Regression Analysis of Factual Weather in Szeged

Team Mates:

1. Lakshmikanth, Vishakh
2. Mukthinuthalapati, N K Vidya Shanmukha Maruthi Sasidhar
3. Parekh, Jayminkumar Chandreshbhai
4. Rai, Keertika

Guide: Prof. Nandhini Gulasingam

Dataset:

Dataset Name: WeatherHistory.CSV

Description: In this dataset, we have 11 Variables (Date-Time, Summary, Precipitation Type, Temperature, Apparent Temperature, Humidity, Wind Speed, Wind Bearing, Visibility, Cloud Cover and Pressure).

The Dependent Variable is Temperature (which we will be predicting) and the Independent Variables are Time, Summary, Precipitation Type, Apparent Temperature, Humidity, Wind Speed, Wind Bearing, Visibility, Cloud Cover and Pressure.

In this dataset, we have 96,453 Observations and there are no missing values.

URL**:** https://www.kaggle.com/budincsevity/szeged-weather

Problem Description:

Our project goal of analysis is to predict/project the future values of the Temperature Variable (Y variable) given the set of values for independent variables (X-variables).

During the analysis, we noticed that all other variables have a direct effect on Temperature predictor hence the dependent variable

Proposed Methodology:

Initially, we would split the data set into Training and Test set using 80-20 Split approach.

Then, we will perform Exploratory Data Analysis to get more insight into the dataset wherein we examine the statistics of the continuous variables. In our next step, we will identify and remove the insignificant predictor using the standardized estimates, correlation coefficient and then we would analyze the data model assumptions using the help of histograms, scatterplots and residual plots.

We will then build the multiple linear regression model using various selection methods and narrow down to a couple of models on which we will check the model assumptions and Validate them. If the observed Adjusted R-Squared value is less than 60% then we will apply transformations or add Interaction Terms to come up with a better model and then we will perform hypothesis testing on the selected model. We will analyze the presence of outliers, Influential points and then take necessary action on them and then we would rerun the model. In our entire model analysis, we are aimed at proposing an efficient yet effective model that provides the best explanation for dependent variable.

At last, we will generate the final regression model equation and predict Test Data and try to analyze the limitations of the Linear Regression method.

References:

1. Implementation of GOSSTANDART technique for verifying and validating Goodness of Fit and maximal test power:
B.Yu.Lemeshko, S.N.Postovalov, E.V.Chimitova, Rules of Application of Goodness- of – fit in simple and Composite Hypothesis Testing, pg. 126 – 132.

2. Implementation of Vector Regression Modeling technique for better/accurate prediction:
Kavitha S, Varuna S, Ramya R, A Comparative Analysis on Linear Regression and Support Vector Regression, 2016 Online International Conference on Green Engineering and Technologies (IC-GET)

3. Proposes MRDDV (Mutiple linear Regression with Dependent Dummy Variable) that combines quantitative and qualitative variables to estimate the behavior of other predictors
N.J.Park, K.M. George-N.Park, A Multiple Regression Model for Trend Change Prediction, Proceedings of the *ITI 2010 32nd Int. Conf. on Information Technology Interfaces,* June 2010.