April 19, 2017

Version 1.0

*Prepared by*

**Inbox Blues**

|  |  |
| --- | --- |
| inbox_logo  Inbox Consulting (Private) Limited  6th Floor Dawood Center  Tel: 111-551-551  [www.Inboxbiz.com](http://www.Inboxbiz.com) |  |

Contents

[1 Background 3](#_Toc480379012)

[2 Scope 3](#_Toc480379013)

[3 Acceptance Criteria 3](#_Toc480379014)

[4 Key Stakeholders 3](#_Toc480379015)

[5 Technical Solution 4](#_Toc480379016)

[5.1 Data Explanation 4](#_Toc480379017)

[Stores: 4](#_Toc480379018)

[Train: 4](#_Toc480379019)

[Features: 4](#_Toc480379020)

# Background

Predicting future sales for a company is one of the most important aspects of strategic planning. We would like to analyze how internal and external factors of one of the biggest companies in the US can affect their Weekly Sales in the future.

The data collected ranges from 2010 to 2012, where 45 Walmart stores across the country were included in this analysis. It is important to note that we also have external data available like CPI, Unemployment Rate and Fuel Prices in the region of each store which, hopefully, help us to

# Scope

Achieve an approximate weekly sales prediction looking at previous years’ performance per Store on a weekly basis.

# Acceptance Criteria

Model should predict most accurate result when as per following criteria:

1. Coefficient of determination near to 1.
2. Relative Squared Error near to 0.

Further, Model should produce more or less equal predictions when compared to same models on different platforms (e.g. Azure ML, Python, R)

# Key Stakeholders

|  |  |
| --- | --- |
| Sponsor | Inbox Consulting. |
| Client | Walmart |

|  |  |
| --- | --- |
| **Name** | **Role in Project** |
| Muhammad Salahuddin | Project Manager / Data Aggregation / Feature Engineering |
| Saba Muhammad Ali | Front End Development / Feature Engineering / Project Documentation |
| Hammad Ali Khan | Front End Development / Data Aggregation / Feature Engineering |
| Ashar Burney | Model Implementation / Model verification on different platforms |
| Zawar Khan | Model Implementation / Model verification on different platforms |
| Shayan Ishaque | Model Implementation / Model verification on different platforms |
| Saad Khan | Model Implementation / Model verification on different platforms |
| Hunain Wali | Model Implementation / Model verification on different platforms |
| Muhammad Faizan | Front End Development / Back End Development |
| Muhammad Saif | Front End Development / Back End Development |

# Technical Solution

## Data Explanation

Stores:

* Store: The store number. Range from 1-45.
* Type: Three types of stores ‘A’, ‘B’ or ‘C’.
* Size: Sets the size of a Store would be calculated by the no. of products available in the particular store ranging from 34,000 to 210,000.

Train:

* Date: The date of the week where this observation was taken.
* Weekly\_Sales: The sales recorded during that Week.
* Store: The store which observation in recorded 1-45.
* Dept: One of 1-99 that shows the department.
* IsHoliday: Boolean value representing a holiday week or not.

Features:

* Temperature: Temperature of the region during that week.
* Fuel\_Price: Fuel Price in that region during that week.
* MarkDown1:5 : Represents the Type of markdown and what quantity was available during that week.
* CPI: Consumer Price Index during that week.
* Unemployment: The unemployment rate during that week in the region of the store.

## Model Selection

We perform cleaning, aggregation, feature engineering and splitting of data in R and after that we perform Linear Regression Model first, which did not give us expected values may be because of complexity of data.

Since, LM did not predict results (27% only) with accuracy as per our expectations; we change our approach and considered python (Scikit-Learn) for further proceedings.Decision Trees was our first choice to train our model by specifying target and independent variables in Python and it gives satisfactory result (93%). We did not stop here and tried few more models in python i.e. Random forest, which also gives satisfactory result (93%) but Gradient Boosting Method ends with 39% result after optimization.

Finally, we chose Azure ML for further verification, because it provisioned with Web Service to interact our trained model with external systems. We chose Decision Forest Regression to train our model, which satisfies our requirement and gives the predicted value near to 94% with very less Relative Squared Error. We also perform cross validation on trained data set and test data set as well and we obtained same accuracy level.

## Workflow



## Work Breakdown Structure

