**CMPE 256 - Project Report Template**

* Report should be prepared using the [IEEE proceedings template](https://www.ieee.org/conferences/publishing/templates.html) (you may use word or latex)
* Report **should not exceed 8 double-column pages** (Less is fine, just make sure to include all requested details below. Additional details if needed can be included in an Appendix. However, the data in the appendix will not contribute to/against your final grade and should be only informative.)
* You should update/adjust the subsections of the template below according to your project’s focus and details.

**Ch.1 Introduction**

* **Motivation**
* **Objectives**

1. Basic data mining tasks for data preprocessing, data analysis and feature

engineering

2. For time series forecasting, we might use either Prophet (Time Series forecasting

algorithm from Facebook) or ARIMA (Auto Regressive Integrated Moving

Average) based on which model gives better performance.

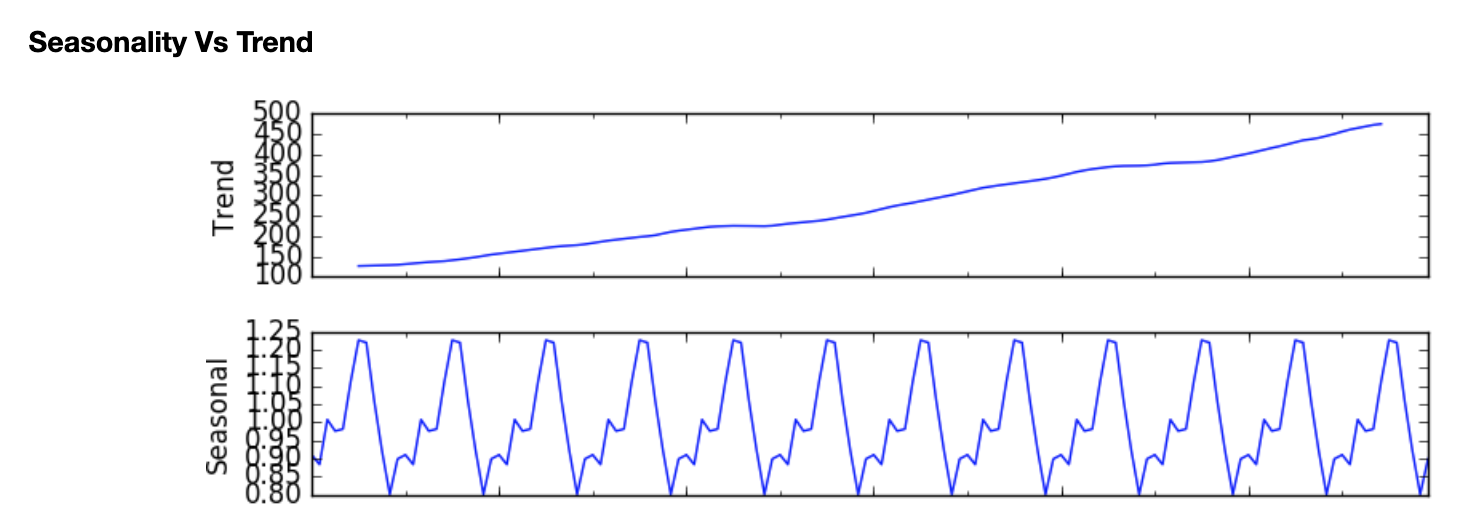
3. Classification model to predict whether the route is accident prone

**Ch.2 System Design & Implementation details**

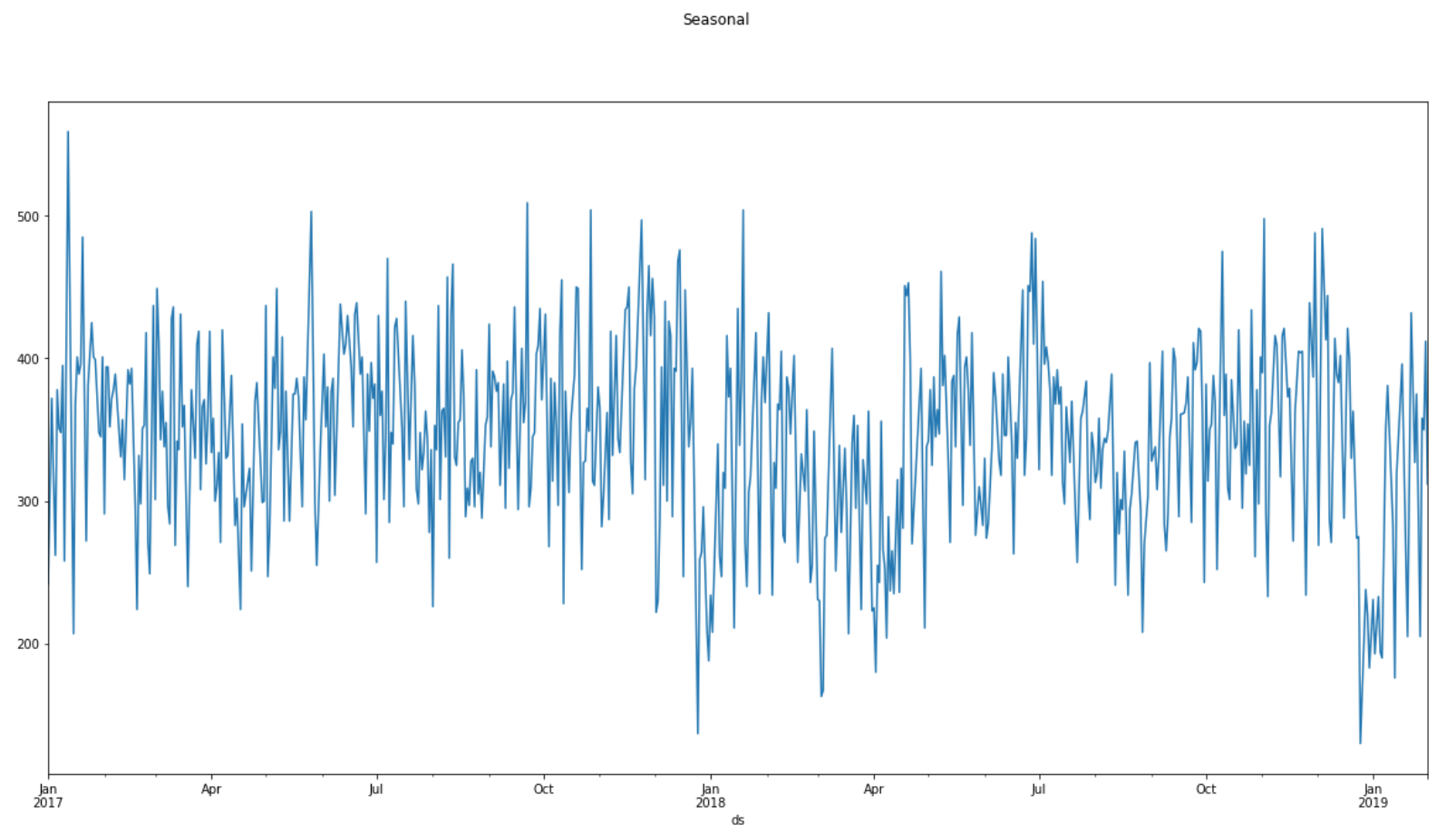
* Algorithm(s) considered/selected (and why)

**Time Series Forecasting:**

|  |  |
| --- | --- |
| **Algorithm** | **Reason** |
| **SARIMAX** | ARIMA, which stands for the AutoRegressive Integrated Moving Average model, is a mixture of two model forms, namely the AutoRegressive model and the Moving Average model, and the differentiation method. ARIMA and related models presume some kind of causal link between past values and past failures and future time series values. Sometimes, more configuration is required for algorithms such as ARIMA to achieve realistic results, which is out of the scope of many people who are not experts. We wanted to explore. Seasonality and exogenous influences (seasonality and normal ARIMA do not blend well) are the contrast between ARIMA and SARIMAX. We tested the model with both simple ARIMAX and SARIMAX, SARIMAX performed better for our data. The reason for this is that our dataset possess Seasonal characteristics. |
| **Prophet**  ***(Selected)*** | Facebook Prophet is not searching for any causal links between the past and the future. Instead, it simply attempts to use a linear or logistic curve and Fourier coefficients for the seasonal components to find the best curve to match the results. There is also an aspect of regression, but it is for external regressors, not the time series itself. The regressors help in adding additional features while model building. In our application, we needed a model to understand the effect of features such as ‘Road Type’ was important.  By design, Prophet can provide the trend variable with uncertainty intervals by simulating potential trend shifts in your time series. You should run a few hundred iterations (which takes a few minutes) if you want to model uncertainty regarding future seasonality or holiday impact, and your predictions may provide projections of seasonal uncertainty. Prophet provides an option to set strong multiple “human-scale” seasonalities: day of week and time of year. |

****

**Plot for our dataset:**

****

**Classification:**

* Technologies & Tools used (and why)

1. Pandas

Pandas is basically utilized for data analysis. Pandas permits bringing in information from different files and we have used it to read data from csv files. We also performed different data manipulation operations such as reshaping, merging, as well as data cleaning.

1. Numpy

Numpy gives supporting capacities that make working with ndarray exceptionally simple.

We used it while encoding the features acceptable to our model. Arrays are regularly utilized in data science, where speed is very critical.

1. matplotlib.pyplot and seaborn

We made use of visualizations while exploring the different data files - Accident, Casualties, Vehicles. Various plotting options provided by matplotlib and seaborn helped us understand the relationship between features.

**Time Series Forecasting:**

1. statsmodels.api

Statsmodels is a Python module which provides classes and functions for many different statistical models to be calculated, statistical tests to be carried out, and statistical data exploration to be carried out.

Statsmodels.tsa(part of statsmodels.api) includes groups and functions of the algorithm that are useful for study of time series. Univariate autoregressive (AR), vector autoregressive (VAR) and univariate autoregressive moving average (ARMA) models are simple models.

1. fbprophet

Prophet is a method focused on an additive model for forecasting time series data where non-linear patterns match with annual , monthly, and regular seasonality, plus holiday results. It fits well for time series that have clear seasonal influences and historical evidence from many seasons. Prophet is immune to lost data and pattern fluctuations, and usually manages outliers well.

Prophet is open source software published by the Core Data Science team of Facebook.

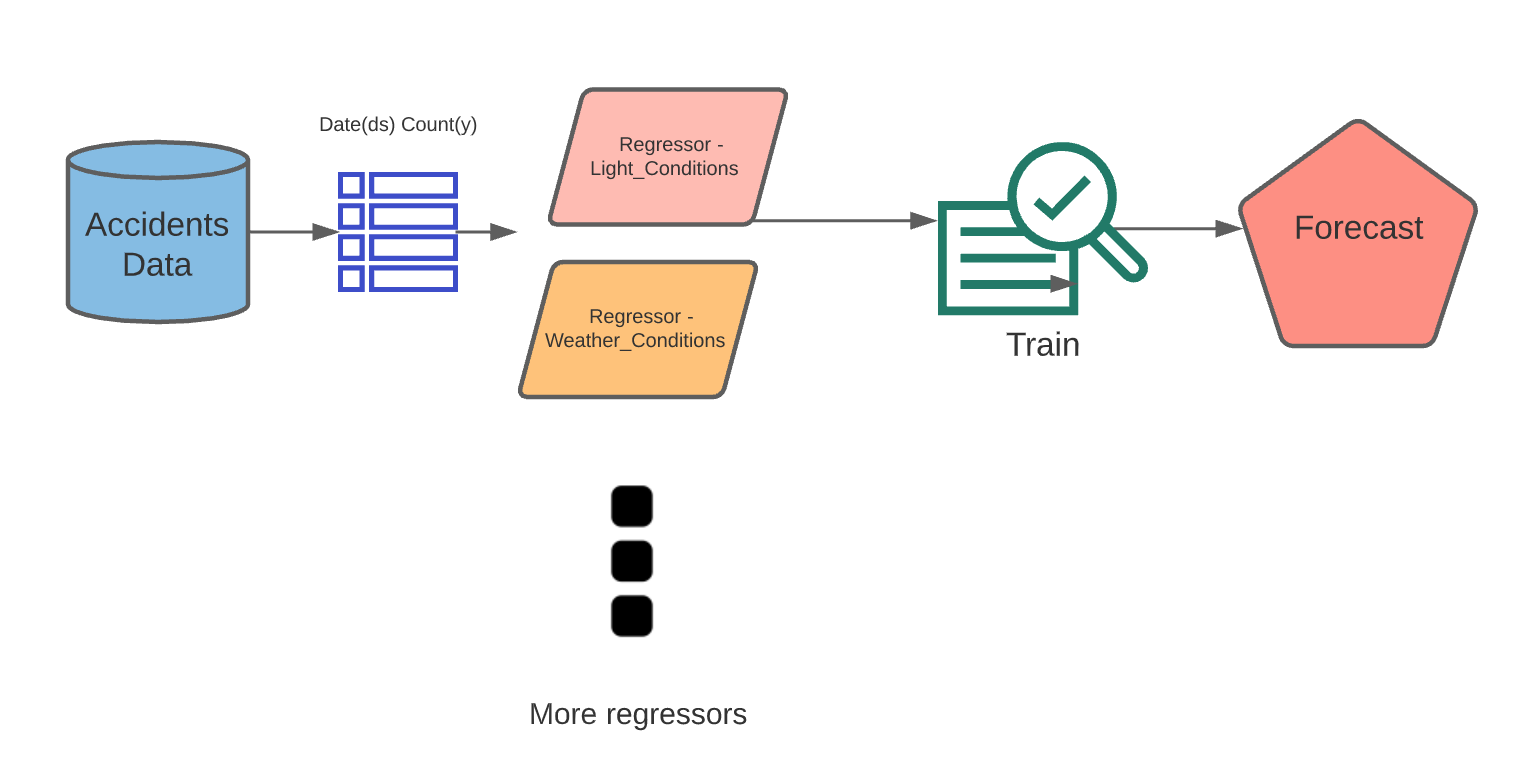
“Facebook recommends using Prophet for forecasting with data that has the following characteristics:

* hourly, daily, or weekly observations with at least a few months (preferably a year) of history
* strong multiple “human-scale” seasonalities: day of week and time of year
* important holidays that occur at irregular intervals that are known in advance (e.g. the Super Bowl)
* a reasonable number of missing observations or large outliers” [3]

**Classification:**

* System design/architecture/data flow/workflows as applicable (you may use diagrams with some supportive text)

**Time Series Forecasting:**



**Classification:**

* Visualization/UI/ GUI/screenshots (only if applicable)

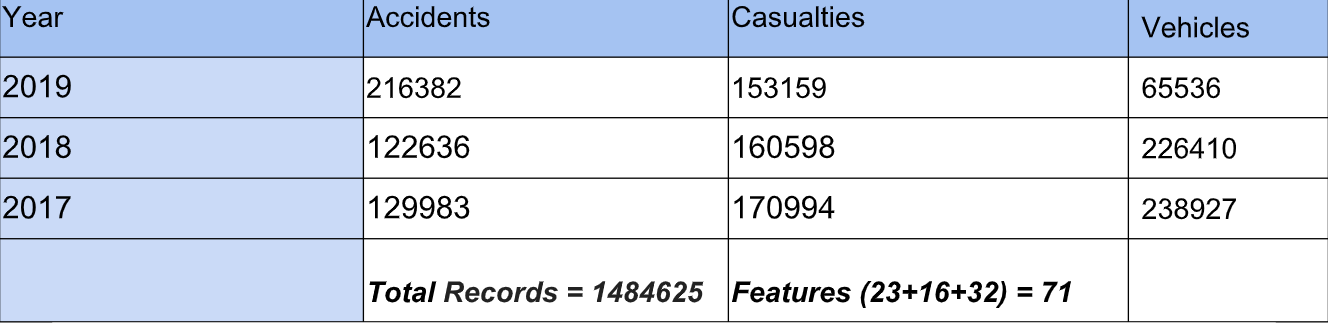
**Ch.3 Experiments / Proof of concept evaluation**

(This chapter will differ depending on what has been proposed and discussed between you and your instructor)

* Dataset

The dataset consists of accident and vehicle information for the year 2019

There are three different CSV files with following number of records and features:



The features in these three files have been referenced using a common

“Accident\_Index” column. The dataset also contains a variable look-up excel sheet

which provides information about various features in the three CSVs.

The dataset is available on UK’s government website: <https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data>

* Data preprocessing decisions

**Classification:**

* Feature Selection

**Time Series Forecasting:**

After the initial data exploration, we found out the important contributors to frequency of accidents. Furthermore, finding correlations between different features helped us narrow down the input features for the model.

Following are the selected features for Time series forecasting model :

*Day\_of\_Week*

*Light\_Conditions*

*Weather\_Conditions*

*Speed\_limit*

*Pedestrian\_Crossing-Physical\_Facilities*

*Special\_Conditions\_at\_Site*

*Number\_of\_Vehicles*

*Police\_Force*

*2nd\_Road\_Class*

**Classification**

* Evaluation methodology followed (e.g. n-fold-cross validation, number of n, size of training/test set etc.)

**Time Series Forecasting:**

We split the data by date since time series need the date in appropriate order.

Training Set - 01-01-2017 to 31-01-2019

Test Set - 01-02-2019 to 31-12-2019

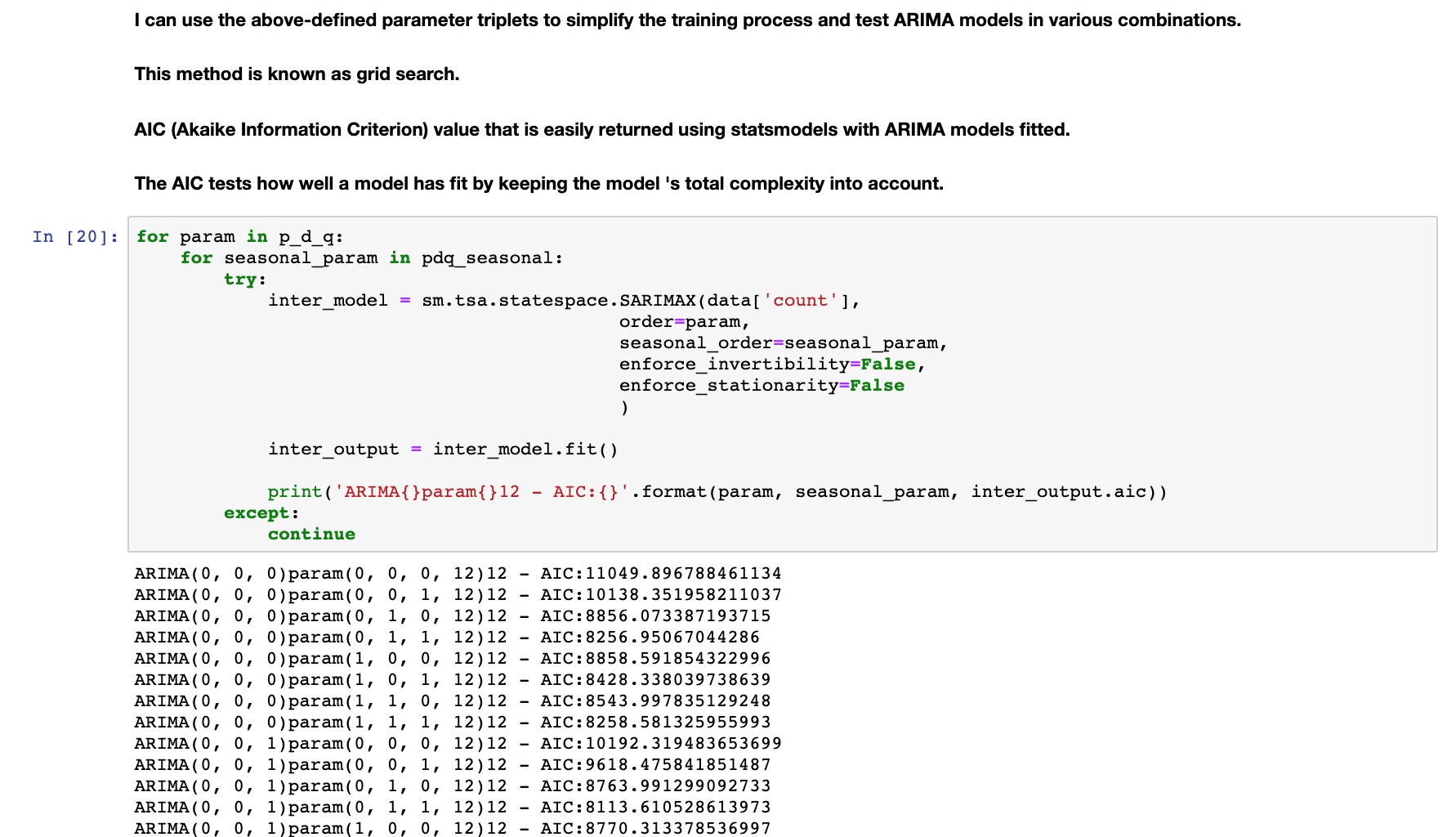
**Classification**

* Graphs showing different parameters/algorithms evaluated in a comparative manner, along with some supportive text

**Time Series Forecasting:**

For the Prophet, the important contributors were the external regressors(features). We manually added and removed them to understand when the model performed best.

For ARIMA, we performed grid search for different combinations of hyperparameters. The combination having the least AIC is the best option.



**Classification**

* Analysis of results

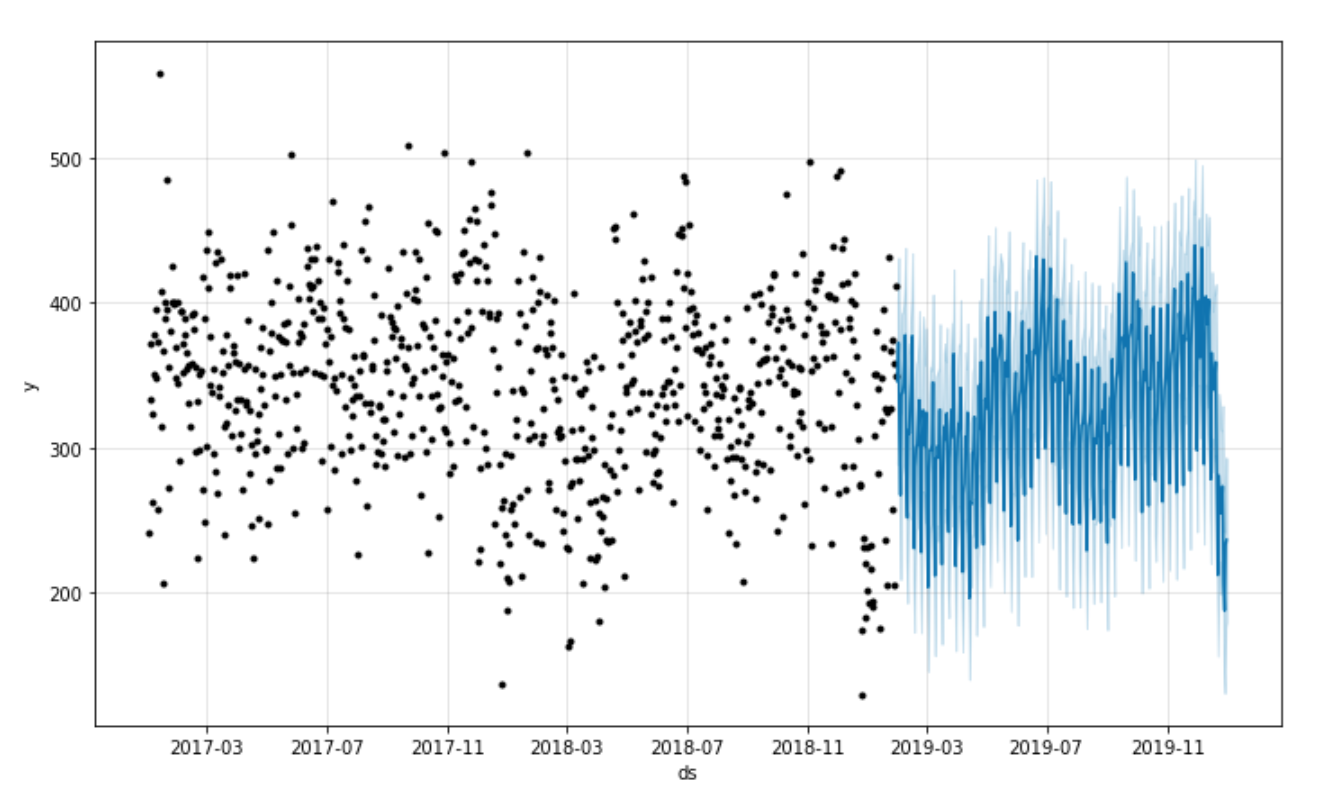
**Time Series Forecasting:**

Following is a screenshot of the forecast generated by Prophet.

Dark blue is yhat. At the top, the light blue is yhat-upper and at the bottom, the light blue is yhat-lower.

*yhat* is the predicted value. *yhat\_lower, yhat\_upper* are the uncertainty interval (predicted range can be between these two values)

If outliers (dots) impact the prediction, they can be excluded.



Historical data \*--------------------------------------------------\* Forecasted data

**Classification:**

**Ch.4 Discussion & Conclusions**

* Decisions made/Things that worked

1. Division of tasks was decided early and that helped members to focus on respective responsibilities.

* Difficulties faced/Things that didn’t work well

1. Exploring time series forecasting algorithms was time consuming.
2. Since we explored both ARIMA and Prophet, it was difficult to understand subtlety initially.

* Conclusions

1. We were able to explore the data and decide important features from a large set of features.
2. We were able to understand how time series algorithms work.

Ch.5 Project Plan / Task Distribution

|  |  |
| --- | --- |
| Tasks | Group Member |
| Dataset Selection | Harshada, Manasa,Pragati |
| Data Exploration - Accidents | Harshada |
| Data Exploration - Casualties |  |
| Data Exploration - Vehicles |  |
| Feature Selection for Time Series | Harshada |
| Feature Selection for Classification |  |
| Preprocessing - Classification |  |
| Model Building - Time Series - ARIMA | Harshada |
| Model Building - Time Series - Prophet | Harshada |
| Model Building - Classification - |  |
| Model Building - Classification - |  |
| Evaluation - Time Series | Harshada |
| Evaluation - Classification |  |
| Report | Harshada, Manasa,Pragati |
| Presentation | Harshada, Manasa,Pragati |

**References:**

[1]<https://stats.stackexchange.com/questions/472266/inference-in-time-series-prophet-vs-arima>

[2]<https://facebook.github.io/prophet/docs/quick_start.html>

[3]<https://mode.com/blog/how-facebook-made-business-forecasting-scalable-for-the-masses-with-prophet/>