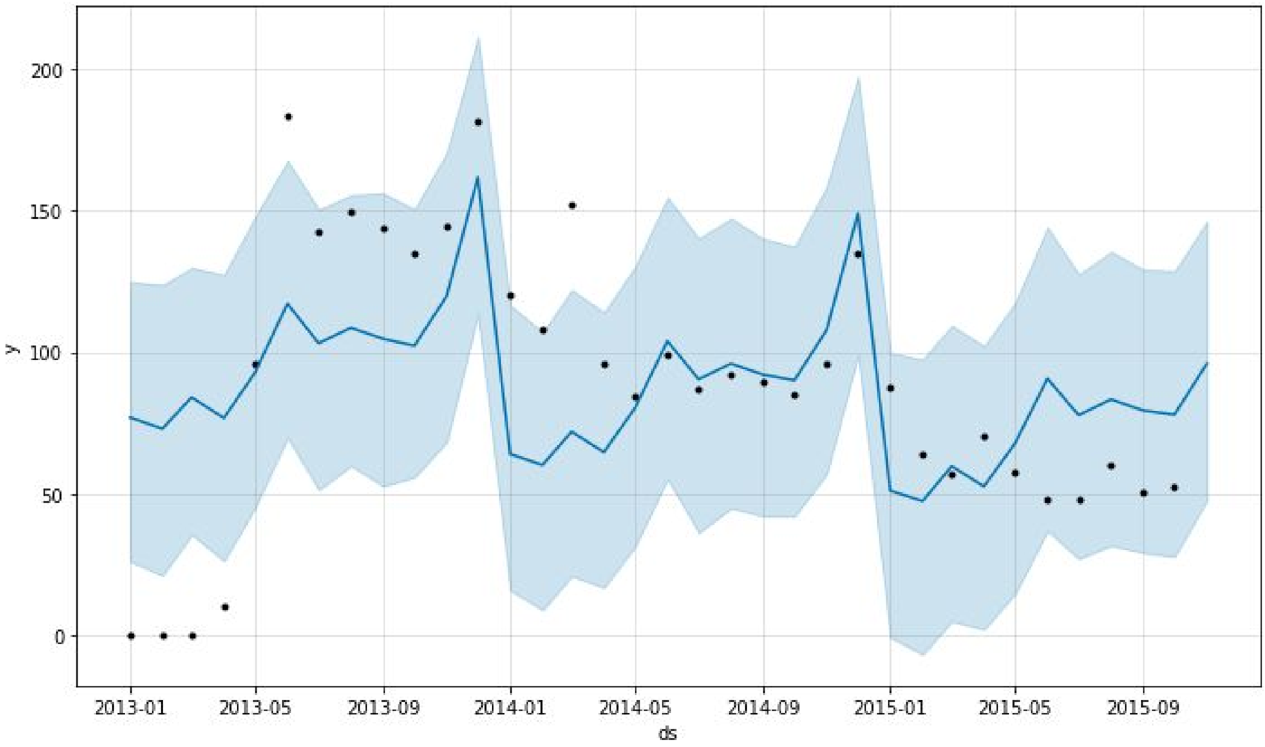
Contact: bilalemadeldin@gmail.com



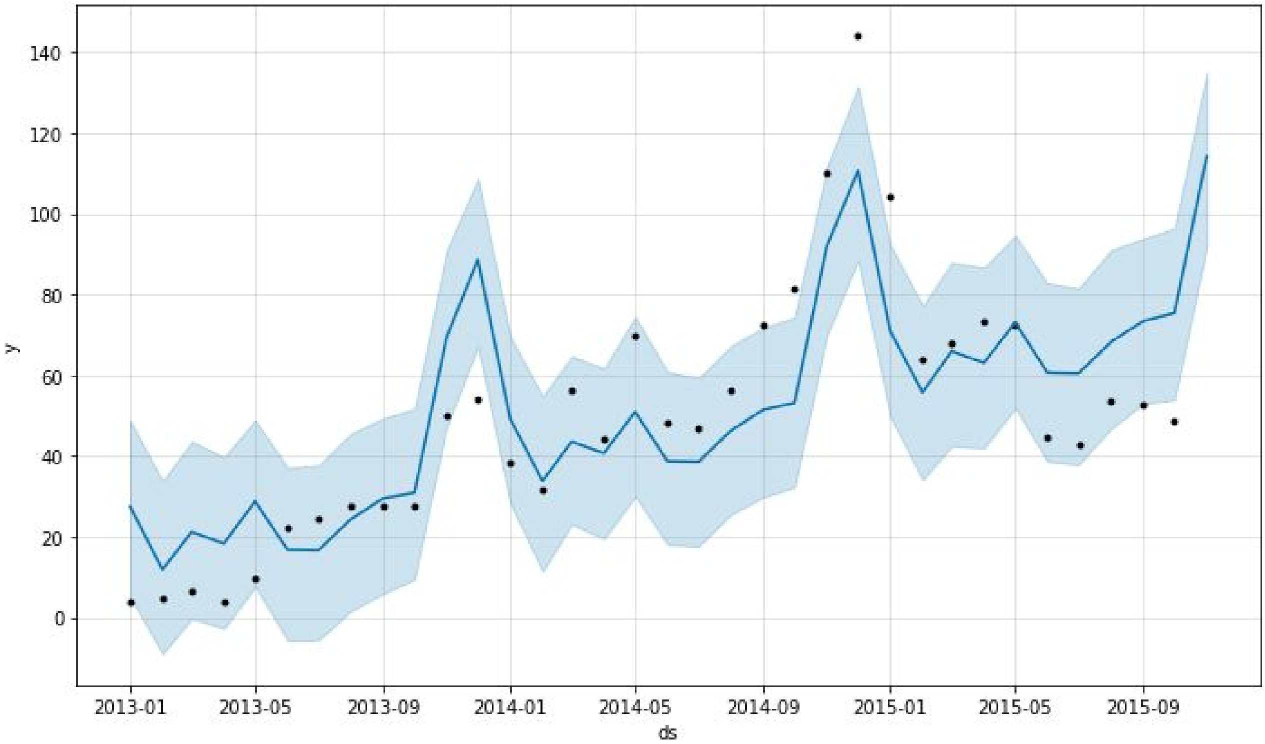


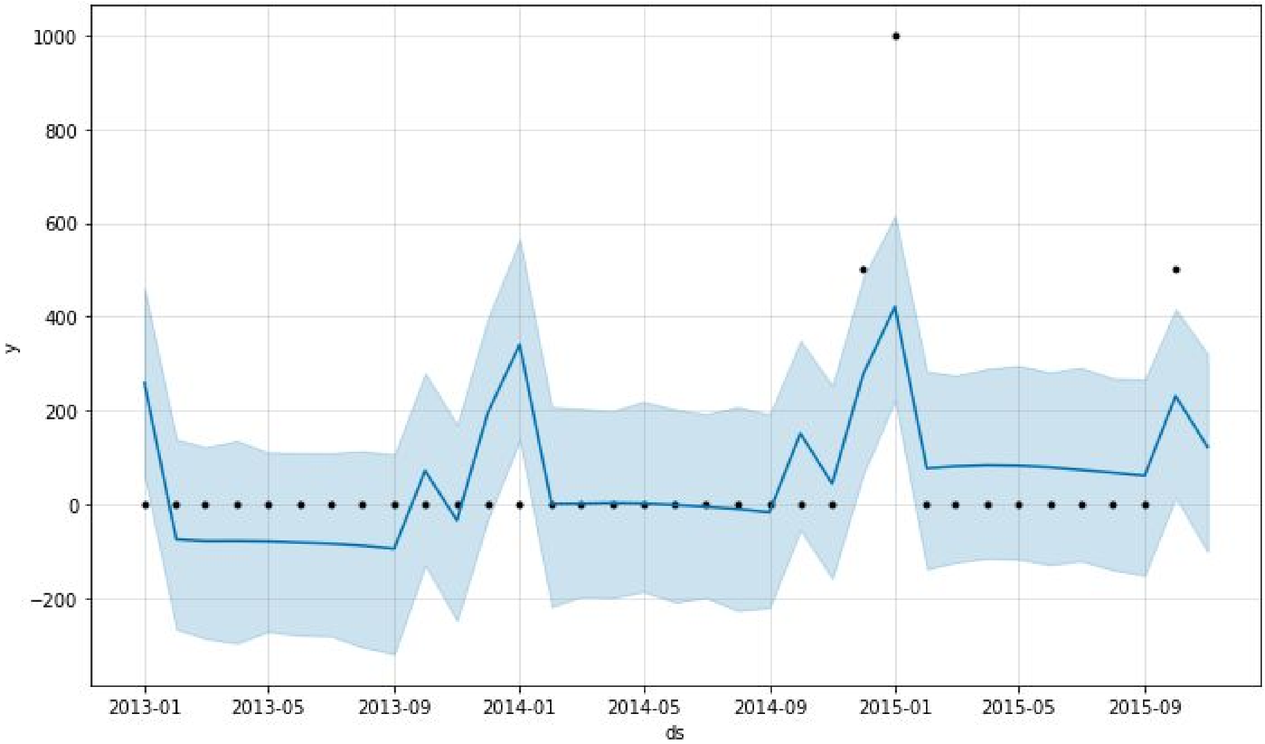
Contact: bilalemadeldin@gmail.com



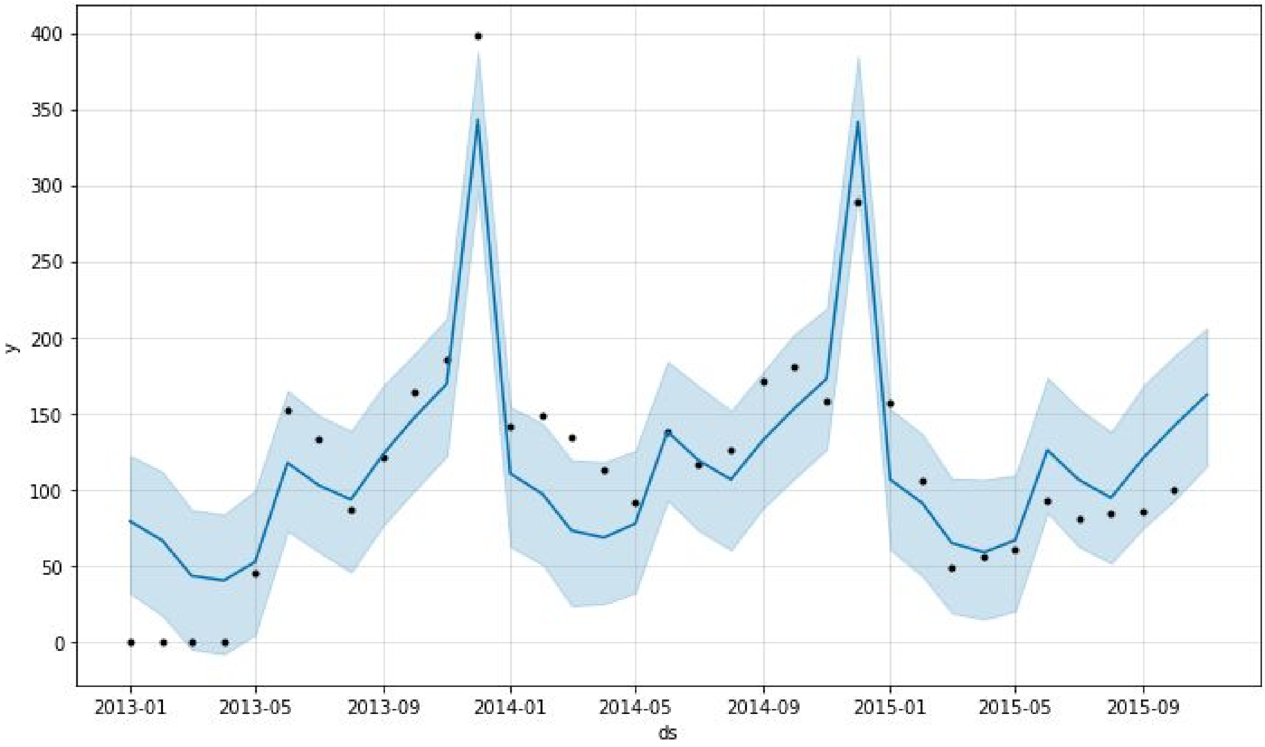


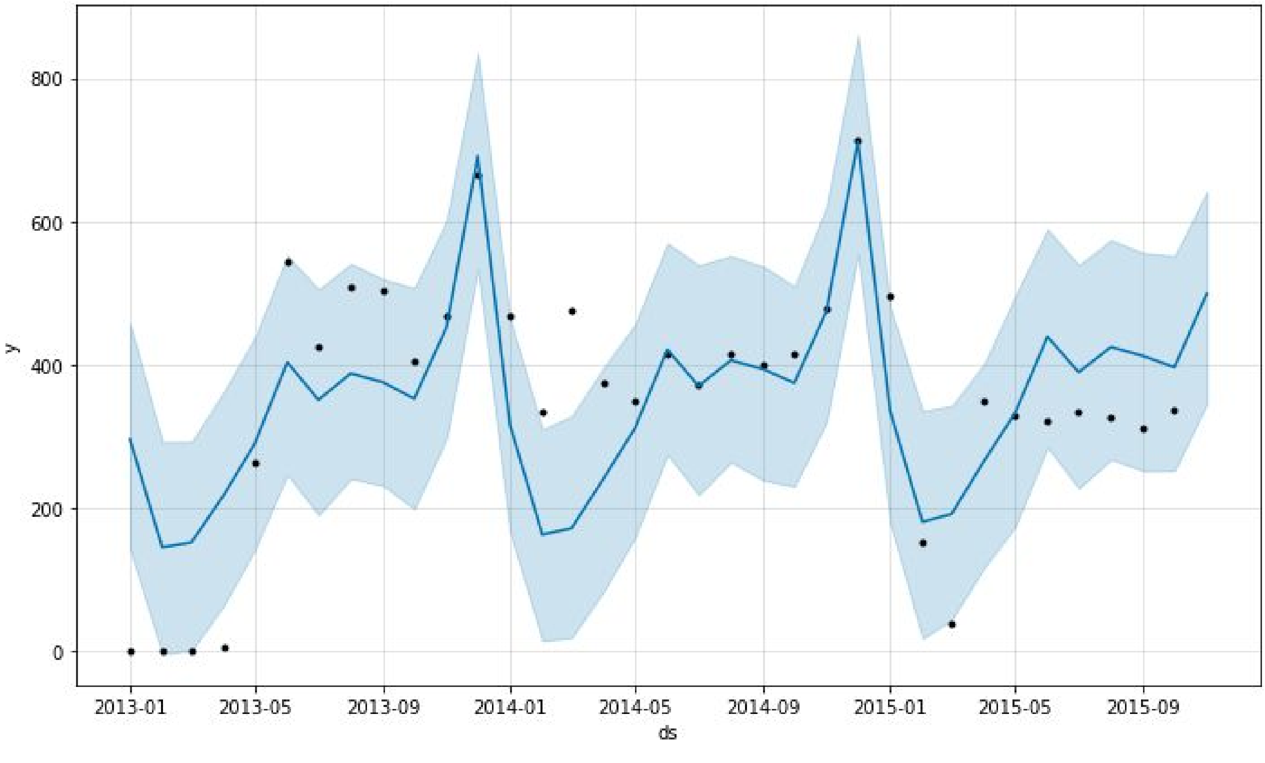
Contact: bilalemadeldin@gmail.com



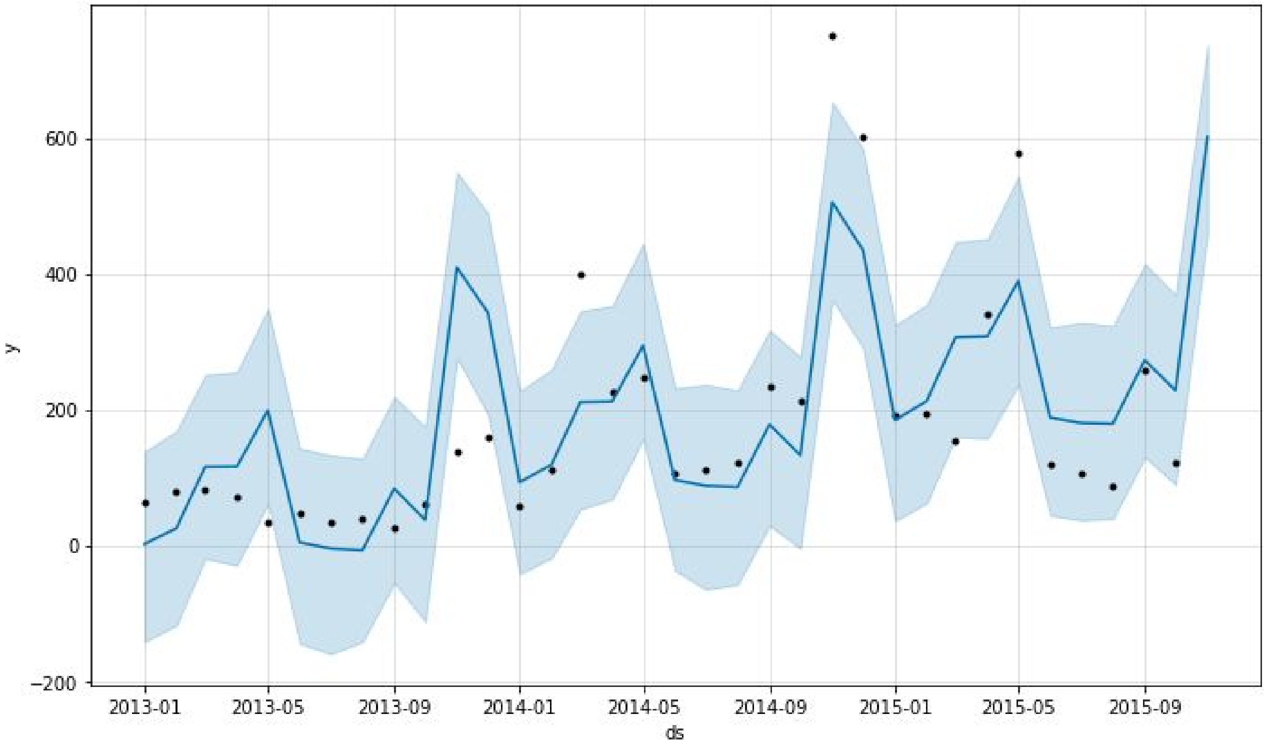


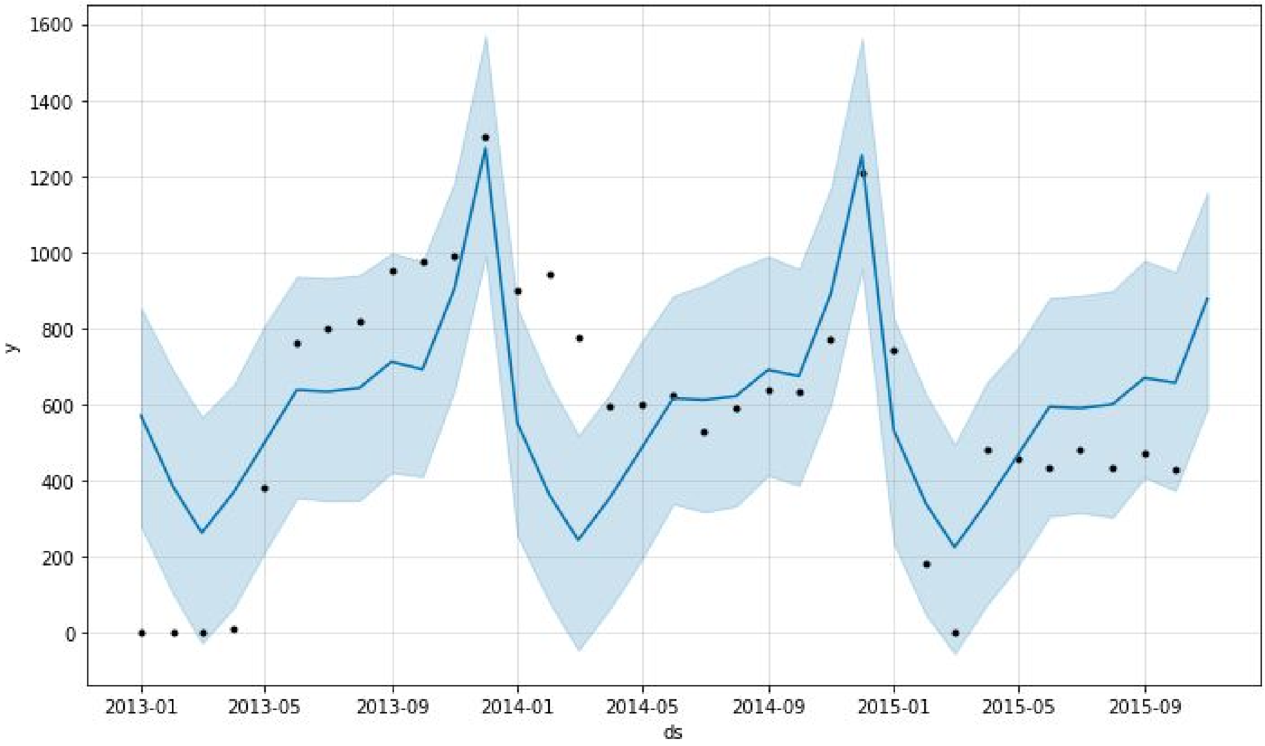
Contact: bilalemadeldin@gmail.com





Contact: bilalemadeldin@gmail.com





The most important part of our business insight comes from these forecasts.

Not only do they guide the marketing campaigns of the company, they also predict the amount of consumption of each item so that the company can make sure their stock is capable of supporting the load.

Contact: bilalemadeldin@gmail.com

In the above forecasts, both clusters 0 & 1 suffer from a negative sale rate. The company can use this information to increase the advertisements for these products or add enhancements to them.

Same with the clusters 10-13. They all have a very high forecast value. This can be useful to the company to boost the sales of other items by piggy backing them on these products.

● Unsuccessful trials

Our first approach was using traditional ARIMA time series forecasting. This was deemed inappropriate since we would have to calculate the parameters for over a million series generated from the Shop-Item combination.

Another solution instead of clustering the data would have been modelling each time series on Prophet and forecasting a different value for each Shop-Item. This however proved to take an extremely long amount of time. The exact number of series we needed to model was 214200. Our collaboratory was able to train 100 models in close to 10 mins. At this rate, training a model for each combination would have taken over two weeks. Thus we came up with the clustering solution.

● Enhancements and future work

Given enough time as mentioned above, we could train a customised model for each shop-item combination which would result in a much more accurate result.

For the future work, additional sales data can be collected to increase the volume of the dataset. At a certain point, a deep learning approach would be a much better solution than an analytical approach, especially with this huge volume of data. Deep learning would be able to find patterns that normal approaches fail to see. It also provides us with an incremental learning option to improve the model day by day.

● Conclusion

By the time we were done, we had learned a lot about our dataset, how time series forecasting works more deeply, and how we can utilize what we learned during this semester in real life along with integrating it with both computer related and unrelated aspects for future analysis. A corporate like 1C can take the info provided here, along with the code for any future improvements by their analysts and the presentation to form a strategy that will help predict future sales, determine low selling items, boost sales, and create a marketing strategies to generate higher profit through appropriate stocking and grouping of merchandise.

● References

Kaggle Competition:

https://www.kaggle.com/c/competitive-data-science-predict-future-sales

Dataset:

https://www.kaggle.com/c/8587/download-all

Prophet:

https://facebook.github.io/prophet