IAF 603 Final Report

Investigating Mountain Climbing Success Rate Based on Weather Conditions.

Etibar Aliyev: **Etibar** worked diligently with **Amos** and **Monir** in cleaning the data and also helped with tidying the final report

Monir Almotairy: **Monir** worked alongside **Abimbola** in data exploration and analysis report writing. He also assisted Amos in the data cleaning process.

Amos Anzele: **Amos** helped in data cleaning alongside **Etibar** and also assisted **Chongyong** in SQL queries

Chongyong Lyu: Chongyong worked with Amos and Abimbola and **Etibar** in SQL queries and also worked on Spark in putting up the predictive model.

Ogungbire Abimbola: **Abimobla** helped **Chongyong** on the predictive model development and also worked with Monir in data exploration and final report writing.

N.B- This is a team work and everyone participated in ensuring the success of this project.

Investigating Mountain Rainier Climbing Success Rate Based on Weather Conditions

Etibar Aliyev (e_aliyev@uncg.edu), Monir Almotairy (mmalmota@uncg.edu), Amos Anzele (aaanzele@uncg.edu), Chongyong Lyu (c_lyu@uncg.edu), Ogungbire Abimbola (arogungb@uncg.edu)

Introduction

Exploration and discovery have been the driving force of humans and so many attempts have been made to reach the summit of Mount Rainier. Weather is believed to be a key factor in determining whether or not a climber makes it to the summit. In an effort to contribute to the body of knowledge in this field, we sought to understand the factors that impede the success of climbers in making it to the summit of Mount Rainier. Here is an attempt to use historical weather data and climbing records to make insightful meaning and answer basic research questions about the data. This project provides mountaineers additional information on which weather condition is best for mountain climbing. Data visualization and statistical methods were used to determine patterns among variables of interest. More specifically we addressed the following research question: Which weather condition is the most influential factor for the success of climbing?

The Data

Mount Rainier is one of the top places that climbers visit in the west side of the US. Two datasets on weather/climate conditions and the number of climbing attempts of Mount Rainier were collected via Kaggle. Thus, our team saw it a great opportunity to conduct this project using these two datasets.

The weather dataset has 465 observations on the following weather factors: battery voltage, temperature, relative humidity, wind speed, wind direction, and solar radiation. Each observation is a date, and the average of each of these factors was reported. The climbing data has 4,078 observations on the climbing route, number of overall attempts, number of successful attempts, and the success rate - which was calculated in the dataset by dividing the number of successful attempts by the number of overall attempts. The weather data was originally available at the Northwest Avalanche Center website, and the climbing data was provided at the Mount Rainier National Park Climbing and Mountaineering website.

Methods

Both datasets were provided through the Kaggle website. Thus, the datasets were downloaded (extracted) on a shared group file between the team members. No metadata were provided at Kaggle for any of the two datasets, but the simple structure for each dataset was discussed between group members.

A common variable that exists in both datasets was the date (the climbing date in the climbing dataset, and the weather date in the weather dataset). Thus, this variable was used as a foreign key, in each of the datasets, to integrate/join both datasets in one table. Both datasets were integrated using this common variable via SQL. The joining approach we used is the inner joining as we only want dates that have data on climbing and weather datasets, thus the unmatched data were not included in the new SQL table.

Missing data was checked as part of the data preparation, and the data that were matched into the new SQL table does not have missing data. Heading, tail, and structure commands were used to understand the structure of the data. We checked normality by producing a histogram for each variable, and we checked for extreme outliers through boxplots. All of the variables were not normally distributed (either right or left skewed), thus a standardization was used for these variables to bring them all to a universal scale. Standardized variables showed normality when plotted, and no extreme outliers were noticed.

During data exploration, several unusual data were found which have succeeded rate higher than 1.0, this value seem impossible, so we deleted the unusual data. In addition we found several route name typos treated during the data cleaning process.

Analysis Section

Visualization

The plot below in Fig. 1 shows that the most used route is Disappointed Cleaver.

Fig. 2 shows the route with the highest average success rate. Tahoma cleaver has a 100% success rate, although it is not often attempted.

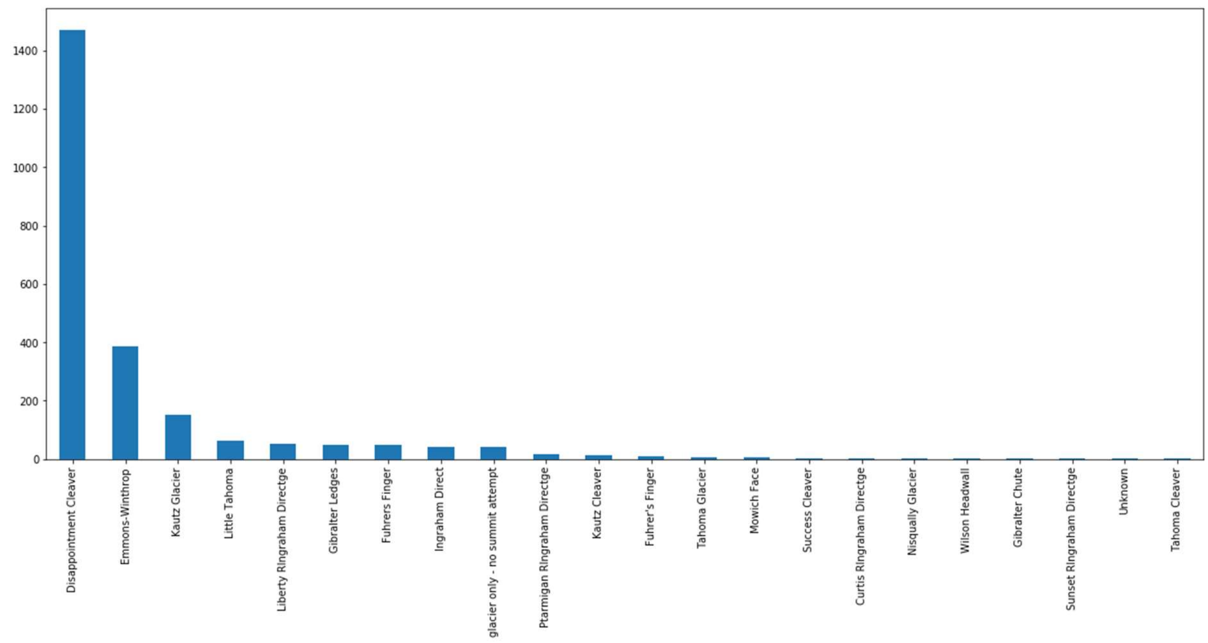

Fig. 1: frequency of route use

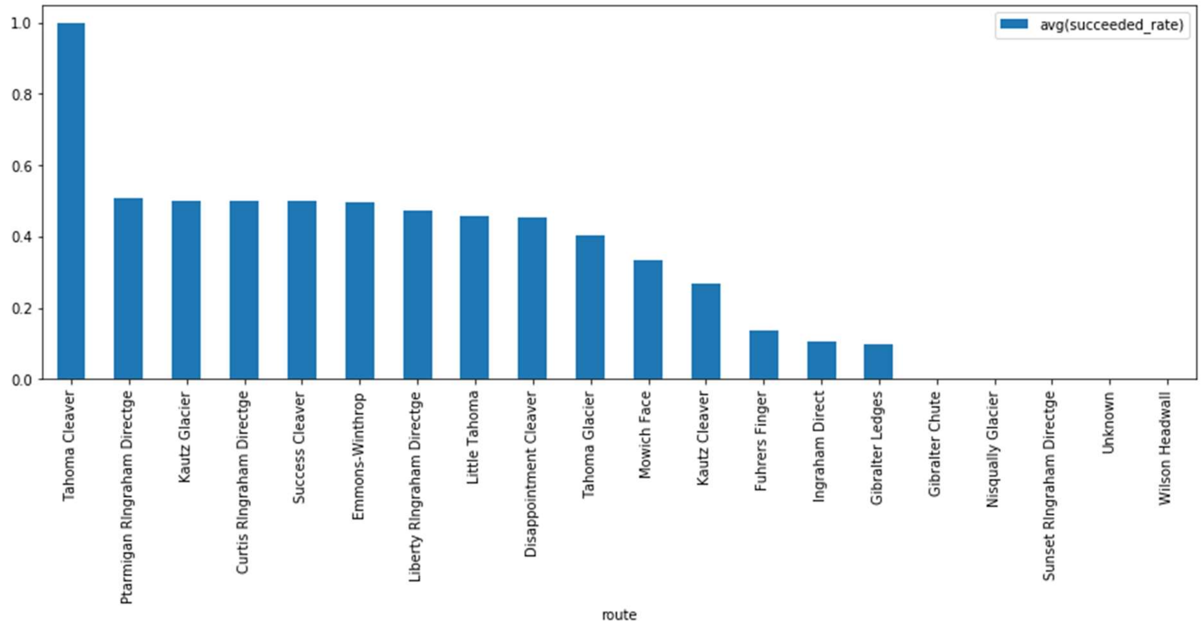

Fig. 2: Average success rate of different routes

The table below shows the result of the route with the highest success of mountain climbing. While Disappointed Cleaver is the best attempted and succeeded route in Rainier mountain climbing. It doesn’t have the highest success rate as the number of failed attempt is enormous

	Route	*Sum(attempted)*	*Sum(succeeded)*	*Avg(succeeded_rate)*	*Wind_direct*
1	**Curtis Ringraham Directge**	**8.0**	**4.0**	**0.500000**	**N**
1	**Dissapointment Cleaver**	**14929**	**7228**	**0.454**	**N**
2	**Emmons-Winthrop**	**2907.0**	**1557.0**	**0.494**	**N**

Table 1: sum of succeeded attempt in different route

Underlying Assumptions

Prior to developing our model, the data was visualized to ensure that the assumptions underlying the multi-linear regression model are strictly adhered to. The distribution of each variable was assessed to ensure the normality, homoscedasticity and independence assumption is reasonable for the data available. Fig. 3 shows the boxplot of the variables of concern. The eyeball inspection of this figure indicate that the normality and equal variance assumption does not seem reasonable, hence, transformation of the variables will be useful.

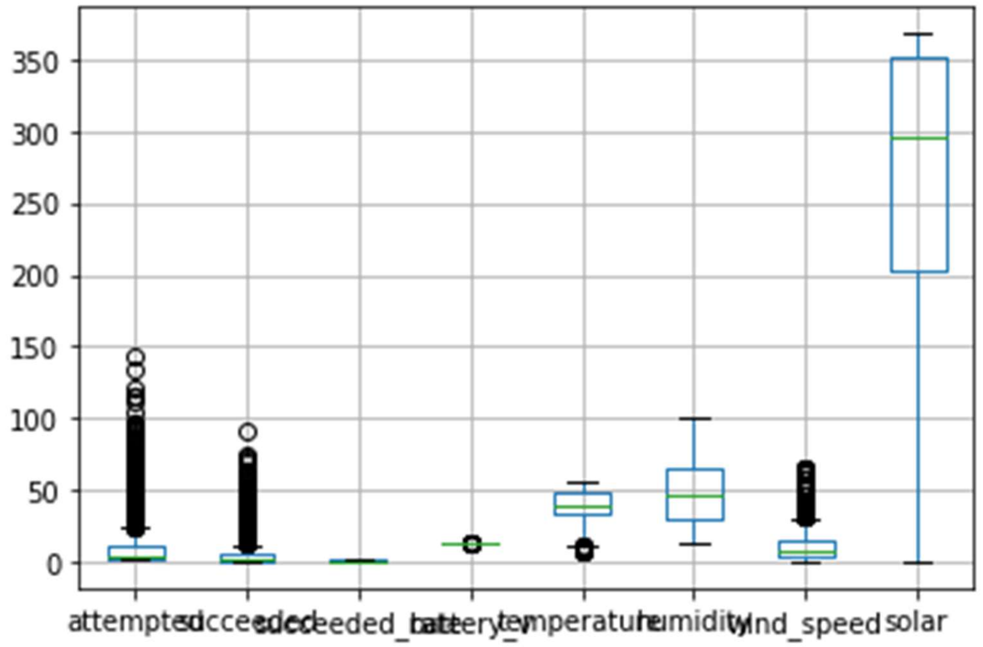

Fig. 3: Box plot before transformation

The standardization of variables was done to bring all variables to the same range. The boxplot following the transformation of data is as shown below.

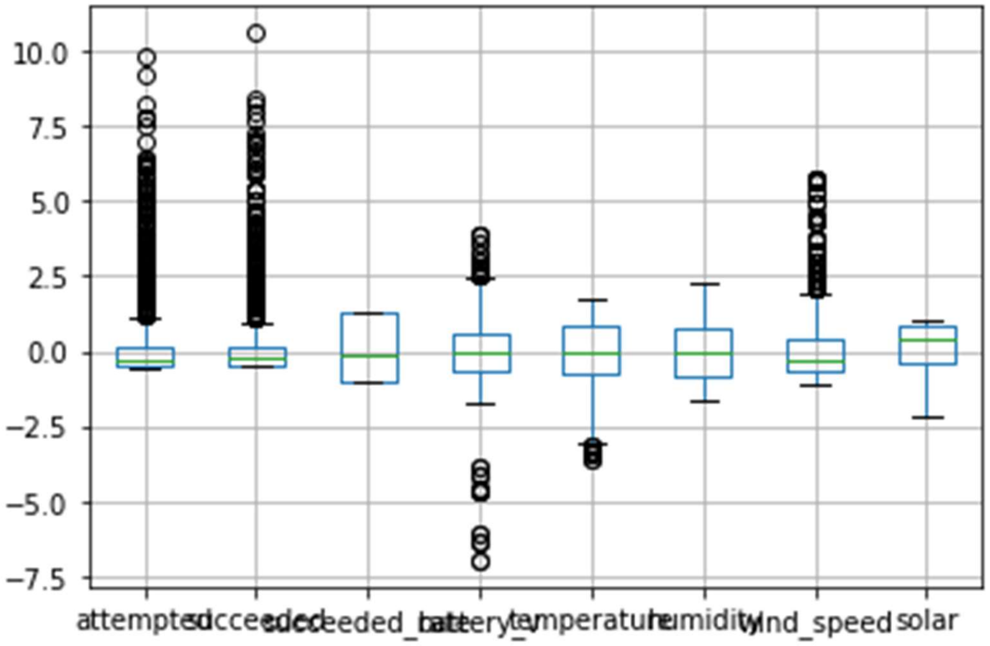

Fig. 4: Boxplot after transformation

Model Development

In order to ascertain which variable is important to succeed in mountain climbing, the relationship between all explanatory variables of concern is made against the response variable (success rate). Fig. 5 below shows the correlation between all variables. Inspection of this plot show that there is a slight moderate relationship between success rate and other predictor variables. Thus a multiple regression analysis will be adequate in predicting the success rate of mountain climbing using these weather parameters.

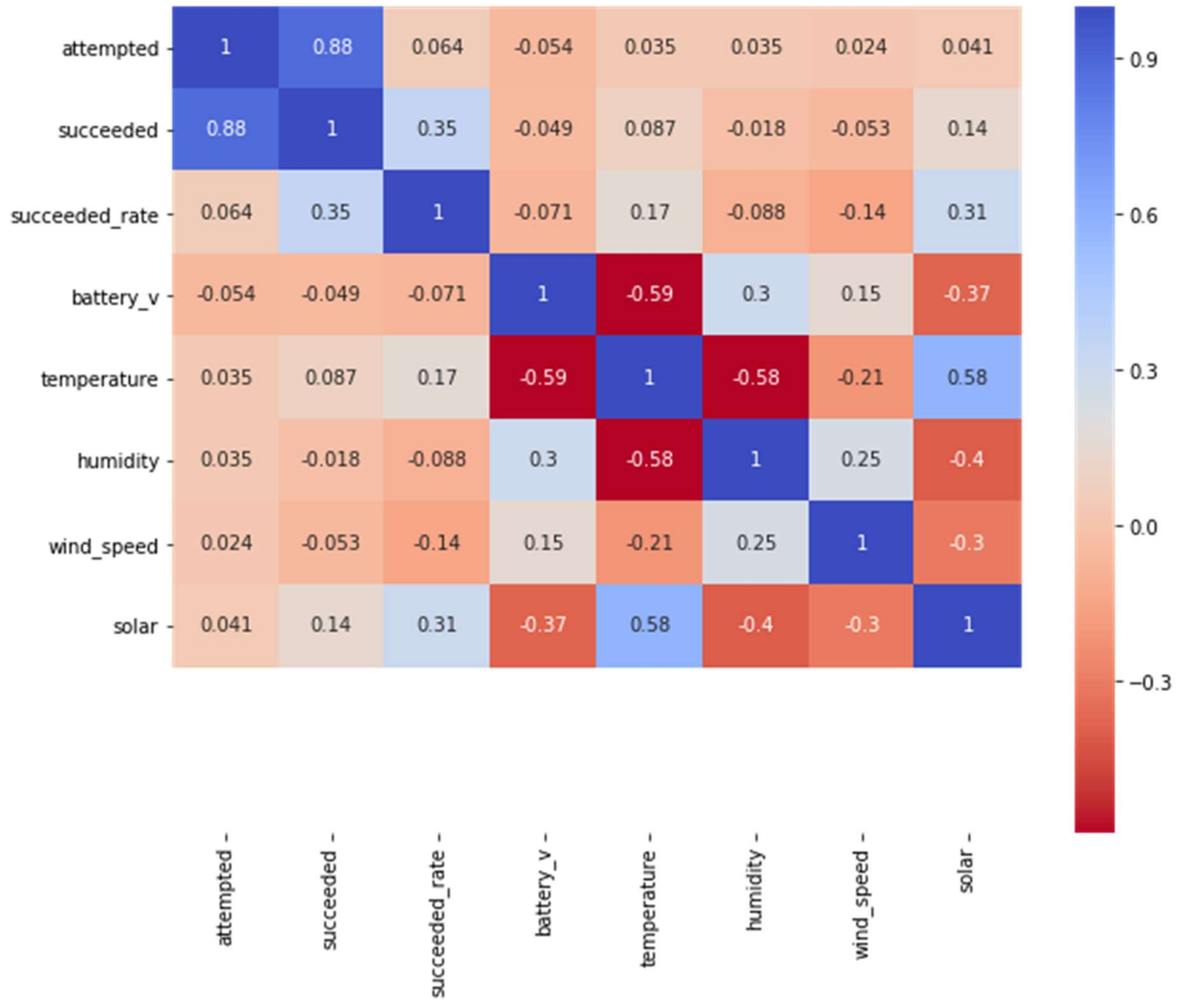

Fig. 5: Correlation matrix plot of all variables.

This plot also enable us to eliminate redundant variable by checking highly correlated explanatory variables in order to avoid multi-collinearity among variables.

The regression model:

$$success\ rate = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

The estimated coefficient from the regression analysis is given as:

```
The coefficient of the model is : DenseVector([0.3508, 0.0025, 0.0015,
-0.002, 0.0012])
```

```
The Intercept of the model is : -4.748641
```

The above result represent the corresponding values of α and β_1 to β_5 (Battery, temperature, humidity, wind_speed and solar) respectively. Below is the final regression model

$$success\ rate = -4.748641 + 0.3508 * Battery + 0.0025 * temperature + 0.0015 * humidity - 0.002 * wind_{speed} + 0.0012 * solar$$

Prediction

The regression model is further used for prediction using the test data to obtain the following result. Below is the result following our prediction using the test data. N.B- The display shows the result of only the top 20 rows.

```
--------------------+--------------+--------------------+
|          Attributes|succeeded_rate|          prediction|
+--------------------+--------------+--------------------+
|[13.16,32.31,100....|           0.0|0.053335540858988395|
|[13.3775,51.13958...|   0.916666667| 0.44304820280809665|
|[13.3775,51.13958...|           1.0| 0.44304820280809665|
|[13.39666667,54.9...|           0.0|  0.5472268539592431|
|[13.39666667,54.9...|           0.0|  0.5472268539592431|
|[13.39666667,54.9...|           1.0|  0.5472268539592431|
|[13.39666667,54.9...|           1.0|  0.5472268539592431|
|[13.4,24.97708333...|           0.0| 0.38229071393875813|
|[13.4,24.97708333...|           0.0| 0.38229071393875813|
|[13.4,24.97708333...|           0.0| 0.38229071393875813|
|[13.40291667,53.1...|           1.0| 0.39476630826283277|
|[13.41041667,51.1...|           0.0| 0.46103300953510207|
|[13.41125,56.1537...|           0.0| 0.49704763865792767|
|[13.41125,56.1537...|           0.0| 0.49704763865792767|
|[13.41125,56.1537...|           0.0| 0.49704763865792767|
|[13.41125,56.1537...|           0.0| 0.49704763865792767|
|[13.41125,56.1537...|   0.583333333| 0.49704763865792767|
|[13.41125,56.1537...|           1.0| 0.49704763865792767|
|[13.41125,56.1537...|           1.0| 0.49704763865792767|
|[13.41125,56.1537...|           1.0| 0.49704763865792767|
+--------------------+--------------+--------------------+
```

Model Evaluation

In order to know how effective our model is for prediction, model evaluation is needed. We will be making use of the RMSE and R^2 value to know the residual standard error and how well our model fit the data respectively.

Following the evaluation of the model, the following results were obtained in term of RMSE and R^2 value.

RMSE:	0.420
R^2:	0.113

Table 2: RMSE and R^2 value of the Regression Model

RMSE value of 0.420 shows that the model didn't do a good job in predicting the test values right. The aim is to achieve the least root mean square error possible.

Subsequently, an R^2 value of 0.113 indicate that only 11.3% of the variability in the response variable is being explained by the explanatory variable. This is not a very good model as much because a lot of the variability in the response variable remain unexplained.

Because of the high unexplained variability in the response variable, it is pertinent that we discuss the limiting factors involved. It only make sense that weather factors will not be able to tell much about the success rate of climbing mountains. While weather parameters are good factor to consider while predicting the success rate of climbing mountains, various factors are as well important. Factors such as physical data about climbers such as weight, height and general physique information.

Conclusion

The result of the analysis shows that that solar radiation is statistically significant in predicting the success rate of climbing mountains. Although, it may not be practically significant, it is very important in predicting the success rate of climbers.

Our analysis shows that the most attempted route is Disappointed Cleaver. It also have the highest number of success in term of mountain climbing. However Disappointed Cleaver does not have the highest success rate due to a large number of failed trial in climbing mountain Rainier through this route. The route with the highest success rate is therefore Tahoma cleaver with a 100% success rate.

Our multi-linear regression model is not as good as expected and various other factors would be needed to account for the unexplained variability in success rate.