Georgia Institute of Technology

AMAZON DATA CHALLENGE

Team We R

Kai Li	IE Master
Kaiwen Luo	IE Master
Su Yu	ME Phd
Zexing Song	Stats Master

03-06-2020

Contents

1 Introduction

Weather events can vary significantly in type and severity, thus affect the delivery operations. To address this factor, each team develops a model to determine the impact of weather on Amazon site operations for Master Data Science Challenge (MDSC) 2020. Participating teams compete to predict site impact scores of 2019 at 91 locations using weather data in 2017-2019 and historical site impacts in 2017-2018. Accurate predictions will allow Amazon to better plan and prepare for site operations.

Our team approaches this challenge with a nested model to extract the weather effect from other factors impacting site operations, such as time, location, and site itself. For this reason, our model contains four stages. Stage I identifies outliers using data inspection and random forest. Stage II models the time effect using time series. Stage III models the sites and locations using averaging and K-means clustering. Stage IV models the integrated weather impact using LASSO. Using this model, our team fits the 2017-2018 site impact scores with a RMSE of 1.15, and forecasts cite impacts for 2019.

2 Data Management

2.1 Data Understanding

In this project, raw business data are provided in three categories - impact scores, mapping files and weather forecasts. Daily impact scores are given at the zip5 level in the range from -1 to 35 between 2017 and 2018. Geographical data are given on 91 Fulfillment centers to map site zip-code with longitude and latitude. Weather forecasts data provided by Global Forecast System (GFS) contains 111 predicting variables such as: wind speed, pressure and temperature recorded four times per day from 2017 to 2019. Geographical distribution and average impact score of each site is plotted in Figure 11 in Appendix A for initial data understanding. Further data understanding is performed on two folds.

2.1.1 Weather Events Hypothesis

Based on a general understanding of weather events and preliminary weather research, our team hypothesizes that approximately 28 variables will have a major contribution to site operations, and 14 more variables may have a minor effect (Figure 15). Because these numbers are much smaller than 111, a dimension reduction process is included. This approach is further discussed in Section 3.4.

2.1.2 Impact Score Outliers Definition and Assumptions

Our team finds outliers in site impact scores. As shown by a histogram of all sites' impact scores in original data (Figure 5), the data distribution is symmetrical and tends to be normalized. However, there are abnormal values in the head and tail of the distribution which is composed of only three values, -1, 0.52 and 35. To better understand their significance, our team particularly researches an operation site in zip code 8085 (New Jersey) and find it opened at mid-September 2017. However, the same impact scores are assigned for this site prior to this date in one of these three values (-1, 0.52 and 35). We assume that consistent impact scores produced before site openings are not related to weather events or regular site operations. Therefore, these three values are not reasonable in this analysis and are defined as "outliers" to be excluded from regular regression models. A classification model is developed in Section 3.1 for outlier identification.

2.2 Data Manipulation

2.2.1 Data Aggregation

For the training data preparation, our team treats the 2017 and 2018 data as a whole. Weather data is mapped to zip code using longitude and latitude, then combined with respective impact scores for each day. As the weather data has 4 observations per day while the impact score only has 1, the weather data is averaged

for each day prior to combination. For the testing data, 2019 data set is prepared using the same method, but without the response variable, impact score. After data combination, training (2017-2018) and the testing (2019) datasets have different dimensions on weather variables. Multiple weather factors in training dataset are not included in testing dataset, and vice versa. To assure model consistency, a benchmark was set for the weather factor selection criterion. Only intersections of weather factors in both datasets are used for model development. Other factors are removed from each dataset. This is the first process for data manipulation.

2.2.2 Missing Value processing

The second step processes the missing values ("N/A"). Our team treats the missing values in the training dataset and testing dataset differently. For N/As in the training dataset, our team first removes weather factors with over 20% of "N/A" by column, then deletes observations with "N/A"s existence by row. For N/As in the testing dataset, our team also removes weather factors removed in the training dataset, but replaces the rest of N/As with the column means or local means from the nearest times at the same site depending on N/As' distribution. The reason that we use different methods for training and testing data is to ensure the genuineness of data. It is better to delete N/As instead of adding aggregated values with additional error for training, as the data is rich in observations. For testing, nevertheless, adding aggregated values is the only option to ensure a complete prediction. Now, the datasets are complete for further analysis.

3 Modeling

To extract the weather impact from distractions such as time of operation, site location, or site's business suspension, a nested model with four stages is developed. They are Outliers Identification, Time Effect, Site Location, and Weather Impact. The Classification method is used for Outliers Identification, and regression methods are used for other stages. The nested model as defined in Equation 4 is illustrated in Figure 1. Here, z, t, w, Y are data input on location (zip-code), time, weather, and site impact score. $\bar{w}$ is the weather averaged per day. $\hat{Y}_{outlier}, \hat{Y}_{time}, \hat{Y}_{zip}, \hat{Y}_{weather}$ are the estimates generated from four stages of modeling to forecast future site impact, $\hat{Y}$. In this document, the symbol "+" means a set combination. Details and methods of each sub-model are described as followed.

$$\hat{Y} = \hat{Y}_{outlier}(Y, z, t, w) + \hat{Y}_{weather}(\hat{Y}_{time}(Y, t), \bar{Y}_{zip}(z, Y), \bar{w}) \tag{1}$$

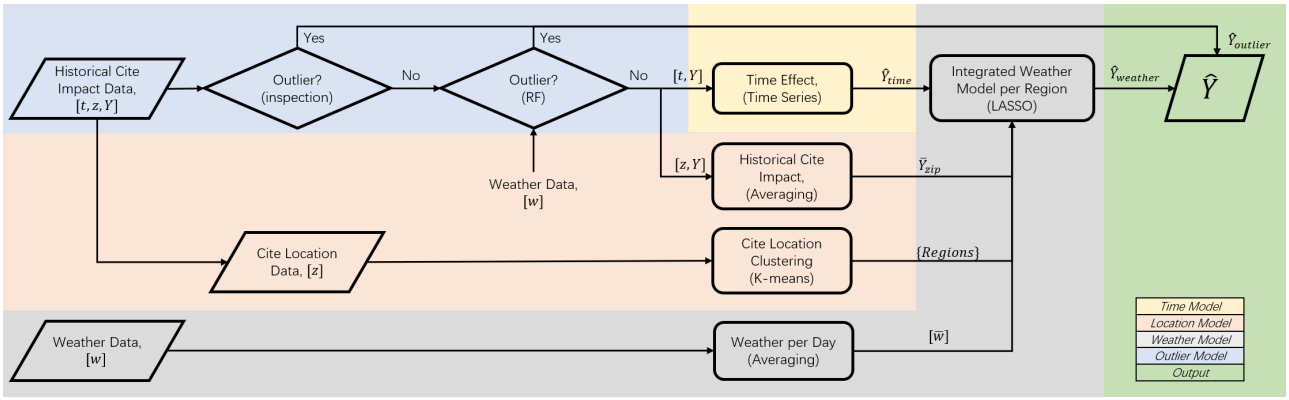

Figure 1: Modeling Illustration with Flow Diagram

3.1 Outliers Identification

3.1.1 Data Inspection

$$\hat{Y}_{outlier} = f_{zip}^{Inspection}(z, t, Y) + f^{RF}(z, t, w, Y) \tag{2}$$

Stage I identifies outliers (impact score = -1, 0.52 or 35) using Equation 2 to excluded them from regression modeling using data inspection, Random Forest classification. From inspection, there do exist 3 evident patterns on outlier distributions after we plotted all outliers in the 26 zip-codes they exist in the training dataset (Figures 2, 3, 4). Zip-codes with outlier Pattern I are assigned "-1" for all future impact scores in 2019. Zip-codes with outlier Pattern II have their outliers manually removed from training dataset, and the rest observation returned for regression modeling. They predicted regularly using regression models for 2019. Zip-codes with outlier Pattern III do not have a univariate trend, and is difficult to predict from a glance. Data of these sites are extracted and classified using the Random Forest algorithm for predictions.

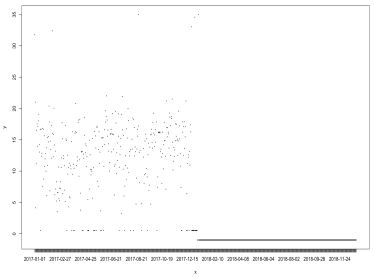

Figure 2: Time-Series Pattern I

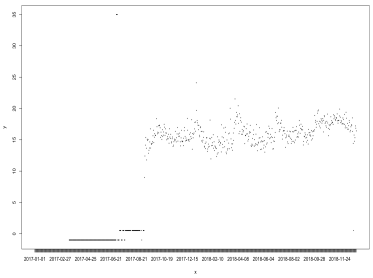

Figure 3: Time-Series Pattern II

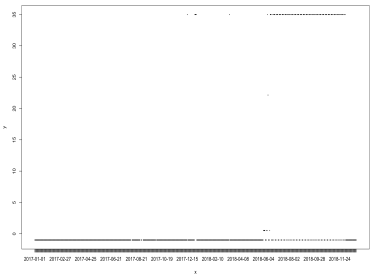

Figure 4: Time-Series Pattern III

3.1.2 Random Forest

Many outliers still exist that are incapable or difficult to be identified from inspection. Therefore, our team used classification methods from machine learning based on "scikit-learn" package in Python. Different methods such as SVM, Neural Network, Random Forest, KNN are compared, and model selection is performed based on data explanatory ability and RMSE. Since this project focuses on a real-world problem, the model needs to give a reasonable explanation of extracting the outliers. Models like Neural Network do have better performance on RMSE, but they lack data explanatory ability. From the above considerations, our team finally choose Random Forest as the outlier classifier. A cross-validation method is designed to find the optimum hyper-parameters, such as max_tree_depth, $criterion$, etc.

3.2 Time Effect

$$\hat{Y}_{time} = f^{TS}(t, Y) \tag{3}$$

Stage II addresses effect of time on impact score as Equation 3 As shown in Fig.12, the impact scores have several significant patterns related to time, which is observed in frequencies of "per year" (e.g. Christmas), "per season", and "per week". For example, at the beginning of each season, the impact scores are the highest, then continuously decrease with time until the end of the season. These patterns can be modeled as a time series. By calling Forecasting package in R, we used *auto.arima* to fit a time series model and to generate estimator $\hat{Y}_{time}$

3.3 Spatial Effect

3.3.1 Location Clustering

$$Region = f^{K-means}(z) \tag{4}$$

Stage III models the spacial effects on site and locations. The given data shows that there are 91 sites operating all over the country. After conducting online research, our team noticed that the proximity of places may lead to similar trends of site operation results. On the one hand, to address regional response to weather events,

such as a 5-inch snow in Northeast vs. South, our weather model need to be more specialized about locations. On the other hand, training models for each individual site is much too time-consuming, and will greatly reduce the model robustness. Thus, we clustered 91 sites into 10 regions for weather model development using K-means.

3.3.2 Historical Site Performance

$$\bar{Y}_{zip} = mean(Y_{z,t1}, Y_{z,t2}, ..., Y_{z,tn}) \tag{5}$$

In addition to location clustering, the impact of each individual site is addressed in Equation 5. Site 1 does not necessarily have the same impact as site 91 due to differences in multiple factors, such as local population density, rural/urban district, number of a delivery man, etc. The sites' differences in performance can distract the modeling on weather impact. To address the site difference, estimator $\bar{Y}_{zip}$ is developed as the mean of historical site performances after outlier removal.

3.4 Weather Impact

$$\hat{Y}_w = f_{region}^{LASSO}(\hat{Y}_{time}, \bar{Y}_{zip}, \bar{w}) \tag{6}$$

Stage IV models the weather impact as Equation 6. 85 of 111 provided weather variables are used in the weather impact model. The other 26 variables are dropped because of dataset inconsistency and large percentages of N/As. Because the cite impact score is provided only once a day, the weather data is averaged over 4 before feeding into the LASSO model to match the dimension. The weather impact model describes the effect of weather on top of time and location effects through integration of estimators $\hat{Y}_{time}$ and $\bar{Y}_{zip}$, from previous stages. From Weather Events Hypothesis, half of the variables may have a small effect on the model. Thus, the weather impact is modeled using LASSO for dimension reduction. Note that here, LASSO also reduces the correlation between $\hat{Y}_{time}$ and $\bar{Y}_{zip}$. Ridge regression is an appropriate, but a less effective alternative, of which their comparison is discussed in Section 4.4. Other dimension reduction methods, such as Principal Component Analysis (PCA), Partial Least Square (PLS), are not considered due to lack of data explanatory ability.

A unique weather impact model is fitted for each clustered region defined in Section 3.3. As illustrated in Figure 1, the predicted $\hat{Y}_w$ is combined with $\hat{Y}_{outlier}$ to generate complete forecasts on cite impact.

4 Analysis

4.1 Model Analysis

Outlier Identification Before the Outlier Identification process, raw Ys (impact score) are not normally distributed and outliers are clearly shown in the head and tail(Fig.5). After applying outlier data inspection and Random Forest to exclude outliers, Y data are normalized as the plots show(Fig.6 and Fig.7).

Using the Random Forest classifier, the top ten features which influence the outlier judgment are shown in Fig. 8. The Date and Site (zip5) factors affect the outlier the most, thus verify our observations about the operation condition mentioned above. Fig. 8 also shows that the geographic factors have a great influence on the outliers. The geographic factors include zip code, latitude and longitude, x and y (x and y are provided in weather forecasts, and are assumed as geographic index). The accuracy of the random forest model is 96.5%.

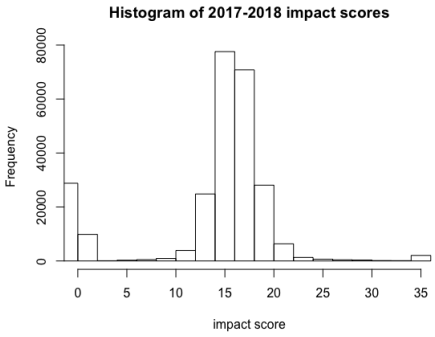

Figure 5: raw Y data in 2017-2018

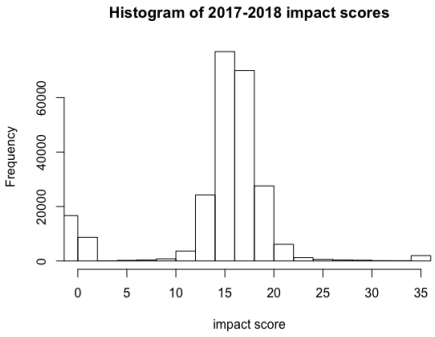

Figure 6: Y after data inspection

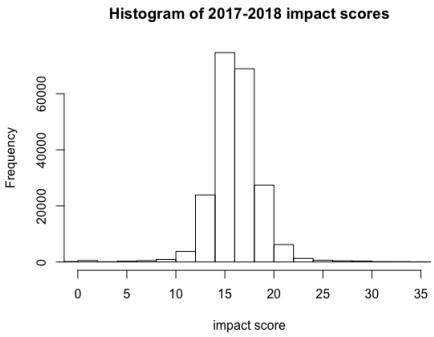

Figure 7: Y after data inspection and Random Forest

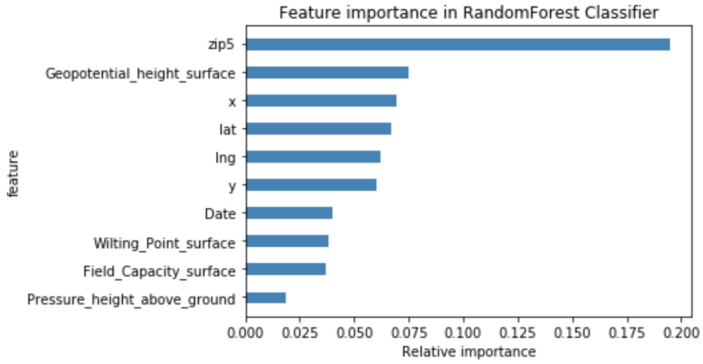

Figure 8: Top Ten Features Determining Outliers

Time Effect The time series model using ARIMA(4,1,1) (Figure 13) demonstrates great fitness to given data, assure complete removal of time effect from weather impact in the integrated weather model.

Location Clustering Location clustering is achieved through K-means by varying k from 6 to 15 clusters. Clusters are displayed on the U.S. Map, and the number of 10 clusters was chosen by visual judgment, as it produces the least group overlap.

Weather Impact LASSO and Ridge regression are compared for weather impact modeling with location clustering and outlier predictions on the entire training dataset. LASSO gives a RMSE of 1.15. Ridge regression gives a RMSE of 1.21. LASSO is used for the final model due to a smaller RMSE. The better performance from LASSO is likely because LASSO's is more aggressive in dimension reduction, which is more effective for highly correlated weather data.

Overall, our model achieves a minimum RMSE of 1.15 on training dataset, as shown in Figure 9.

4.2 Correlation

Correlations between predictors and impact scores are shown in Fig.5. Estimators$\bar{Y}_{zip}$ (0.707) and $\hat{Y}_{time}$ (0.515) have the highest correlation, showing a strong dependence of impact score on site and time. Weather factors have relatively small correlations ($<$0.1) to site impact scores comparing to the other two variables. This is also reflected by the LASSO model (Figure 14). Among weather factors, the top 5 are Ozone Mixing Ratio, MSLP Eta model reduction msl, pressure surface, Pressure reduced to MSL, and Pressure Height Above Ground. These variables are mostly related to air pressure which could be the cause of severe weather

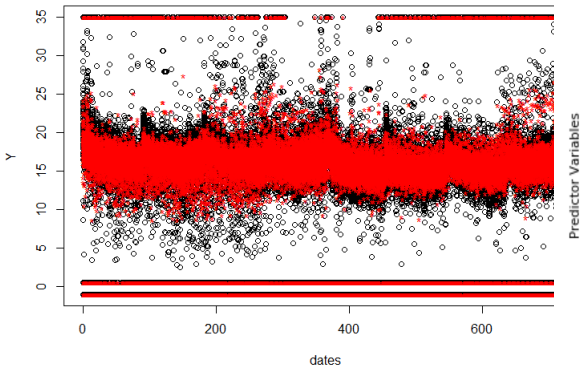

Figure 9: Overall Model Fitting (Red) to 2017-2018 Training Dataset (Black)

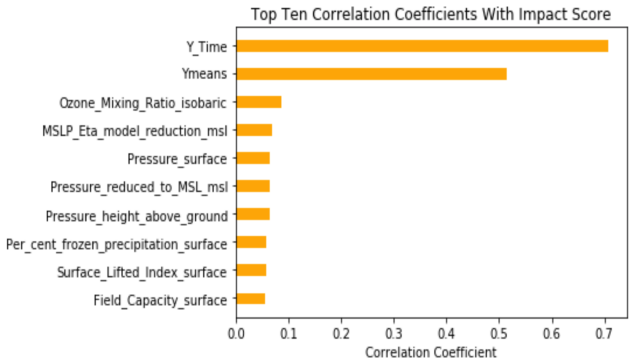

Figure 10: Correlation Coefficients

events. For example, stormy weather are caused by low pressure. When the storm becomes stronger and joins with other nearby thunderstorms, it becomes a hurricane and significantly affects the human's life, especially transportation.

5 Discussion

From the data model, one sees that Amazon's site operation impact is primarily determined by human factors such as time and the site itself. Gladly, weather events have a much smaller influence on Amazon's operation. Among all weather events, pressure related factors demonstrate the most correlation to site operation. This is hypothesized as a potential connection between pressure and extreme weather events. Though commonly underestimated, local forecast on pressure should be closely monitored for improvement in future cite operations.

Improvements can be made on current model by including more extreme weather information. The final LASSO model shows the impact scores are lightly affected by the daily weather data offered by NOAA Global Forecast System (GFS). Nevertheless, ground delivery can be harshly halted under extreme weather events such as hurricane, tornado, blizzard and floods. Due to time limitation, our team did not find more weather data on severe weather events from 2017 to 2019 to augment the data set. If the predictor variables related to these unusual weather events were added to the model, the prediction accuracy could be improved. Other factors, such as sequence of weather events, weights on day averaging are also sources of further model improvements.

6 Conclusion

In conclusion, our team models the impact of weather on Amazon's site operations using a nested model with four stages - outlier identification, time effect, spatial effect, and weather impact. This model is supported by a RMSE of 1.15 on 2017-2018 data, and is used to forecast 2019. Our model concludes that pressure is the primary weather factor to Amazon's site impact score. It also finds that overall, impact of weather is small comparing to site and time.

Appendix A: Additional Figures

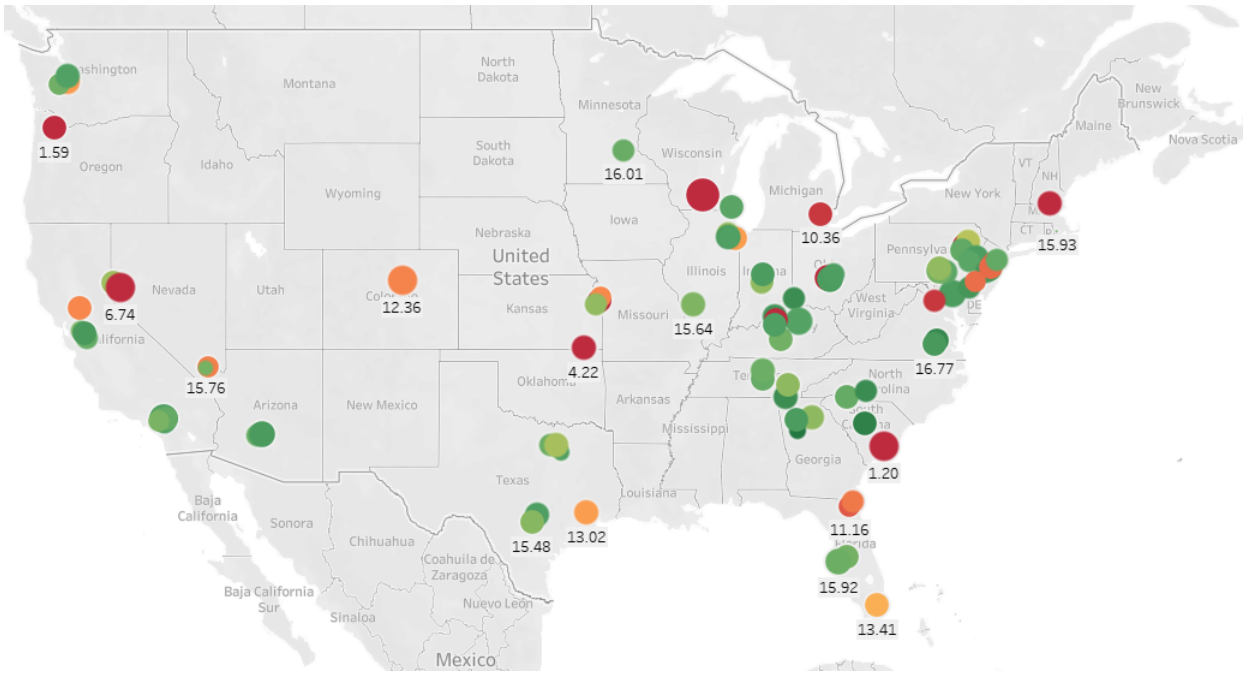

Figure 11: Distribution of Amazon Sites with Average Impact Scores (Note) and Zipcode Gross Income (Color)

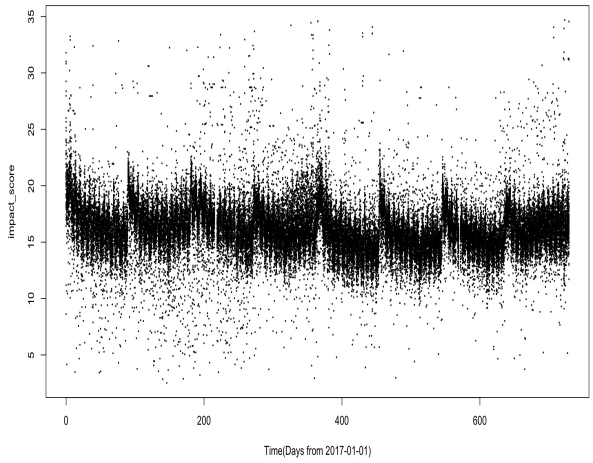

Figure 12: Impact Score(Without outliers) VS. Time

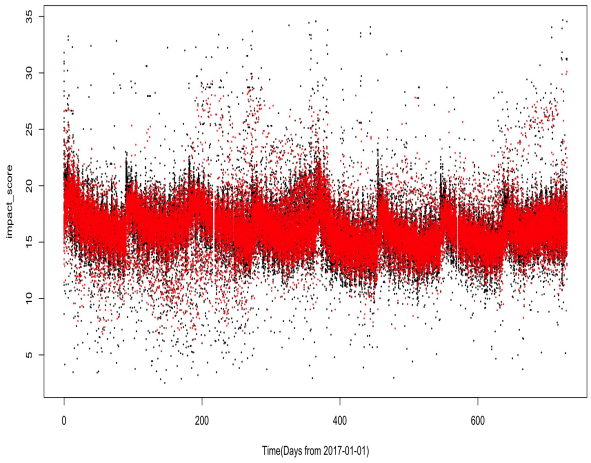

Figure 13: Fitted Time Series

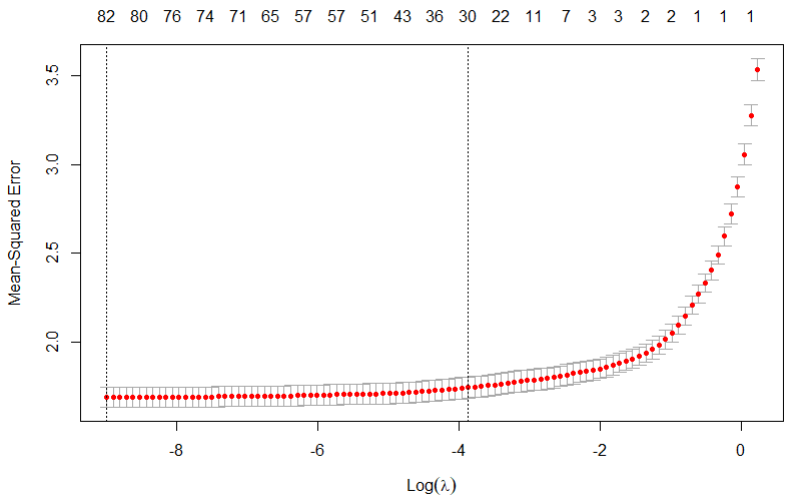

Figure 14: Plot of Error vs. $\lambda in LASSO for Weather Impact Model$

Appendix B: Resources

	Expectation	Comments	
5_Wave_Geopotential_Height_isobaric			sea level
Absolute_vorticity_isobaric	0.01	low	sum of vorticity
Apparent_temperature_height_above_ground	1		
Best_4_layer_Lifted_Index_surface			
Categorical_Freezing_Rain_surface	1		
Categorical_Ice_Pellets_surface	1		
Categorical_Rain_surface	1		
Categorical_Snow_surface	1		
Cloud_mixing_ratio_hybrid			
Cloud_mixing_ratio_isobaric			
Cloud_water_entire_atmosphere_single_layer			
Composite_reflectivity_entire_atmosphere			
Convective_available_potential_energy_pressure_difference_layer			
Convective_available_potential_energy_surface			
Convective_inhibition_pressure_difference_layer			
Convective_inhibition_surface			
Dewpoint_temperature_height_above_ground		?	
Field_Capacity_surface	1	soil moisture	
Geopotential_height_highest_tropospheric_freezing	0.01	logic iso>freeze	
Geopotential_height_isobaric	0.01		
Geopotential_height_maximum_wind			
Geopotential_height_potential_vorticity_surface	0.01		
Geopotential_height_surface	1		
Geopotential_height_tropopause			
Geopotential_height_zeroDegC_isotherm	?		
Graupel_snow_pellets_hybrid	1	might be correlated	
Graupel_snow_pellets_isobaric	1	might be correlated	
Haines_Index_surface	1		
ICAO_Standard_Atmosphere_Reference_Height_maximum_wind			
ICAO_Standard_Atmosphere_Reference_Height_tropopause			
Ice_cover_surface	1		
Ice_growth_rate_altitude_above_msl			
Ice_water_mixing_ratio_hybrid	1		
Ice_water_mixing_ratio_isobaric	1		
Land_sea_coverage_nearest_neighbor_land1sea0_surface	1		
Land_cover_0__sea_1__land_surface			
MSLP_Eta_model_reduction_msl			
Ozone_Mixing_Ratio_isobaric			
Per_cent_frozen_precipitation_surface	1		
Planetary_Boundary_Layer_Height_surface			
Potential_temperature_sigma			
Precipitable_water_entire_atmosphere_single_layer			
Precipitation_rate_surface	1		
Pressure_height_above_ground	1		
Pressure_maximum_wind	1		
Pressure_of_level_from_which_parcel_was_lifted_pressure_difference_layer			
Pressure_potential_vorticity_surface			
Pressure_reduced_to_MSL_msl			
Pressure_surface			
Pressure_tropopause			
Rain_mixing_ratio_hybrid	1		
Rain_mixing_ratio_isobaric	1		
Relative_humidity_entire_atmosphere_single_layer	1		
Relative_humidity_height_above_ground			
Relative_humidity_highest_tropospheric_freezing			
Relative_humidity_isobaric		Correlated	
Relative_humidity_pressure_difference_layer			
Relative_humidity_sigma			
Relative_humidity_sigma_layer			
Relative_humidity_zeroDegC_isotherm	0.01		
Snow_depth_surface	1		
Snow_mixing_ratio_hybrid	0.01		
Snow_mixing_ratio_isobaric	0.01		
Soil_temperature_depth_below_surface_layer			
Specific_humidity_height_above_ground			
Specific_humidity_pressure_difference_layer			
Storm_relative_helicity_height_above_ground_layer			
Sunshine_Duration_surface	1		
Surface_Lifted_Index_surface			
Temperature_altitude_above_msl		Correlated	
Temperature_height_above_ground		Correlated	
Temperature_isobaric	1		
Temperature_maximum_wind		Correlated	
Temperature_potential_vorticity_surface		Correlated	
Temperature_pressure_difference_layer		Correlated	
Temperature_sigma		Correlated	

	Temperature_surface		Correlated
	Temperature_tropopause		Correlated
	Total_cloud_cover_isobaric		
	Total_ozone_entire_atmosphere_single_layer		
	U_Component_Storm_Motion_height_above_ground_layer	0.01	
	V_Component_Storm_Motion_height_above_ground_layer		
	Ventilation_Rate_planetary_boundary		
	Vertical_Speed_Shear_potential_vorticity_surface		
	Vertical_Speed_Shear_tropopause		
	Vertical_velocity_geometric_isobaric		
	Vertical_velocity_pressure_isobaric		
	Vertical_velocity_pressure_sigma		
	Visibility_surface	1	
	Volumetric_Soil_Moisture_Content_depth_below_surface_layer		
	Water_equivalent_of_accumulated_snow_depth_surface	1	might be correlated
	Wilting_Point_surface		
	Wind_speed_gust_surface	1	
	u_component_of_wind_altitude_above_msl	0.01	might be correlated
	u_component_of_wind_height_above_ground	0.01	might be correlated
	u_component_of_wind_isobaric	0.01	might be correlated
	u_component_of_wind_maximum_wind	1	might be correlated
	u_component_of_wind_planetary_boundary		
	u_component_of_wind_potential_vorticity_surface		
	u_component_of_wind_pressure_difference_layer		
	u_component_of_wind_sigma		
	u_component_of_wind_tropopause		
	v_component_of_wind_altitude_above_msl		
	v_component_of_wind_height_above_ground	0.01	
	v_component_of_wind_isobaric	0.01	
	v_component_of_wind_maximum_wind	0.01	
	v_component_of_wind_planetary_boundary		
	v_component_of_wind_potential_vorticity_surface		
	v_component_of_wind_pressure_difference_layer		
	v_component_of_wind_sigma		
	v_component_of_wind_tropopause		
	Summary	28.14	
	1 - Hypothesized to have an effect		
	0.01 - Hypothesized to have a minor effect		