Predicting the Solar radiation level in Hawaii

Ruchi Parab

Northeastern University

Week 6: Project Presentation

03/10/2019

ALY – 6040 Data Mining Applications [CRN – 21288]

Instructor: Justin Grosz

Table of Contents

[Introduction 3](#_Toc4678876)

[Analysis 4](#_Toc4678877)

[Data set 4](#_Toc4678878)

[Data Cleaning 4](#_Toc4678879)

[Exploratory Data Analysis 5](#_Toc4678880)

[Linear modelling 6](#_Toc4678881)

[Optimization of Linear Model 6](#_Toc4678882)

[Stepwise Regression and Gradient Boosting Model 8](#_Toc4678883)

[Conclusion 9](#_Toc4678884)

# Introduction

A company named ABC Corp. is trying to set up a solar power plant in an area in Hawaii. They need to know the levels of solar radiation emitted in that particular area as it will help in designing and sizing of the photovoltaic power systems that are used in the generation of electricity from solar energy. We as data analysts are predicting the solar radiation in the Hawaii islands based on the various atmospheric factors of that area. We are trying to build a predictive model with a high accuracy which can estimate the radiation in that area with the help of the meteorological data present in that area.

This paper demonstrates the various data analysis techniques performed on the given dataset required to build a predictive model which forecasts the level of solar radiation emitted in Hawaii based on the input parameters. These techniques involve data cleaning, exploratory data analysis and various statistical modelling techniques. The data set chosen for this analysis was extracted from Kaggle and contains the meteorological data from the HI-SEAS weather station in Hawaii from September through December 2016. The steps involved in the analysis are mentioned the chapter below.

Files for reference:

# Analysis

## Data set

The data set contains 32,686 entries and a total of 11 columns. The column parameters are: UNIXTime(time in seconds since Jan 1, 1970), Data, Time, Radiation, Temperature, Pressure, Humidity, WindDirection(Degrees), Speed, TimeSunRise and TimeSunSet (Hawaii time). Since we are building a model to prediction the radiation levels in that area, we take “Radiation” as the target variable for this analysis.

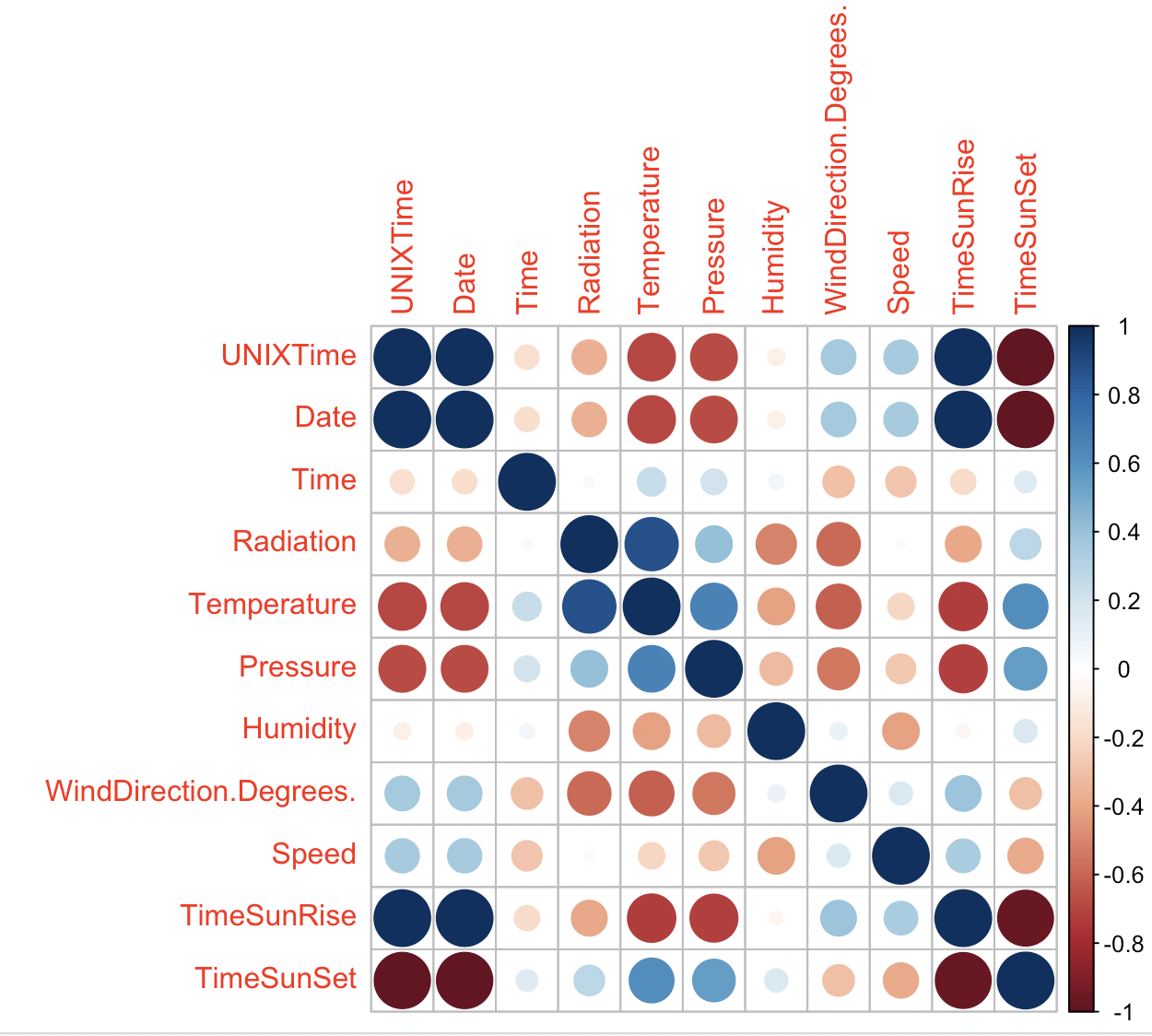
Data Cleaning

In order to visualize the data, we require it to be clean, i.e. without any missing values or typos. So, we started off by cleaning the data which was done by checking for NA values, missing values, typos and outliers.

The columns were cleaned by removing the junk values from the string of their original values. The junk value of “12:00:00 AM” was removed from the time column as it wasn’t of significant importance in the data set. The missing values were checked in Excel but no variables were found containing missing values. The NA values were checked by but none of the variables contained any NA values in the data set. The outliers were tested by creating box plots of the relatable columns. Since it was found that not many outliers were present as they lay close to the box plots and no data point had a significant difference in values compared to the other data points, none of them were removed from the plots. The last step of data cleaning was checking for typos in all the columns. The column name “Data” which was considered a typo as it was misspelled, was later corrected to “Date”. The variables “Date”, “Time”, “TimeSunRise” and “TimeSunSet” which were of the data type “factor” were converted to “numerical” data type as they contained only numerical values.

## Exploratory Data Analysis

After the data has been cleaned for visualization, we performed an exploratory data analysis (EDA) on it to find patterns, trends, outliers and the correlation between the data points. We created a scatter plot plotting radiation with respect to month. From the graph a decreasing trend was clearly observed from September to December with an onset of winter. To understand the significance between the variables and to gain further insights on them we constructed a correlation matrix by the medium of a heat map function.

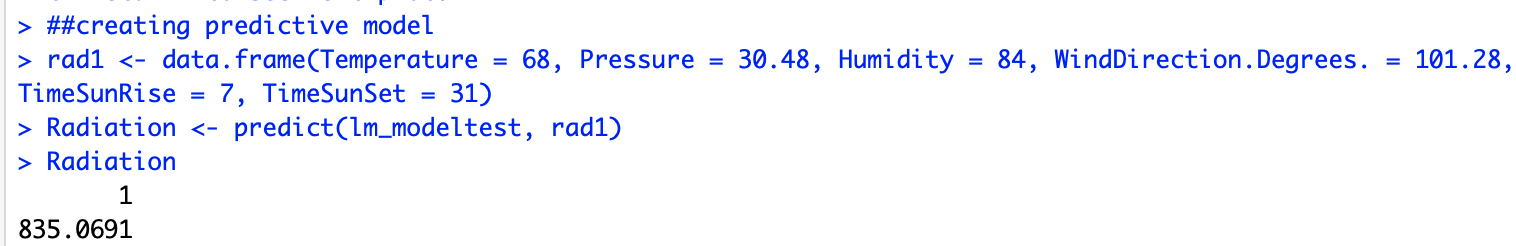


As we have considered “radiation” as our dependent variable, we checked the correlation of radiation with all the other variables. The variables “Time” and “Speed”, which did not show any significant correlation with radiation were removed from the data set. Also, as “UnixTime” and “Date” did not provide any statistical significance in predicting the value of radiation, they too were removed from the data set. Thus, only by observational inferences we removed 4 of the 11 irrelevant variables from the data set. From the plot it was also seen that the dependent variable “radiation” had the highest correlation with “temperature”. This suggests that with the rise in temperature, the value of radiation will also increase. Further, a histogram was created

## Linear modelling

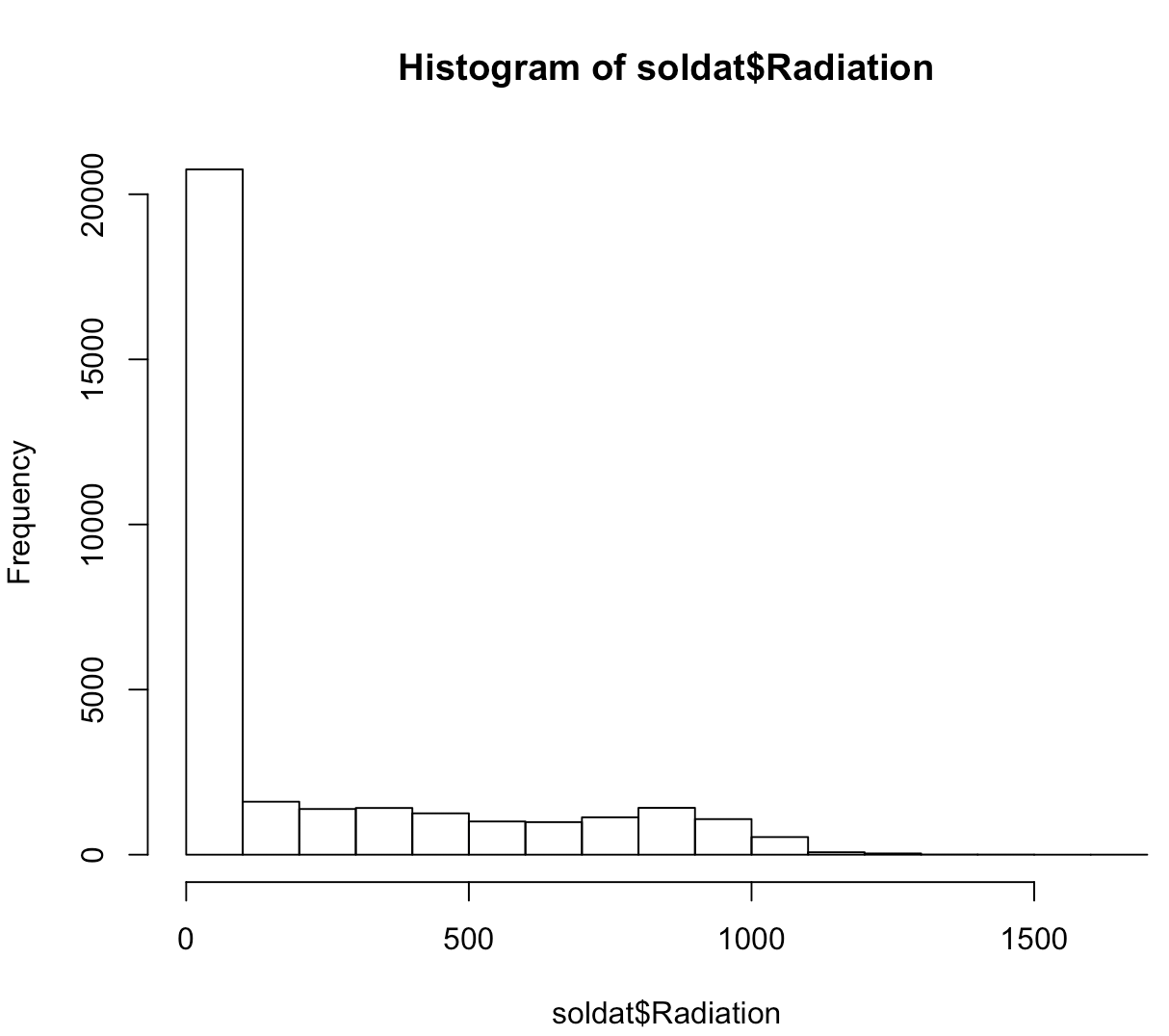
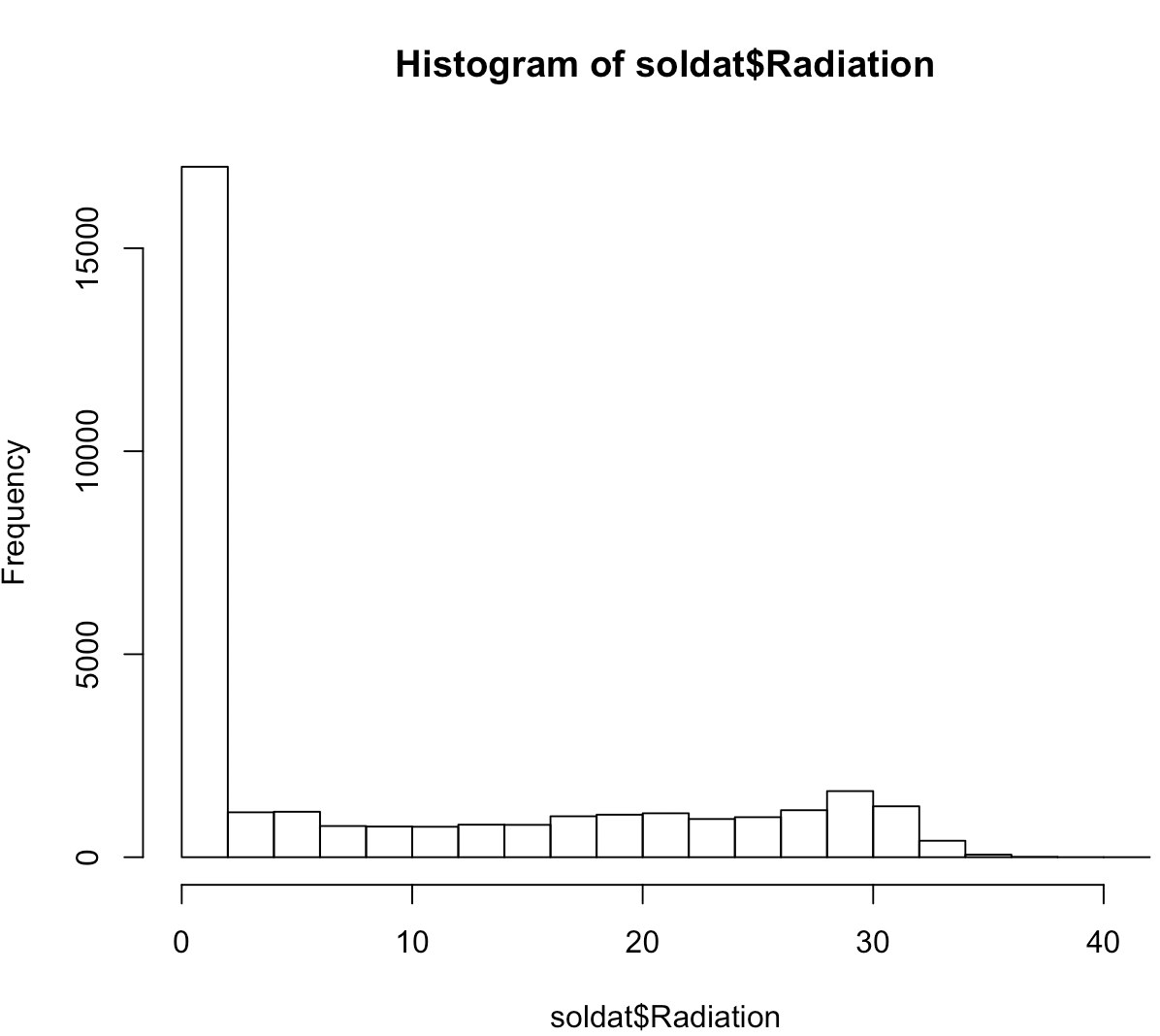
After performing the data optimization techniques, we performed the linear modelling techniques on the data set. For this, the data was divided into two parts: training data set and testing data set. The training data set consisted of 80% of the values of the original data set. A linear model was created for both data sets after the values for each being selected randomly. An accuracy of 59.13% was obtained from this model as seen from the R-squared value. The significantly lower p-value obtained in the summary suggested that there is a high significance between the dependent variable and the other independent variables. As the nature of the training model was considerably significant, the testing data set was regressed. The accuracy of this model obtained was 59.96%. Since, the p-value of all the variables obtained were significantly low, there was no need to remove any variables from the data set.

Also, the multicollinearity of the variables was checked by computing the vif (variance inflation factor) values. As none of the obtained values were found to be above 10, we can say that none of the independent variables were overfitting in the data set. As the accuracy of the testing model obtained was considerable, a predictive model was created which gave out an estimated radiation value of 835.0691 based on the input values of the independent variables.



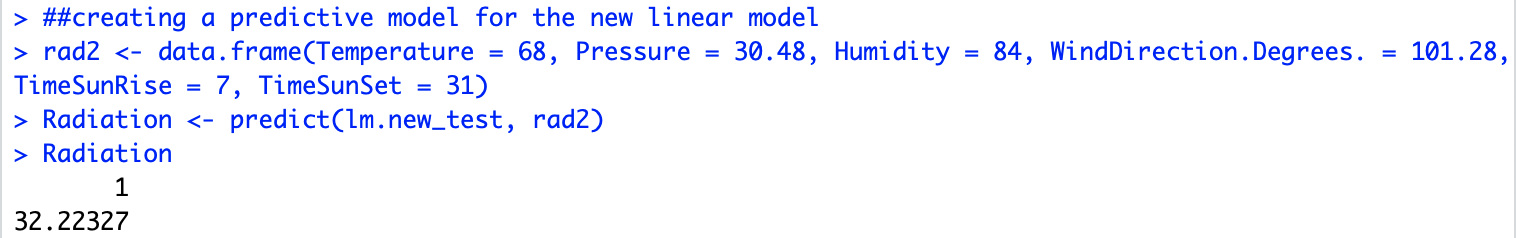
## Optimization of Linear Model

Now, to improve the accuracy of the model we have used a few model optimizing and data transforming techniques. As the dependent variable “radiation” is skewed towards the right, we will make use of square root transformations to normalize the data and reduce the range of data points.

From the above graphs\_\_\_\_\_\_\_\_\_\_.

A new linear model was created keeping the same 80:20 ratio for the train and test data set used for the earlier linear model. The accuracy obtained from the training data set of this new model was 64.04% as seen from the R squared value, which was an almost 5% increase in the accuracy from the initial training model. This was attained by applying square root transformation as it reduced the range of data points which normally distributed the dependent variable across all the data, leading to a higher accuracy. As all the p-values were now less than 2.2e-16 and the significance codes displayed “\*\*\*” for all the variables, we can now say that all the variables were highly statistically significant. The accuracy obtained from the testing data set was 62.94%, which was an almost 3% increase in the accuracy from the initial testing model. Also, as none of the VIF values obtained from both the models were found to be above 10, their corresponding variables were not overfitting to the data. Therefore, we created a new predictive model involving the said optimization techniques with an improved accuracy in their findings.



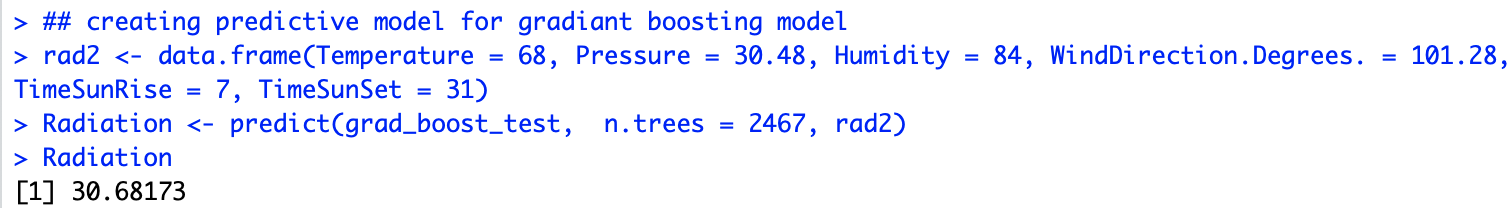
On running the predictive model, we got an estimated value of radiation as 32.22327, which was actually the square rooted value of the original radiation. Thus, we had to take the square of the value to obtain a value of 1038.33913. The value 12.97786 was the log value of the original price and thus, by taking the antilog of the value, we arrived at a value of 432,726. This value obtained from the new model with a higher accuracy was 203.27 Watts per square meter more than the original estimated value.

There was a change in the predicted value when the accuracy of the model was increased by 3%.

## Stepwise Regression and Gradient Boosting Model

In order to remove the redundant variables and to help in further optimization of the model, we regressed the model stepwise using backward selection technique to further reduce the irrelevant variables if any found. The variables will be reduced in the model till the threshold R2 value is attained. On performing the stepwise regression, it was found that none of the variables present in the data set were statistically insignificant and so none of them had to be dropped out of the data set.

Now, we used the gradient boosting method to further improve the accuracy of the model with initially considering 10,000 iterations for the training data set. It was observed that none of the variables had zero influence on the model and also that the temperature variable had the maximum influence in predicting the solar radiation level. Thus, it can be said that the temperature of the given area highly contributes in affecting the value of solar radiation of that same area. Also, it was observed that, to predict the values of \_\_\_\_\_\_\_\_\_\_\_ in the training data set, 10,000 iterations were required. On comparing the newly predicted values that were obtained with the optimal number of iterations with the original predicted values, it was found that the mean of both these values was almost equal. Now, similar techniques were applied on the testing data set and similar observations were recorded concerning the variables. The optimal number of iterations required to predict the radiation were 2,467 \_\_\_\_\_\_\_\_\_\_\_\_. We, then built a final predictive model based on the same values that were used in the previous models to compare the change in the value of radiation. Based on this gradient boosting model, we readjusted the data frame depending on the relative influence of the variables.



The value obtained on running. the model was the square root value of radiation. We, therefore, had to the take the square of this value and the new value obtained was 941.368556. On comparing the predicted value of both the gradient boosting and the original linear regression model, a difference of 106.299456 was observed. Thus, as the accuracy of the model is increased, \_\_\_\_\_\_\_\_\_\_\_\_\_\_.

# Conclusion