Success Rate Prediction for Mount Rainier Dataset

Aalap Doshi
Electrical and Computer Engineering
Arizona State University

Poojan Patel
Electrical and Computer Engineering
Arizona State University

Varadaraya Ganesh Shenoy
Electrical and Computer Engineering
Arizona State University

***Abstract*—Our motive and objective of this dataset is to predict the rate of success of the climb given the route and the weather condition of a day. We take the dataset from kaggle and implement Linear Regression, Polynomial Regression, and MLP by manipulating dataset. We calculate RMSE for Linear Regression.[Si20]**

***Index Terms*—Linear Regression, MultiLayer Perceptron, Polynomial Regression,RMSE**

I. Introduction

Regression analysis is a set of statistical processes to estimate the relationships between a dependent variable and one or more independent variables. One of the most popular implementation of regression analysis is linear regression, in which a researcher attempts to find a line (or a complicated linear combination) that fits the data most closely according to a specific mathematical criterion. This method of analysis is used for two conceptually distinct purposes. First, regression analysis is mainly used for prediction and forecasting, substantially overlapping with the field of machine learning. Second, it helps to infer causal relationships between the independent and dependent variables. Importantly, regressions by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed dataset.[Wikc]

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.[Wika]

Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y.

Multi-Layer Perceptron is a class of feed-forward artificial neural network. It typically refers to networks composed of multiple layers of perceptrons with threshold. Multilayer perceptrons with a single hidden layer are ccolloquially called "vanilla " neural networks. It usually contains at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node has a neuron that uses a nonlinear activation function.This model helps to distinguish data that is not linearly separable.[Wikb]

Root Mean Square Error (RMSE) is the standard deviation of the prediction errors. Prediction errors also referred to as,residuals are a measure of how far the data points are from the regression line.RMSE signifies the concentration of data around the line of the best fit. [SH]

In this project we have developed a linear regression model for the Rainier Dataset. Different parameters such as Temperature, Humidity, Wind Speed. The training of the model was done by using a package called Linear regression in python and then predicting the value of our data based on the actual data given. The accuracy of our predication based on actual data was done by measuring the amount of error. We measured different types of data such as root mean square error (RMSE), Mean square error (MSE) and Mean absolute error (MAE). By doing this error calculation we were able to determine how good our prediction is based on trained data.

In the second part of our project we trained a multilayer perceptron with two hidden layers with 5 neurons each. All the analysis such as Linear regression was done on multilayer perceptron and polynomial regression[Pol] also. The analysis of multilayer perceptron in python by using the package MLPRegressor [Mlp]. After training both the models, we compared the RMSE for the three models and saw how closely related they are.

II. Dataset and Model Parameters

The dataset used is the Mount Rainier Weather dataset with aprroximately 4000 samples. [**kaggle**]

In this paper, the features used for the analysis: Temperature,Humidity,Wind Speed. It is implemented on the linear regression model using [Sci].It is also implemented using polynomial regression [Pol] of degree 2. For the MLP model created using [Mlp] , all the features are used. Finally, RMSE is recorded for both models using Eq(1).

$$RMSE = \sqrt{\overline{(f-o)^2}} \tag{1}$$

where,

f = expected values,
o = observed values

III. Methods

Variables present in climbing statistics dataset are Date, Route, Attempted, Succeeded, Success Percentage. So the attempted means the people that tried to climb the mountain but failed to do so. That means the Attempted and Succeeded add up to the total number of people that climbed the mountain. We feel that each entry in the climbing dataset is the attempt to climb that route and depending on that we see whether the group has success or not. So, we have combined the data according to the Date stamp. This will be easy to merge the data as in the Rainier Weather dataset we have 464 rows while in the climbing, we have 4078 rows. So, merging it directly won't solve our problem. So, 'Date' is the object we need to convert it to time stamp. This will solve the problem and give 1895 entries.

The Linear Regression Model used for the analysis takes the form of Eq(2).

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon \quad (2)$$

where ,

i = 1,...,n

n = number of statistical units

$$y_i = \beta_0 + \beta_1 x + ... + \beta_p x^i + \epsilon \quad (3)$$

where ,

i = 1,...,n

n = number of statistical units

The MLP model used here has 15 neurons in its hidden layer to extract the features of the dataset.

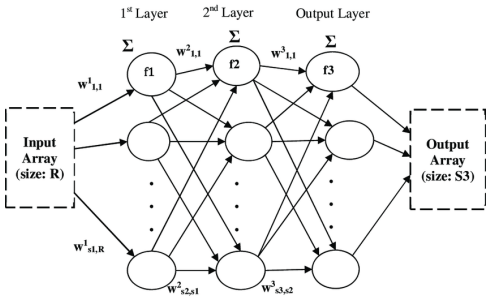

Fig. 1: MLP model with 15 neurons in hidden layer

IV. Results

Algorithm	Train-Test RMSE
Linear Regression	3.067
Polynomial Regression	3.071
MLP Regression	3.091

(a) RMSE for the Regression Models using 3 features

From metrics obtained, Linear Regression, Polynomial Regression and MLP model achieves the similar RMSE. Thus, for this dataset the linear regression model reveals better relation between dependent and independent variables.

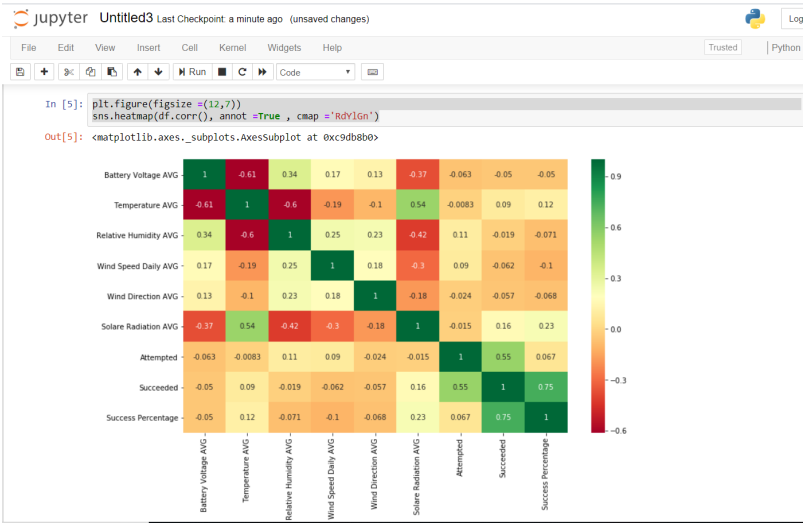

Fig. 2: Correlation Matrix of Dataset

```
]: #'Date' is object we need to convert it to time stamp
   df['Date'] = pd.to_datetime(df['Date'])
   df['Month'] = df['Date'].dt.month
   df.head()
```

]:

	Date	Battery Voltage AVG	Temperature AVG	Relative Humidity AVG	Wind Speed Daily AVG	Wind Direction AVG	Solare Radiation AVG	Route	Attempted	Succeeded	Success Percentage	Month
0	2015-11-27	13.643750	26.321667	19.715000	27.839583	68.004167	88.496250	Disappointment Cleaver	2	0	0.0	11
1	2015-11-21	13.749583	31.300000	21.690708	2.245833	117.549667	93.660417	Disappointment Cleaver	3	0	0.0	11
2	2015-10-15	13.461250	46.447917	27.211250	17.163625	259.121375	138.387000	Disappointment Cleaver	2	0	0.0	10
3	2015-10-13	13.532083	40.979583	28.335708	19.591167	279.779167	176.382667	Little Tahoma	8	0	0.0	10
4	2015-10-09	13.216250	38.260417	74.329167	65.138333	264.687500	27.791292	Disappointment Cleaver	2	0	0.0	10

Fig. 3: Modified Dataset

```
In [7]: df.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1895 entries, 0 to 1894
Data columns (total 12 columns):
Date                    1895 non-null datetime64[ns]
Battery Voltage AVG     1895 non-null float64
Temperature AVG         1895 non-null float64
Relative Humidity AVG   1895 non-null float64
Wind Speed Daily AVG    1895 non-null float64
Wind Direction AVG      1895 non-null float64
Solare Radiation AVG    1895 non-null float64
Route                   1895 non-null object
Attempted               1895 non-null int64
Succeeded               1895 non-null int64
Success Percentage      1895 non-null float64
Month                   1895 non-null int64
dtypes: datetime64[ns](1), float64(7), int64(3), object(1)
memory usage: 185.1+ KB
```

Fig. 4: Description of the Dataset

V. Conclusions

In this paper, the dataset is analyzed using regression models. The linear regression, polynomial regression, MLP Regression achieves similar RMSE. Although, this would be affected by other factors for that particular year the relation can help realize the significance of each feature to the success rate.Another regression model, Logistic Regression can also be applied. It is said that Logistic Regression is highly recommended when the response variable is categorical in nature, it can still be applied to investigate the relevance of these type of models to examine the correlation of the

dependent variable to independent variable.

REFERENCES

[Si20] Jennie Si. "Team Midterm Competition Project". In: *EEE 511- Spring 2020* (2020).

[Sci] *sklearn.linear$_m$odel.LinearRegression*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

[Mlp] *sklearn.neural$_n$etwork.MLPRegressor*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html.

[Pol] *sklearn.preprocessing.PolynomialFeatures*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html.

[SH] Statistics-HowTo. *RMSE-Root Mean Square Error*. URL: https://www.statisticshowto.datasciencecentral.com/rmse/.

[Wika] Wikipedia. *Linear Regression*. URL: https://en.wikipedia.org/wiki/Linear_regression.

[Wikb] Wikipedia. *Multilayer Perceptron*. URL: https://en.wikipedia.org/wiki/Multilayer_perceptron.

[Wikc] Wikipedia. *Regression Analysis*. URL: https://en.wikipedia.org/wiki/Regression_analysis.