Predictive Group Project

M5 Forecasting - Accuracy

Estimate the unit sales of Walmart retail goods

by Addit, Aditya, Rahul, Shubham, Sisi

AGENDA

1

Background and Data Exploration

- Business Objective
- Data Overview
- Exploratory Data Analysis

2

Data Preperation

- Data aggregation
- Feature engineering

3

Modeling Analysis

- Model valuation Metrics
- Predictive Analysis

4

Results

- LightBGM
- Model results

AGENDA

1

Background and Data Exploration

- Business Objective
- Data Overview
- Exploratory Data Analysis

2

Data Preperation

- Data aggregation
- Feature engineering

3

Modeling Analysis

- Model valuation Metrics
- Predictive Analysis
- LightBGM

4

Results

- LightBGM
- Model results

BUSINESS OBJECTIVE

The main goal is to as accurately as possible estimate point projections of the unit sales of numerous items that Walmart sells in the USA.

WHY IS INVENTORY MANAGEMENT IMPORTANT

Retail inventory management ensures a retailer has enough inventory to meet customer demand so that they don't end up with too little or too much merchandise.

It's essential to avoid situations where a retailer runs out of popular items or ends up with excess items that nobody is buying

WHY NOT JUST LOOK AT LAST YEAR'S NUMBERS

Outrightly referring to last year's numbers does not account for holidays, changes in spending patterns of customers due to business growth, regionality, and special events.

PROBLEM AT HAND

It becomes important to find a way to accurately predict future sales using machine learning algorithms with domain knowledge to lower the inaccuracy of prediction.

WAY FORWARD

Our model will use the Sales and external data with LightGBM to accurately predict the unit sales of numerous items so that Walmart can avoid situations where a they runs out of popular items or ends up with excess items that nobody is buying

Data Overview & Exploratory Data Analysis

DATA OVERVIEW

Unit sales of **3049** products

10

stores in the US

- 4 in California
- 3 in Texas
- 3 in Wisconsin

3

product categories

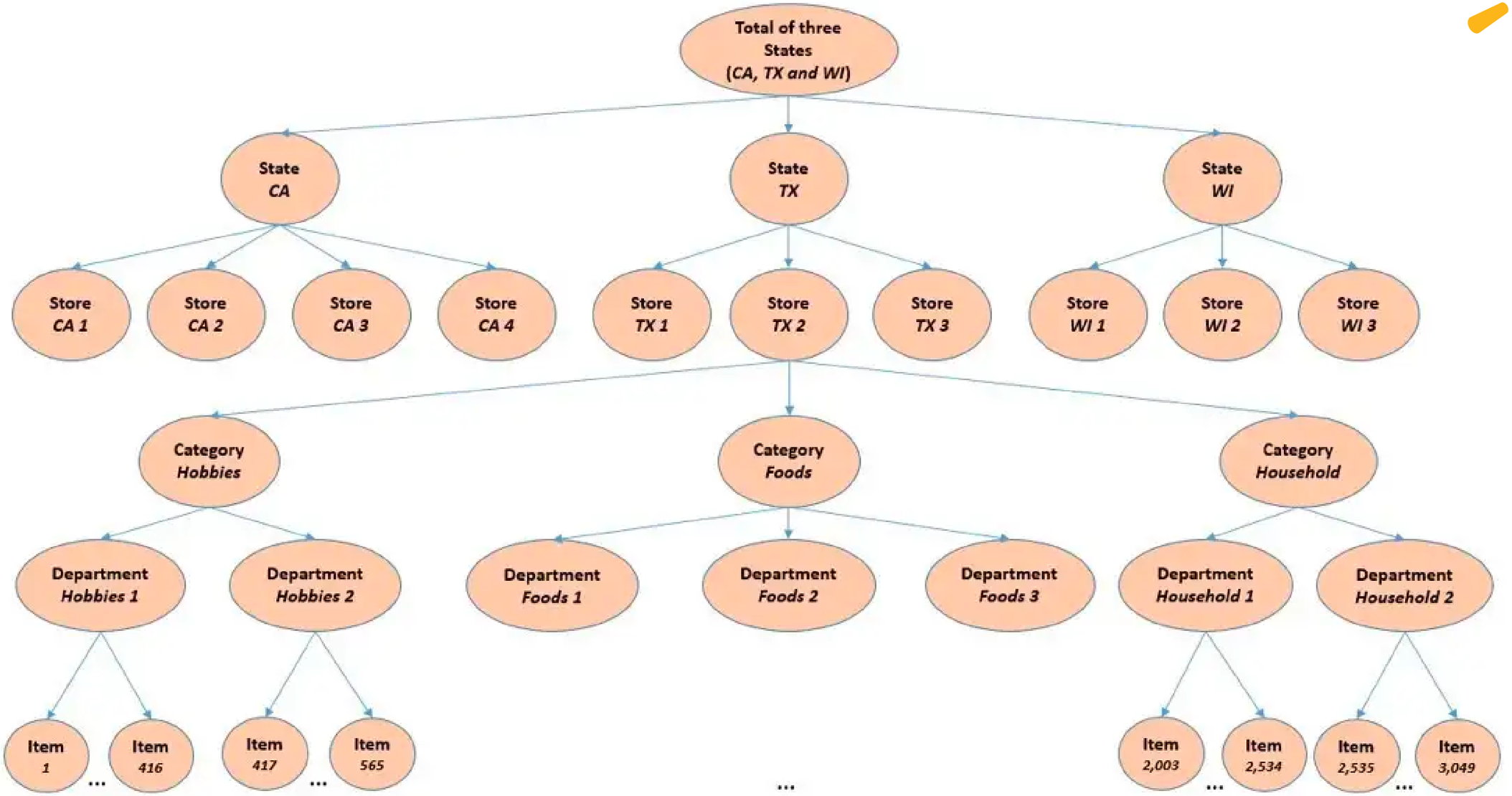
- Hobbies
- Foods
- Household

Explanatory variables

- price
- promotions
- day of the week
- special events

Sales data for 5 years (29th Jan 2011 to 22nd April 2016)

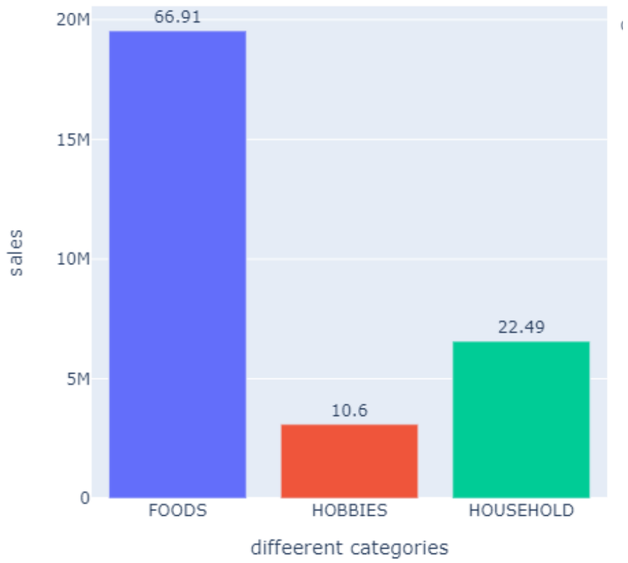
DATA OVERVIEW



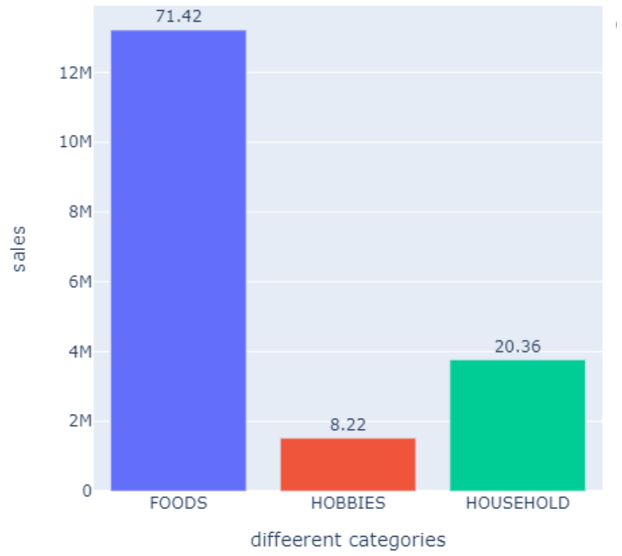
EXPLORATORY DATA ANALYSIS

Analysis at State level

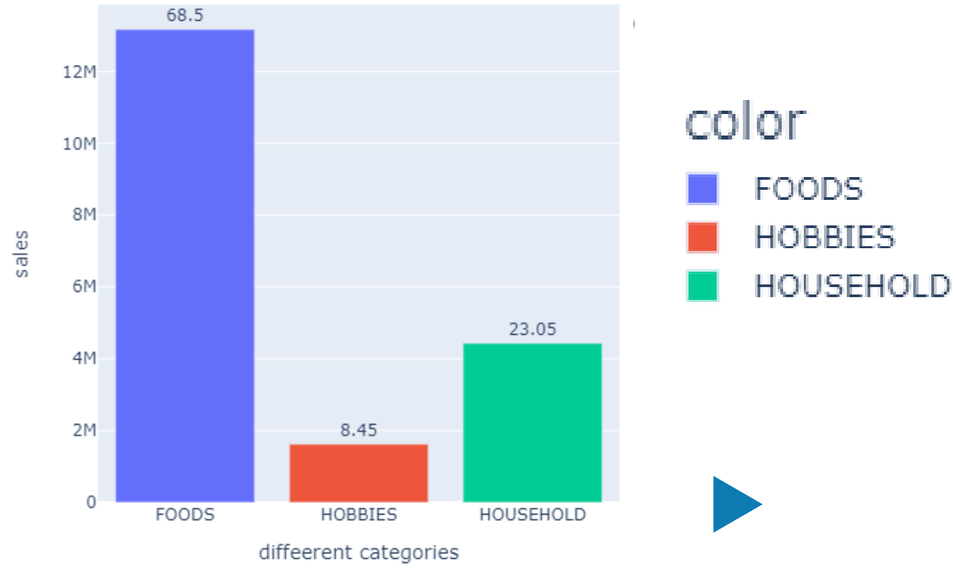
most happening sales category in CA



most happening sales category in WI



most happening sales category in TX

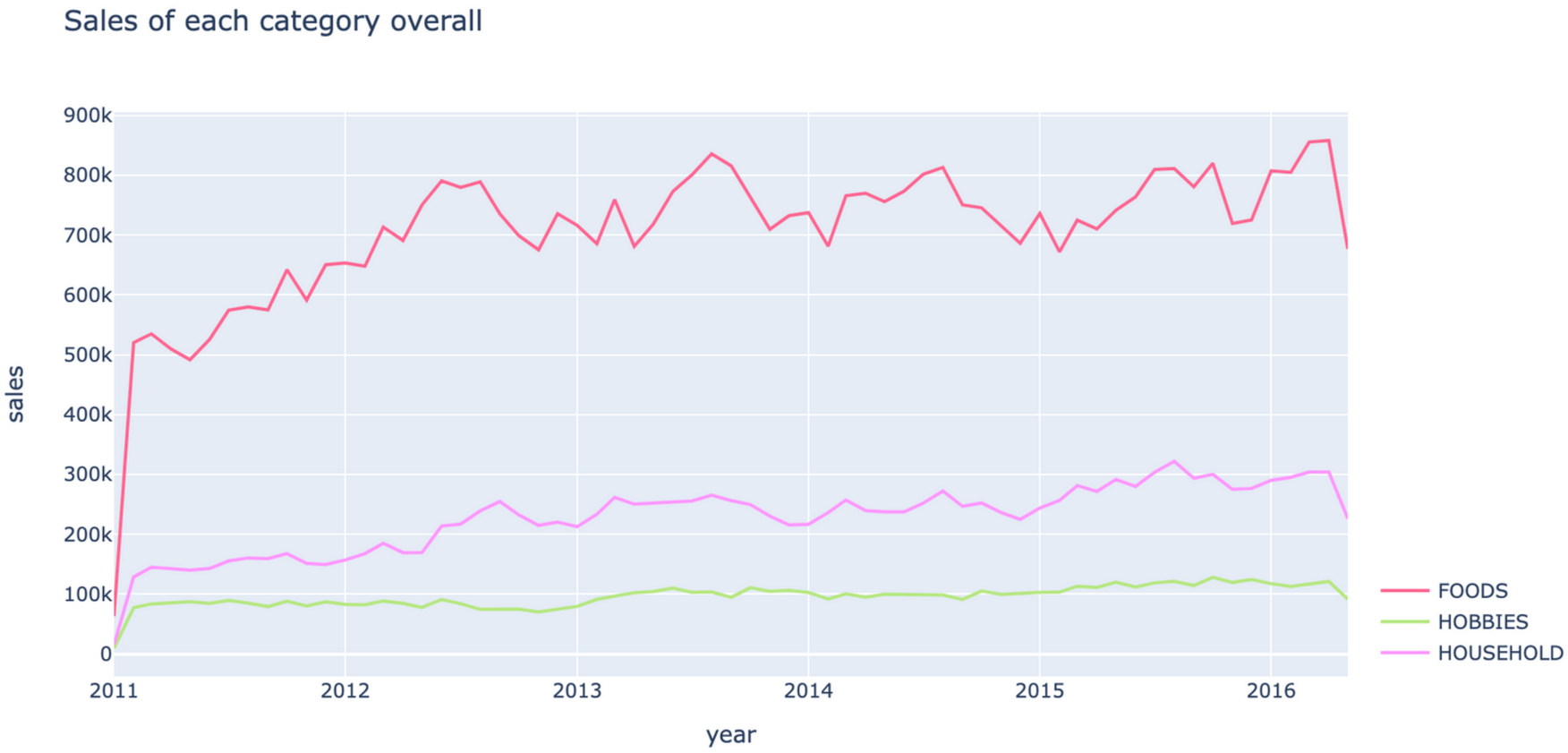


In all states, food category has the highest sales and household category has the least sales.

Analysis at Store level

All stores in each state had almost same percent of sales with slight variation except store 3 in CA.

Analysis at Category level



Food item are consumed the most across all the years then followed by household items.

AGENDA

1

Background and Data Exploration

- Business Objective
- Data Overview
- Exploratory Data Analysis

2

Data Preperation

- Data aggregation
- Feature engineering

3

Modeling Analysis

- Model valuation Metrics
- Predictive Analysis

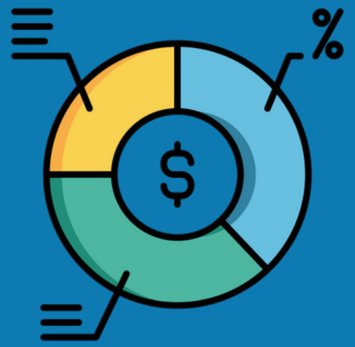
4

Results

- LightBGM
- Model results

DATA AGGREGATION

We combine the data from 3 sources and consolidate them together:



Sales Data

Calendar Data

Prices Data

Final Data

FEATURE ENGINEERING

We do some transformation on the current dataset and make it easier to analyze:

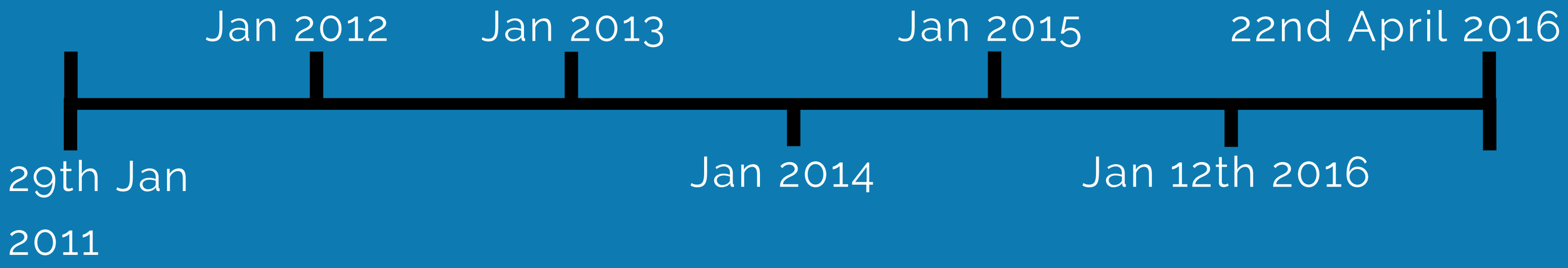
Lag/Shift Features

Rolling Features

Moon Phase

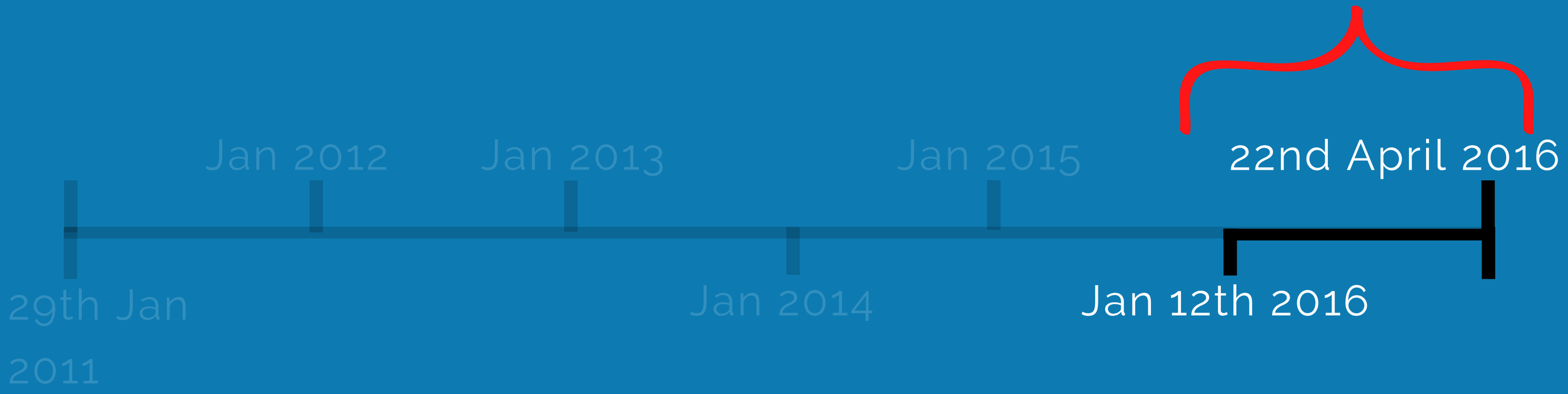
PREDICTIVE ANALYSIS

Before we can run the model on full data, we need to finalize the model that we will be moving forward with:



PREDICTIVE ANALYSIS

We selected last 100 days of data for finalizing.....

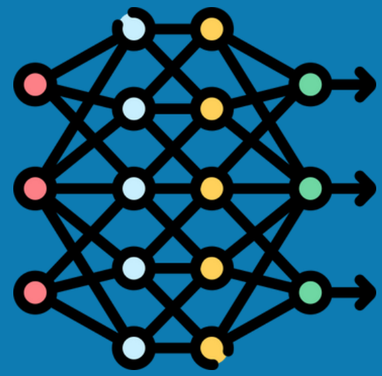


PREDICTIVE ANALYSIS

..... with the below models:

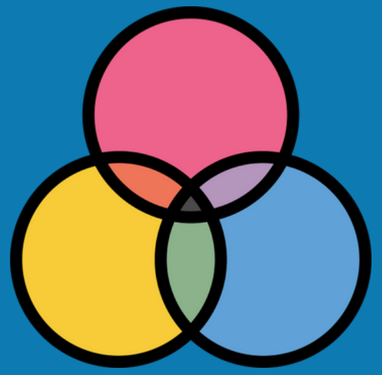
Linear Regression

Light GBM

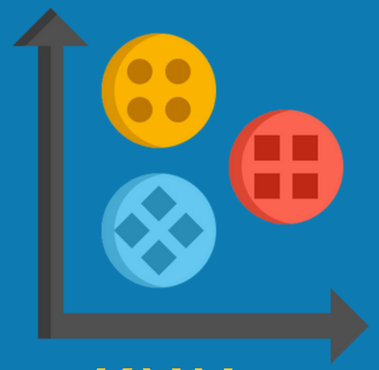


Neural Network

LSTM



XG Boost



KNN

MODEL EVALUATION METRICS

RMSE:

$$\sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$$

Root Mean Square Error(RMSE) is the measure of the differences between values predicted by a model and the values observed

Tweedie Loss:

It is designed to deal with right-skewed data with most of the data distribution "concentrated" around 0.

WRMSSE:

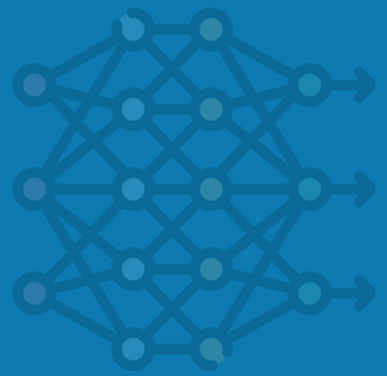
Kaggle uses this to determine the leaderboard score. This metric is designed to deal with right-skewed data with most of the data distribution "concentrated" around 0.

PREDICTIVE ANALYSIS

Out of these model, Light GBM gave us the best results:

Linear Regression

Light GBM



Neural Network

SVM

XG Boost

KNN

AGENDA

1

Background and Data Exploration

- Business Objective
- Data Overview
- Exploratory Data Analysis

2

Data Preperation

- Data aggregation
- Feature engineering

3

Modeling Analysis

- Model valuation Metrics
- Predictive Analysis

4

Results

- LightBGM
- Model results

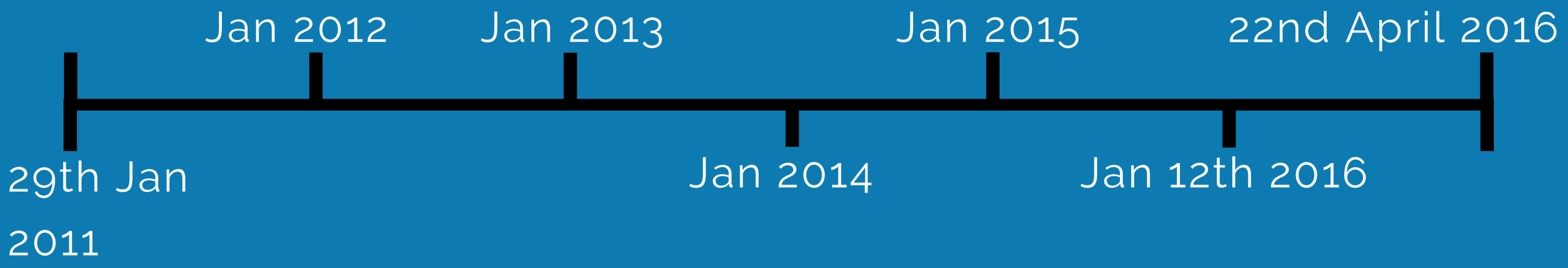
LIGHT GBM

Based on decision tree model, it Increases efficiency in model by using gradient boosting.

Benefits

Ability to provide higher efficiency and more accurate prediction over large-scale data.

THEN, WE RAN THE LIGHTGBM MODEL FOR ALL THE DATA



KAGGLE RESULT

submission_private1.csv

Complete (after deadline) · 8d ago

Score: 0.62114

Private score: 0.60776

Thank You!