UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

3251 Statistics for Data Science

Course Project

Weather Driven Sales Prediction

Regression Analysis

[Group 9]

Adil Alkhateeb

Linda Wong

Mahammad Ali

Submitted on: Apr 21st 2018

Table of Contents

Introduction 3
Data Overview 3
Correlation Analysis 3
Data Preparation 3
Regression Analysis 4
Regression Framework Design 4
Models processing and Output 5
Regression Analysis Results 5
Conclusion 5
References 6
Appendix: Jupyter Notebooks 6

Introduction

Walmart is notable for being one of the largest multinational retail stores in North America and reputable for their large and spacious stores offering various products to the average consumers at a competitive price. According to an article published in 2014, Walmart discovered value in managing their inventory using weather as component of their prediction models.

Following the release of the article, Walmart launched a Kaggle competition challenging participants to predict the sales of 111 potentially weather-sensitive products (like umbrellas, bread, and milk) around the time of major weather events at 45 of their retail locations.

The objective of this course project is to design and implement a sales prediction model by performing a regression analysis utilizing all sales and weather information provided by Walmart.

Data Overview

Walmart has provided three datasets for the challenge as following: daily Sales records, daily Weather records and a Key table linking Weather stations to stores within their area. The sales records provided total daily sales per store for 111 items for the period JAN 2012 to OCT 2014 masked into numbers only to maintain their anonymity and reduce potential prediction bias. The weather records provided measurements of 18 local climatological data provided by 20 Automated Weather Observing System (AWOS) stations covering the subject stores. Below [Table 1] lists all the measured data:

tmax	tmin	tavg	depart	dewpoint	wetbulb	heat	cool	sunrise
sunset	codesum	snowfall	preciptotal	stnpressure	sealevel	resultspeed	resultdir	avgspeed

TABLE 1: RECORDED CLIMATOLOGICAL DATA

Correlation Analysis

A general correlation analysis was conducted to better understand the relationship between the features and how they may influence the regression optimization process [NB4]. It was noticed that the majority of highly correlated features were dropped during the selection process due to their negative effect on the models fitting and variance from actuals. The following heat map [Figure 1] illustrates the overall correlation between all of the subject features.

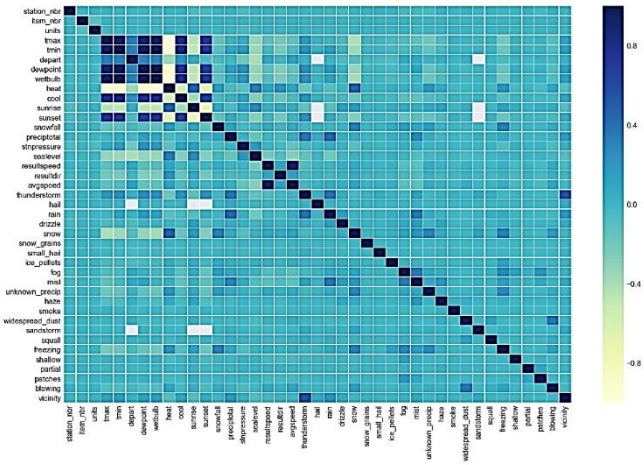

FIGURE 1 - CORRELATION HEATMAP

Data Preparation

The preparation phase included data cleaning [NB1], items sales mapping into the weather stations and finally joining the weather data with items sales per station to produce a final regression ready dataset [NB3].

Sales records were clean with no missing data, therefore active data cleaning was only required for the weather data as the provided dataset had a total of 51,108 dummy and missing readings. All dummy data was carefully replaced by NaNs to facilitate data filling. A simple data backfill approach was found to be inefficient because backfill function may fill readings from other stations. A limited filling logic within the same station was inefficient as well because it uses the previous day (or following day) only as a reference to fill. The best approach was to interpolate the missing readings by

using the surrounding days within the same station. This was achieved by utilizing '*interpolate*' function in Pandas DataFrames with time-series awareness configuration.

The recorded weather data was almost entirely continuous except one parameter '*codesum*'. This was a categorical parameter flagging 32 significant weather phenomena such as rain, heavy fog, snow, thunderstorm and others. Each of the weather flags was binary encoded [NB2.1, NB2.2] in order to provide the ability to incorporate the weather conditions into the prediction model. This has expanded the weather dataset significantly from 18 to 49 feature as every weather flag was accounted in a separate column.

The final preparation step was to link the sales records to the weather data utilizing the provided Key table. All items sales were linked to their respective weather stations by using the stores numbers as the common key. As some stations were covering several stores, items sales were grouped and summed by the related weather stations. Lastly, the weather for each station were joined by their related item numbers, and stores numbers were dropped from the analysis for simplification purposes. The final combined clean DataFrame had a total of 2,038,737 rows and 41 columns with 18,367 records per item.

Regression Analysis

For the purpose of the analysis, the Multiple Linear Regression model was selected using the *'ols'* function within the '*StatsModel*' package to predict sale of unit for given a set of weather forecast.

Regression Framework Design

Due to the large number of predictors involved, and in order to get the best predicting model, *Backward elimination* [NB5] and *Forward selection* [NB6] methods were implemented in parallel for each item in the dataset. Each regression iteration was conducted on Five Folds (80:20) Train/Test split logic to further narrow down the best performing prediction model per selection method per item.

The selection methods were based on improving R2_adj with each iteration, therefore producing the model with the best fit for each fold per item. An overfitting scenario was suspected due to the high Mean Square Error (MSE) result per model. Therefore, the selection process was then modified to evaluate on MSE reading per fold instead of R2_adj to produce the best possible model with the least MSE value. By making the following adjustments, it produced improved models resulting in MSE values that were on average 50% less than the previous selection process that relied on R2_adj values. This was accomplished by slightly reducing the fit of the model to eventually but dramatically improving the predictions. Below [Figure 2] is an illustration of both evaluation criteria results for the top item in sales (# 45):

Evaluation Criteria:			**R2_Adj**		**MSE**		**MSE Improvement**	
Folds	*item*	*selection*	*R2_Adj*	*MSE*	*R2_Adj*	*MSE*	*MSE Delta*	*%*
fold1	45	Forward	0.560251715	13478.00381	0.450041386	6377.37534	-7100.63	-53%
fold2	45	Forward	0.565887125	14433.33723	0.449557781	6731.300762	-7702.04	-53%
fold3	45	Forward	0.563067961	13811.41438	0.449110138	6450.697522	-7360.72	-53%
fold4	45	Forward	0.56039181	14609.54234	0.457753509	7176.632181	-7432.91	-51%
fold5	45	Forward	0.566636608	13935.26545	0.464221146	7389.851892	-6545.41	-47%

FIGURE 2 - ITEM 45 REGRESSION RESULTS

Models processing and Output

The regression analysis code was designed to switch between R2_adj and MSE evaluation criteria dynamically as a configurable parameter. The code runtime was 1 hour and 40 mins on average to generate individually optimized regression models for each item in the dataset. Upon completion, all the generated models were evaluated, and the models producing the lowest MSE values between the folds were shortlisted on a per item basis for every selection process. Finally, the model from the backward and forward selection processes were compared again for their lowest MSE and the best performing models were identified per item. All final optimum models were consolidated in a DataFrame then serialized and saved using the *'Pickle'* package for immediate prediction use [NB7].

Regression Analysis Results

The final produced optimum models revealed that the forward selection process was successful in producing optimal prediction models for items with high unit sales quantities. On the contrast, Backward elimination was more efficient in producing models predicting items with low sales (which was represented by most of the items). The final results had Backwards models selected for **101** items and Forward models selected for **9** items only which had the top selling items in Table 2.

Item	Quantity Sold	Selection
45	1,005,111	Forward
9	916,615	Forward
5	846,662	Forward
44	577,193	Backward
16	226,772	Backward

TABLE 2 - TOP SELLING ITEMS OPTIMUM SELECTION METHOD

Using the results from the regression analysis, prediction of sales of any item given a set of weather measurements and conditions is made possible. The sales prediction vs actual for the top three items in total sales were selected are illustrated below in [Figure 3] on a per month basis:

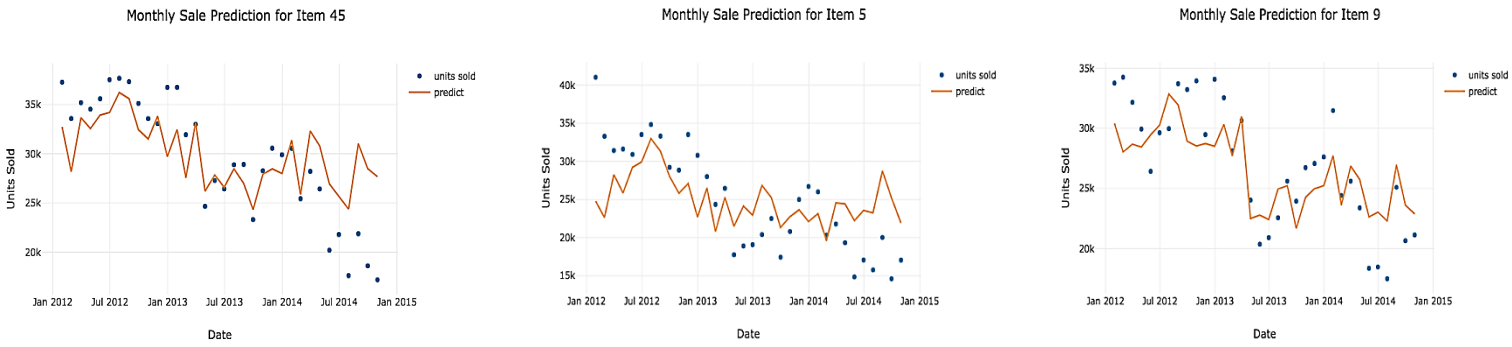

FIGURE 3 - TOP 3 ITEMS PREDICTION VS ACTUAL

The regression analysis was successful in capturing the reducing sales seasonal trend and managed to produce moderately fitted prediction models as shown above [NB8].

Conclusion

In conclusion, though intuitively weather typically should affect the sale of weather-sensitive items, however the analysis proved to show that weather may not be of a great influence to the behavior of consumers. The weather based sales prediction models can assist in planning stock levels better if combined with directly related consumer buying influencers such as day of week, holidays, paycheck days, and promotions.

References

1) http://adage.com/article/dataworks/weather-forecast-predicts-sales-outlook-walmart/295544/
2) https://www.kaggle.com/c/walmart-recruiting-sales-in-stormy-weather
3) https://www.bloomberg.com/gadfly/articles/2016-03-31/walmart-s-first-ever-sales-drop-marks-new-era

Appendix: Jupyter Notebooks

[NB1] Data Cleaning.ipynb

[NB2.1] CodeSum Encoding.ipynb

[NB2.2] Codesum Encoding Map.xlsx

[NB3] Items Weather Linking.ipynb

[NB4] Correlation Analysis.ipynb

[NB5] Regression Analysis - Backward Elimination.ipynb

[NB6] Regression Analysis - Forward Selection.ipynb

[NB7] Regression Analysis Consolidation.ipynb

[NB8] Prediction Charts.ipynb