Walmart Analytics

Predicting Walmart Sales over the Holidays

By

Michael Stroud,
Dustin Meier,
Joshua Freeman

Problem Description

- What was the challenge?
  - Extract meaning sales trends during holidays to benefit customer buying patterns
- Why did the company choose to use analytics instead of other tools/techniques?
  - Walmart contains the world’s largest data set rendering traditional statistical methods obsolete
- What was the business objective and data mining objective?
  - Business objective – Improve customer experience when shopping at Walmart
  - Data mining objective – Find best model & find dominant factors of weekly sales

Analytics Methodology

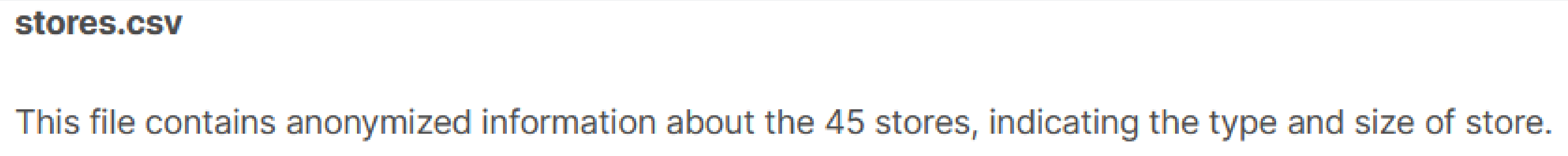
The following four machine learning models have been used:

- Linear Regression
- Lasso Regression
- Gradient Boosting Machine
- Random Forest

Predictive Analytics were performed to predict Weekly sales

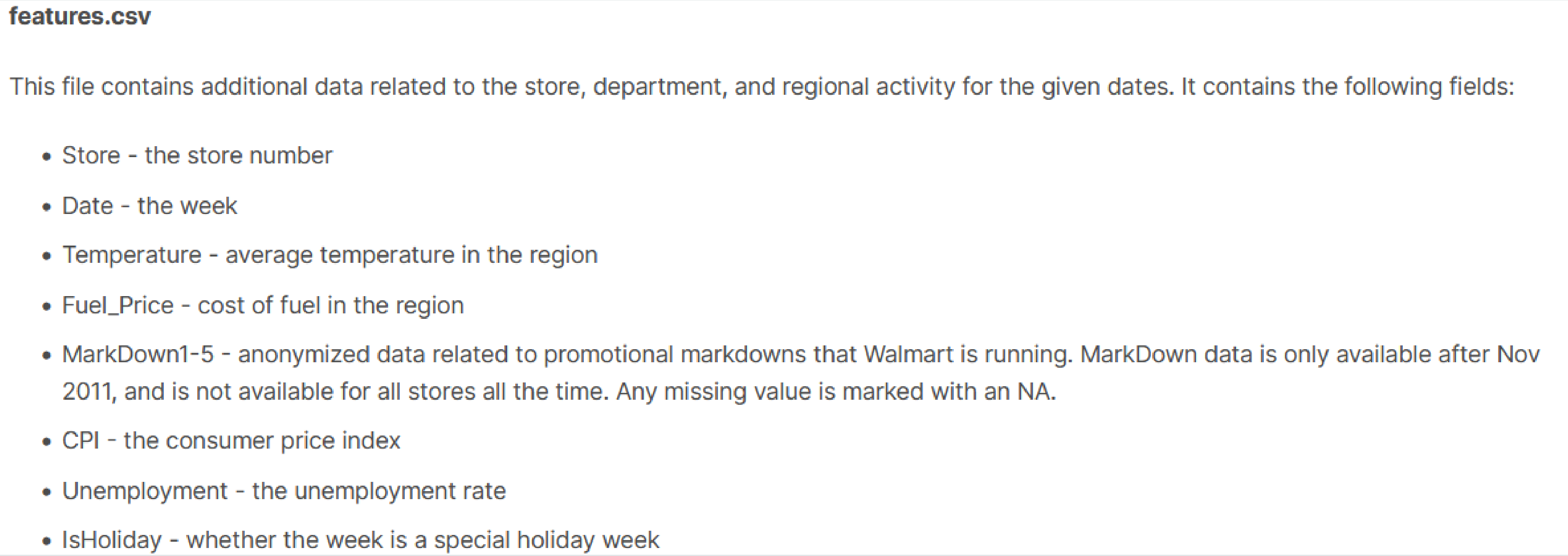
Analytics:
They tried to identify relationships between independent and target variables through **Exploratory Data Analysis.** Components like the correlation matrix and scatter plots, feature importance plots created as part of the random forest and gradient boosting models as well as the interaction effects.

Data Sources: Provided by Walmart



stores.csv

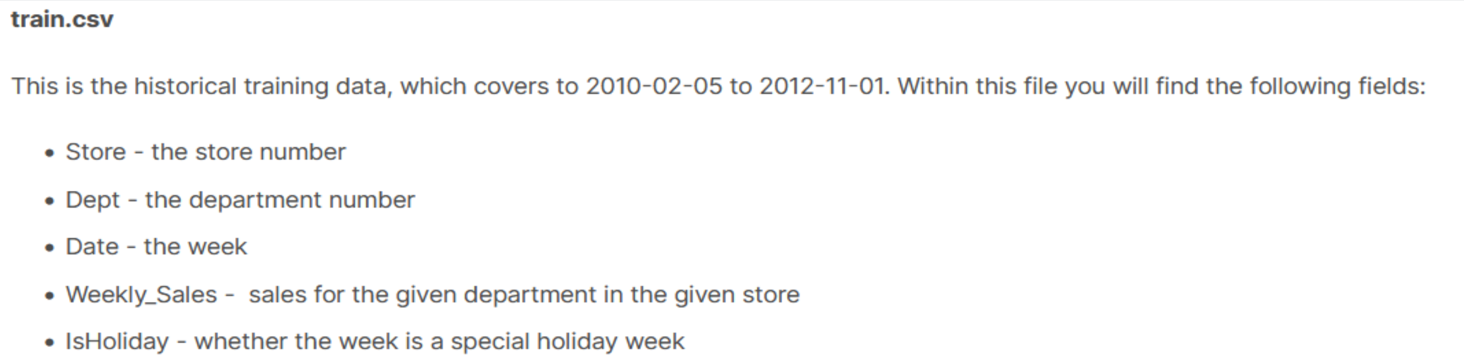
This file contains anonymized information about the 45 stores, indicating the type and size of store.



features.csv

This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:

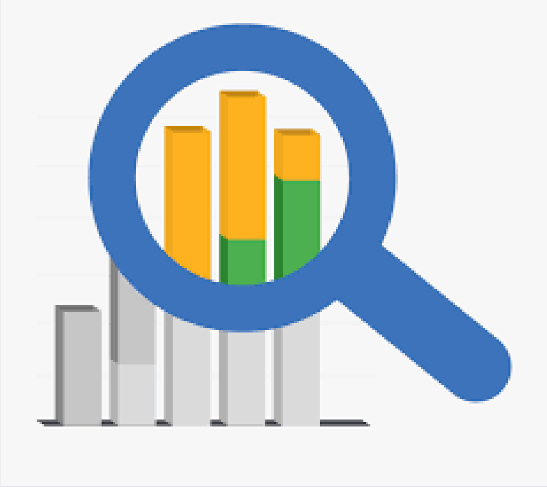
- Store - the store number
- Date - the week
- Temperature - average temperature in the region
- Fuel_Price - cost of fuel in the region
- MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
- CPI - the consumer price index
- Unemployment - the unemployment rate
- IsHoliday - whether the week is a special holiday week



train.csv

This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within this file you will find the following fields:

- Store - the store number
- Dept - the department number
- Date - the week
- Weekly_Sales - sales for the given department in the given store
- IsHoliday - whether the week is a special holiday week



Independent Variables

Not Statistically significant:

Store - the store number
Date - the week
Fuel_Price - cost of fuel in the region
MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running.
IsHoliday - whether the week is a special holiday week
Dept - the department number
Size of store
Type of store

Statistically significant :

Temperature
CPI (Consumer Price Index)
Unemployment

Dependent Variable

<u>Weekly sales</u>

Model Evaluation

Train.csv compared against Test.csv

Train.csv is a dataset with historical weekly sales. A model was trained on this dataset. 4,21,570 rows

Test.csv is a dataset without historical weekly sales. The model's predictive power was evaluated from this dataset. 1,15,064 rows

This competition is evaluated on the weighted mean absolute error

$$\text{WMAE} = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i |y_i - \hat{y}_i|$$

Where:

- n is the number of rows
- (hat{y}_i) is the predicted sales
- (y_i) is the actual sales
- (w_i) are weights. w = 5 if the week is a holiday week, 1 otherwise

If I were to evaluate the model differently, I would use a higher weight for holiday weeks because
the goal of this study is to predict holiday sales.

Managerial Implications and Benefits

1. Improve customer satisfaction by predicting consumer demand
2. Verify correlation between statistically significant independent variables and dependent variable Weekly Sales.
   - Could be used to predict profitability of un-opened locations
3. Increases revenue opportunity by insuring critical items are in stock

Results

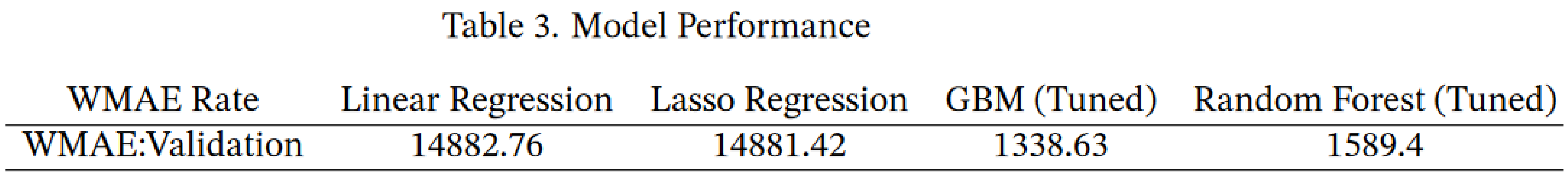


Table 3. Model Performance

WMAE Rate	Linear Regression	Lasso Regression	GBM (Tuned)	Random Forest (Tuned)
WMAE:Validation	14882.76	14881.42	1338.63	1589.4

- Which independent variable(s) had the highest impact on the dependent variable?
  - Temperature, CPI (Consumer Price Index), Unemployment
  - Tuned Gradient Boosting model, with the lowest WMAE score, is the main model used to create the final predictions for this study.
- How long did the freelance researcher take to complete the project?
  - 2014-Present

Conclusions

o **The Goal**: Analyze sales trends to better predict customer buying patterns

- ➢ 4 different machine learning algorithms were used (ex: Linear Regression & Machine Learning)
- ➢ Datasets of recorded data and training data imputed (stores.csv & train.csv)
- ➢ Independent variables, such as 'Temperature', that had the greatest effect on buying patterns were identified

o **Future scope**: Expand e-commerce of Walmart

- ❑ Easier to collect data
- ❑ Expands marketing capabilities

o **Cross-market scope**: These methods and methodology can easily be applied in the e-commerce market of other companies as well

- ➢ Websites such as amazon.com and target.com could benefit from using this data as well; they mostly have the same target audience, the shoppers are just virtual

References

https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109#toc-1

https://www.rit.edu/ischoolprojects/sites/rit.edu.ischoolprojects/files/document_library/Rashmi_Jeswani_Capstone.pdf

https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data