A Machine Learning Based Approach, For Predicting Road Closure Events, Given Data of the US Road Construction and Closure

Authors

- **Amirthavarshini Muraleetharan**
  [677-010-487](#) · [amirthavarshini246](#)
  Department of CEE, University of Illinois Urbana Champaign

- **Thomas Ngare**
  [652-601-317](#) · [thomasNg](#)
  Department of CEE, University of Illinois Urbana Champaign

- **Kapil Shah**
  [668-376-620](#) · [kapilrs2](#)
  Department of CEE, University of Illinois Urbana Champaign

Abstract

A nationwide dataset of road construction and closure events, including data from 49 US states is chosen for the project. The roadwork included in this dataset's construction events ranges from minor paving repairs to significant undertakings that might take months to complete. Several APIs that provide streaming traffic incident (or event) data are used to collect the data between January 2016 and December 2021. These APIs transmit traffic information gathered by several organizations, including the US and state departments of transportation, law enforcement organizations, traffic cameras, and traffic sensors embedded in the road networks. The number of construction and shutdown records in this dataset currently stands at roughly 6.2 million.

In general, this dataset can be used for a wide range of applications, including the prediction of short- and long-term road construction, the prediction of road closures, the study of the life cycle of road construction, the development of insights to help city planners choose construction sites wisely with the most negligible negative impact on traffic flow, and the investigation of the influence of precipitation or other environmental stimuli on the need for road work. The dataset is being updated on an annual basis. The data will be obtained from US Road Construction and Closures (2016 - 2021), Kaggle, and it is available in CSV format. Presently, the dataset contains 6,170,627 observations comprising of features like Construction severity, Latitude and longitude, Precipitation, Traffic signal and many such taking a total of 47 columns. Table 1 elaborates the specifics of this data set. Using this dataset, a machine learning model will be developed to predict the duration of a new road construction project as short, medium, or long term, given inputs of pertinent features derived from Table 1. The developed model will be cross validated in four (4) folds, to be made suitable for accurate and robust predictions. With this model, it is envisaged that contractors, city planners and relevant authorities can categorize potential road construction projects based on expected average weather conditions, for better planning and project delivery.

To achieve this goal, data wrangling will be performed. The essential data frames for the study will be extracted from the original dataset followed by exploratory data analysis (EDA). EDA will enable us to derive insights by forming a pattern for better visualization and exploration. Based on this, pertinent features will be realized to build a classification model that accurately predicts the duration of a road construction project.

Features	Description
ID	Unique identifier of construction record
Severity	Shows the severity of the construction
Start and End Time	Shows the start time of construction
End Time	Shows the end time of construction
Latitude and Longitude	Shows the GPS coordinates
Distance	The length of the road extent affected by the construction
Street Details	Shows the street number, name and right/left side in address field
Address Details	Shows the city, county, state, country and zip code in address field
Time zone	Shows time zone based on the location of the construction event
Weather	Shows the time stamp of weather observation record
Temperature, Wind, Humidity, and Pressure	Shows the temperature, wind chill, humidity, and pressure

Features	Description
Visibility	Shows visibility
Wind Direction and Speed	Shows wind conditions
Precipitation and Weather condition	Shows precipitation and weather condition
Amenity	An annotation which indicates presence of amenity in a nearby location
Bump and Crossing	Annotations which indicate presence of speed bump or hump and crossings
Give way, Junction, railway	Annotations which indicate presence of give way, junction and railway
Exit, Roundabout, Station, Stop	Annotation which indicates presence of no exit, railway, roundabout, and station
Traffic Details	Annotations which indicate traffic calming, signal, turning loop
Light Details	Annotations which indicate sunrise, sunset, civil twilight, nautical twilight, astronomical twilight

Table: Description of Undertaking Dataset {#tbl:features}

2.0 Data Wrangling and Exploratory Data Analysis The data obtained from Kaggle is highly generic and unstructured, hence unsuited for analysis in its raw form. To achieve the goals of our design, an initial data wrangling was performed to put the data in the right format followed by EDA to determine pertinent features that affect road construction duration in USA.

2.1.1 Data Cleaning The csv file obtained from Kaggle was composed of over 6.1 million observations, which would have been computationally expensive to work with. To reduce the burden of computational time, while still retaining a good representation of the data, 1,048,575 observations (15% original data) were randomly selected with R. The data was converted to a data frame format suitable for analysis with Julia. However, some features were discovered with missing entries, which would be adversarial to subsequent codes and functions. To resolve this problem, a package in Julia (Missing package) was leveraged to filter out all observations with one or more missing entries. This process further reduced the size of the dataset to 482,849. Then, an intuitive search was made of pertinent features that could affect road construction duration, like the length of the road extent affected by the construction (denoted by "Distance" in miles), the total amount of precipitation, and other environmental factors. Although some of these pertinent features were discovered, most of the entries of each observation of these features were not in the format suitable for visualization (which requires real numbers) i.e., some variables were Strings, or Boolean values, hence the need to clean the data. Table 2 below shows a segment of the dataset after it has been roughly sampled, and missing entries removed, while Table 3 shows the summary of the dataset.

Table 2: Dataset of road construction projects in USA

Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Pressure(in)	Visibility(mi)	Wind_Direction	Wind_Speed(mph)	Precipitation(in)
String31											
"4/5/2019 16:00"	"9/29/2020 11:53"	32.8384	-93.1524	32.8507	-93.1644	1.1035	29.72	10.0	"S"	3.0	0.0
"11/12/2021 7:59"	"11/12/2021 8:22"	30.2213	-92.0086	30.2166	-92.0038	0.433173	30.09	3.0	"CALM"	0.0	0.0
"10/12/2021 7:17"	"10/12/2021 9:18"	39.6532	-104.91	39.6531	-104.914	0.192266	24.09	10.0	"WSW"	5.0	0.0
"2/10/2021 2:46"	"2/17/2021 23:59"	33.9615	-118.029	33.9619	-118.029	0.0321121	29.92	9.0	"CALM"	0.0	0.0
"12/6/2021 7:50"	"12/6/2021 9:53"	27.8957	-82.7872	27.8946	-82.7853	0.141399	30.12	1.0	"SE"	5.0	0.0
"9/30/2021 1:39"	"9/30/2021 3:40"	41.977	-87.9398	41.9759	-87.9398	0.0732397	29.39	10.0	"CALM"	0.0	0.0
"9/16/2021 15:08"	"9/16/2021 15:40"	38.9209	-77.0365	38.9333	-77.0365	0.853305	30.09	10.0	"N"	20.0	0.01
"12/1/2021 8:51"	"12/1/2021 10:56"	40.8683	-73.9191	40.867	-73.9231	0.226995	30.15	10.0	"W"	16.0	0.0
"7/31/2021 12:11"	"7/31/2021 13:06"	33.4831	-112.074	33.4793	-112.074	0.258762	28.83	10.0	"WNW"	8.0	0.0
"11/19/2021 14:06"	"11/19/2021 16:27"	33.4949	-112.028	33.4954	-112.03	0.130887	28.81	10.0	"ESE"	6.0	0.0
"7/22/2021 10:45"	"7/22/2021 11:31"	31.7066	-92.2451	31.7057	-92.2348	0.607945	30.04	10.0	"W"	9.0	0.0

Table 3: A summary of pertinent information of road construction projects in USA

	variable	mean	min	median	median	max	nmissing	eltype
1	:ID	nothing	"C-1"	nothing	nothing	"C-999999"	0	String15
2	:Severity	2.25385	1	2.0	2.0	4	0	Int64
3	:Start_Time	nothing	"1/1/2016 3:01"	nothing	nothing	"9/9/2021 9:59"	0	String31
4	:End_Time	nothing	"1/1/2017 0:45"	nothing	nothing	"9/9/2021 9:58"	0	String31
5	:Start_Lat	37.2934	21.3551	39.2578	39.2578	49.0008	0	Float64
6	:Start_Lng	-91.436	-157.939	-85.6933	-85.6933	-67.0758	0	Float64
7	:End_Lat	37.1675	21.3548	39.198	39.198	49.005	107566	Union{Missing, Float64}
8	:End_Lng	-91.3286	-157.94	-85.6538	-85.6538	-67.0747	107566	Union{Missing, Float64}
9	Symbol("Distance(mi)")	1.2595	0.0	0.325327	0.325327	424.374	0	Float64
10	:Description	nothing	" and will be enforced by local law en	nothing	nothing	"working behind barrier walls. CHARLES	0	String
⋮	more				more			
47	:Astronomical_Twilight	nothing	"Day"	nothing	nothing	"Night"	99	Union{Missing, String7}

2.2 Wrangling and feature derivation It can be seen from Table 2 that the project duration is not explicitly stated. Hence, the project duration was defined to be the difference between the start and end time, after preprocessing the String data entries to a Date format. Table 4 shows the resulting features after preprocessing.

Table 4: Derivation of project duration from raw data

	Data before wrangling				*Data after wrangling*	
Start_Time	End_Time	Distance(mi)		Start_Date	End_Date	Project_Duration_Days
"4/5/2019 16:00"	"9/29/2020 11:53"	1.1035		2019-04-05	2020-09-29	544
"11/12/2021 7:59"	"11/12/2021 8:22"	0.433173		2021-11-12	2021-11-12	1
"10/12/2021 7:17"	"10/12/2021 9:18"	0.192266		2021-10-12	2021-10-12	1
"2/10/2021 2:46"	"2/17/2021 23:59"	0.0321121		2021-02-10	2021-02-17	8
"12/6/2021 7:50"	"12/6/2021 9:53"	0.141399		2021-12-06	2021-12-06	1
"9/30/2021 1:39"	"9/30/2021 3:40"	0.0732397		2021-09-30	2021-09-30	1
"9/16/2021 15:08"	"9/16/2021 15:40"	0.853305		2021-09-16	2021-09-16	1
"12/1/2021 8:51"	"12/1/2021 10:56"	0.226995		2021-12-01	2021-12-01	1
"7/31/2021 12:11"	"7/31/2021 13:06"	0.258762		2021-07-31	2021-07-31	1
"11/19/2021 14:06"	"11/19/2021 16:27"	0.130887		2021-11-19	2021-11-19	1
"7/22/2021 10:45"	"7/22/2021 11:31"	0.607945		2021-07-22	2021-07-22	1

```julia
begin
    dtf = DateFormat("m-d-y")
    Start_date = Vector{Date}()
    End_date = Vector{Date}()
    for i in 1:length(df.Start_Time)
        k = findfirst(" ", df.Start_Time[i])
        k2 = findfirst(" ", df.End_Time[i])
        k = k[1]
        k2 = k2[1]
        push!(Start_date, Date(replace((df.Start_Time[i][1:k-1]), "/" => "-"),dtf))
        push!(End_date, Date(replace((df.End_Time[i][1:k2-1]),"/" => "-"),dtf))

    end
    df.Start_Date = Start_date
    df.End_Date = End_date
end
```

Julia code snippets 1

Having computed the project durations as shown in Table 4, some anomalies were detected in the data. The 2nd and 3rd observations show that about 0.43mi and 0.19mi (692m and 305m) of road span was constructed in 1day respectively, whereas about 0.03mi (or 48m) was constructed in 8days as shown in the 4th entry which is unrealistic. To resolve this, a further investigation was done on the narrative of the dataset from Kaggle, and it was deduced that some of the observations were just minor repair works on existing roads (which would not take a long time irrespective of the road span) while others were new construction projects which takes longer to complete. Unfortunately, there exist no feature in the dataset that reveals if an observation was a minor repair task or a major construction task. To circumvent this, observations corresponding to project durations less than 50 days were filtered out, resulting in a dataset of 43,134 observations significantly dominated by new road construction projects or at least projects lasting longer than 50days. Furthermore, three (3) categories of projects durations were designated as short-, medium- and long-term new road construction projects. A short-term new road construction project is defined as one lasting less than 100 days, while a medium-term project lasts between 100-300 days, and a long-term project lasts longer the 300 days. Table 5 shows the resulting data after grouping into the specified classes.

Table 5: Dataset for new road construction projects in US, or projects exceeding 50days of duration

Start_Lng	End_Lat	End_Lng	Distance(mi)	Start_Date	End_Date	Project_Duration_Days	Project_Duration_Class
				Date	Date	Int64	Any
-93.1524	32.8507	-93.1644	1.1035	2019-04-05	2020-09-29	544	"long term"
-83.2652	39.7261	-83.2465	1.10024	2020-04-20	2020-07-22	94	"short term"
-73.9652	40.763	-73.974	0.952071	2021-05-10	2021-12-10	215	"medium term"
-80.1912	26.1871	-80.1874	0.24557	2020-11-15	2021-12-30	411	"long term"
-87.724	41.8312	-87.7242	0.453084	2021-07-04	2021-10-08	97	"short term"
-86.4798	39.6149	-86.4798	0.601812	2021-03-10	2021-12-31	297	"medium term"
-112.063	33.5092	-112.066	0.173348	2021-02-27	2021-12-28	305	"long term"
-80.1906	25.8904	-80.1845	0.380199	2021-04-13	2021-12-13	245	"medium term"
-80.3374	25.7482	-80.3356	0.118257	2021-06-16	2021-12-14	182	"medium term"
-90.7869	42.44	-90.8007	0.701292	2021-05-14	2021-10-22	162	"medium term"
-92.4132	41.0376	-92.4158	0.999269	2020-07-20	2020-10-31	104	"medium term"

2.2.1 Feature detection and EDA

The resulting dataset as presented in Table 5, was processed to reveal key features. A rough guess was made that the numeric features quantifying environmental conditions like temperature, humidity, precipitation, wind speed, and pressure together with the road construction span or "Distance" affects the proje t duration. This can be seen from the histogram plots of these features, as there exit a considerable spread in their distributions. It was further assumed that the average amount of these quantities (e.g., Temperature) was recorded during the entire project duration, as this information was not explicitly stated in the dataset description. Thus, given the expected or average environmental conditions, and the span or extent of the road construction, the developed model is expected to predict the class of the completion time of the project as short, medium or long term. Table 6 shows the summary of the initial features selected for the development of the model.

Table 6: Extracted features for EDA

	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)	Project_Duration_Days
1	1.1035	75.0	75.0	58	29.72	10.0	3.0	0.0	544
2	1.10024	70.0	70.0	60	28.89	10.0	5.0	0.0	94
3	0.952071	48.0	46.0	93	29.72	2.0	5.0	0.01	215
4	0.24557	72.0	72.0	78	29.88	10.0	8.0	0.0	411
5	0.453084	64.0	64.0	87	29.37	4.0	0.0	0.0	97
6	0.601812	73.0	73.0	87	29.1	7.0	6.0	0.0	297
7	0.173348	50.0	50.0	31	28.68	10.0	6.0	0.0	305
8	0.380199	71.0	71.0	84	29.95	10.0	0.0	0.0	245
9	0.118257	79.0	79.0	66	30.17	10.0	12.0	0.0	182
10	0.701292	41.0	41.0	65	29.15	10.0	3.0	0.0	162
⋮ more									
43134	0.999269	70.0	70.0	90	29.09	10.0	0.0	0.0	104

An indispensable aspect of EDA is to detect multicollinearity and prevent confounding in the modeling. Julia's "Statistics" package was leveraged to compute and plot the correlation between all the independent (selected features) and the dependent variable (project duration). This plot not only enabled discovery of statistically related features, but also enabled realization of features that are pertinent to predicting the dependent variable. The figures below show the correlation plots for the selected features.
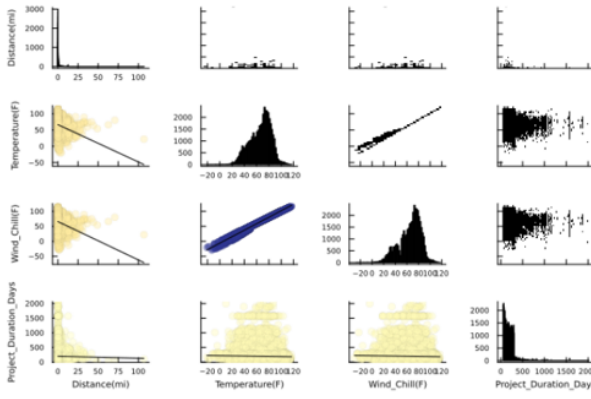
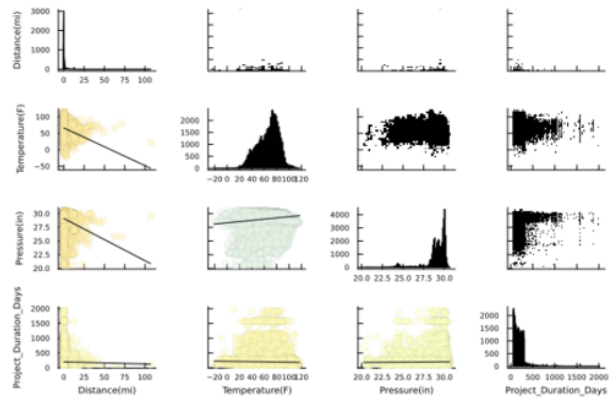Fig. 1: Correlation plot for Distance, Temperature, Wind chill and Project Duration

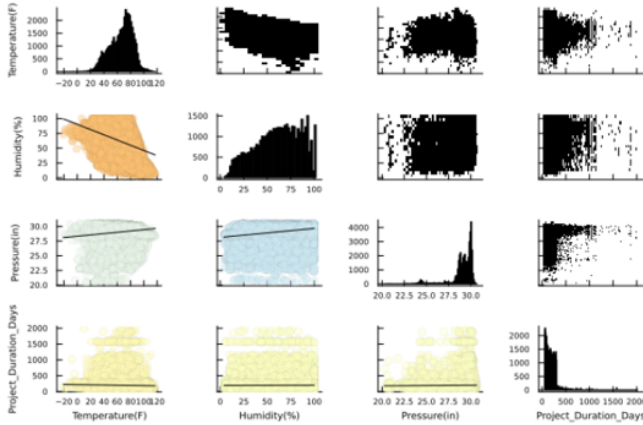Fig. 2: Correlation plot for Distance, Temperature, Pressure and Project Duration

Fig. 3: Correlation plot for Temperature, Humidity, Pressure and Project Duration
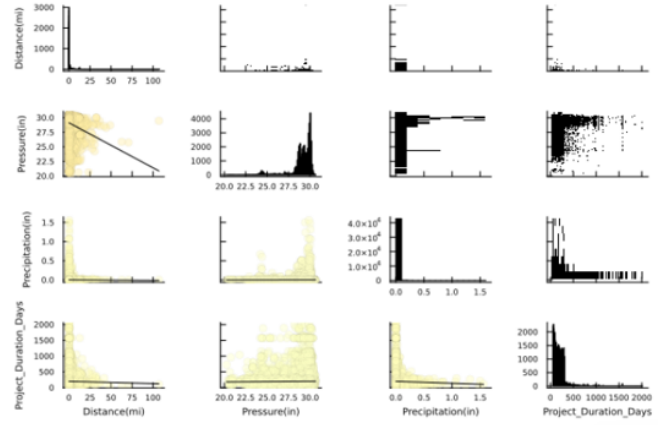

Fig. 4: Correlation plot for Distance, Pressure, Precipitation and Project Duration

The "Combinatorics" package in Julia was leveraged to create three (3) combinations of all features plus the dependent variable in order to generate the correlation plots that enabled the determination of the most relevant statistical features. A total of 56 combinations were generated and plotted in Julia, but due to space constraint, only some of the plots are presented in the figures above. The code snippet below illustrates how the figures above were generated.

```
begin
    arr = collect(1:8)
    plots = []
    combinations_arr = collect(combinations(arr, 3))
    for i in combinations_arr
        push!(i, 9)
        push!(plots, @df df_sel corrplot(cols(i), grid = false, size(10000,8000),
        xtickfontsize=5, ytickfontsize=5, xguidefontsize=6, yguidefontsize=6))
    end
    plots
end
```

Julia code snippets 2

It can be deduced from fig. 1 that a strong correlation exists between Temperature and Wind chill, hence Temperature was retained for the model development, while Wind chill was eliminated by choice and convenience. Following the correlation plots above, the scatter plots in figure 5, show promise in the development of classification tree networks.
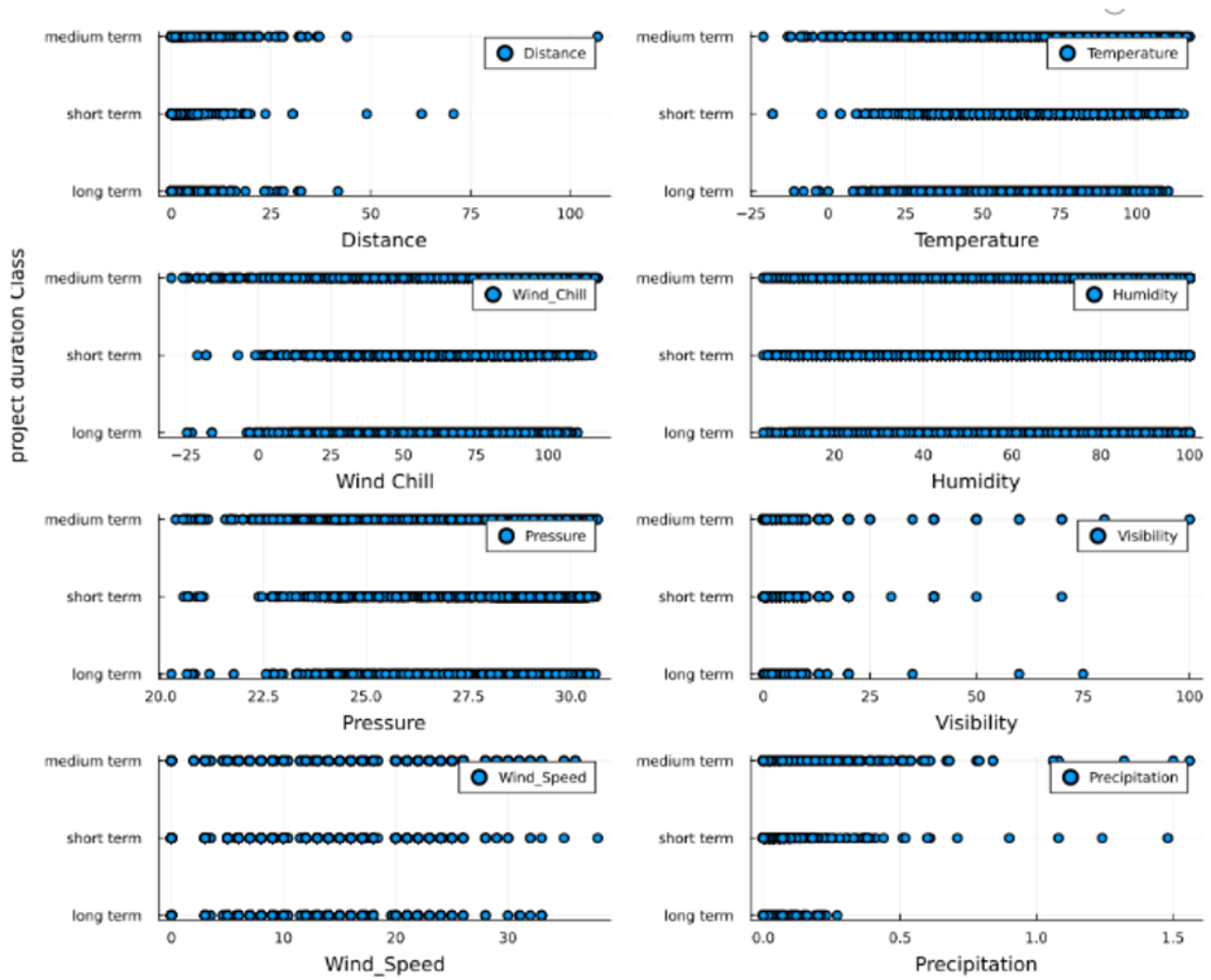
Fig. 5: Scatter plots of pertinent features against Project Duration Class

Figure 5 revealed Humidity as a less promising feature for the project task because a clear cut would be difficult to achieve to develop classification trees. The following box plots in figure 6 reveal outliers in all the feature space except for Humidity. This information would guide in the selection of locations for cuts in creating decision trees. A further investigation of the statistical significance of the selected features would be done through principal component analysis (PCA). PCA will reveal the directions with most significant variance and enable the dimensional reduction of the model.
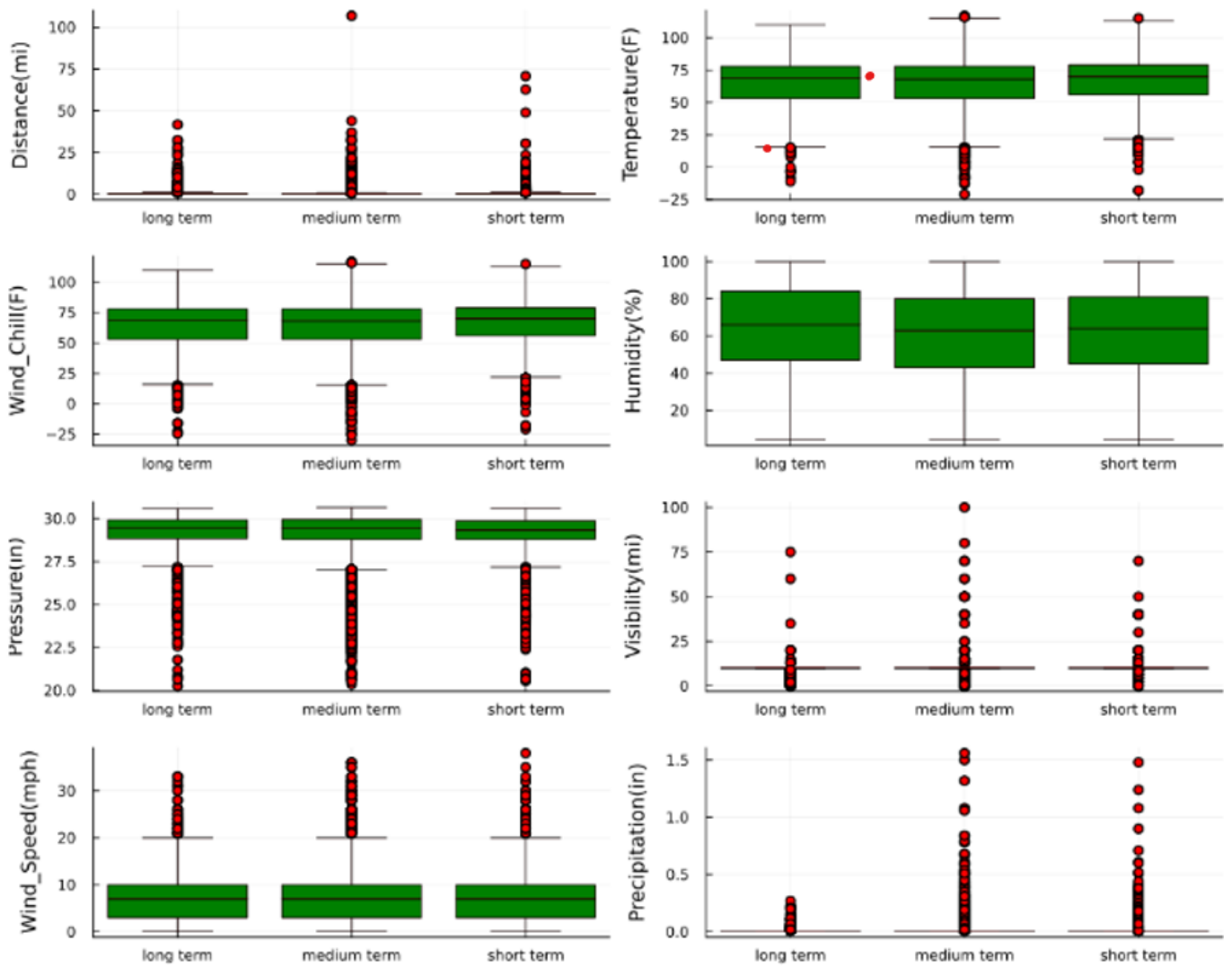
Fig. 6: Box plots for outlier detection in pertinent features

2.3 Dimensionality reduction To further explore the possibility of getting a concise representation of the dataset, and improve the model accuracy, PCA was done to transform the data and reduce the size of the feature space. This analysis revealed that with just two principal components, over 97% of the variance in the original dataset could be captured, which is a good representation of the original data and dramatically reduces the number of features to two in the PCA coordinate system. Code snippet 3 below shows the code and results obtained from the PCA analysis.

	Number_of_Modes	Fraction_of_Variance
1	1	0.640433
2	2	0.970394
3	3	0.991476
4	4	0.995204
5	5	0.9978
6	6	0.998981
7	7	0.999999
8	8	1.0

```julia
begin
    var_pca = [i^2 for i in F.S]
    frac_i = []
    n_modes = []
    for i in 1:length(var_pca)
        push!(frac_i, sum(var_pca[1:i])/sum(var_pca))
        push!(n_modes, i)
    end

    PCA = DataFrame(Number_of_Modes = n_modes, Fraction_of_Variance=frac_i)

end
```

Julia code snippets 3

Indeed, the PCA shows promise of reducing the feature space, as scatter plots corresponding to all PCA coordinate frames shows the potential of developing a classification tree to meet the goals of this project. As with the original feature space, and outlier analysis was done to detect outliers in the PCA coordinate frames. This post process would yield better accuracy in the model development.
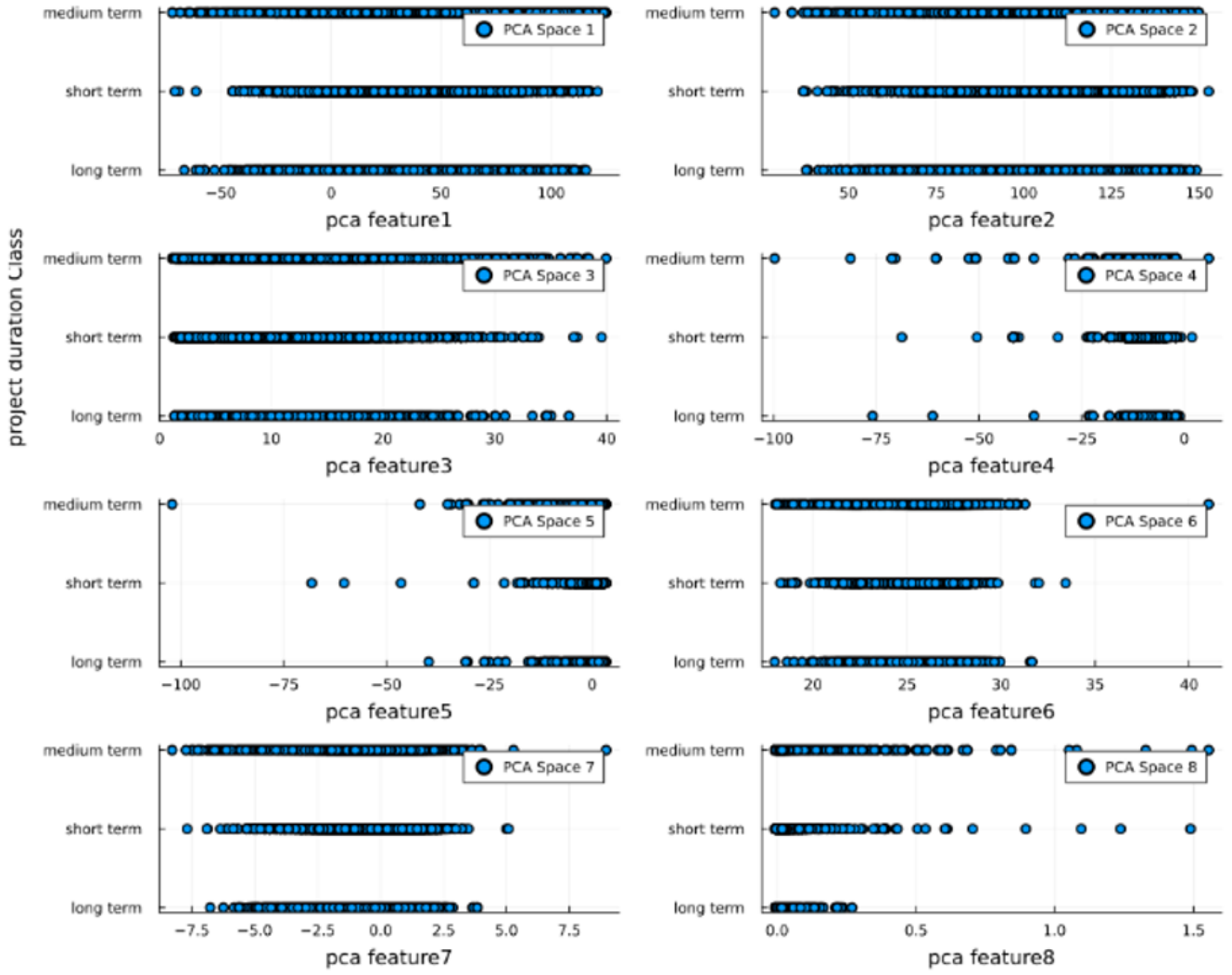
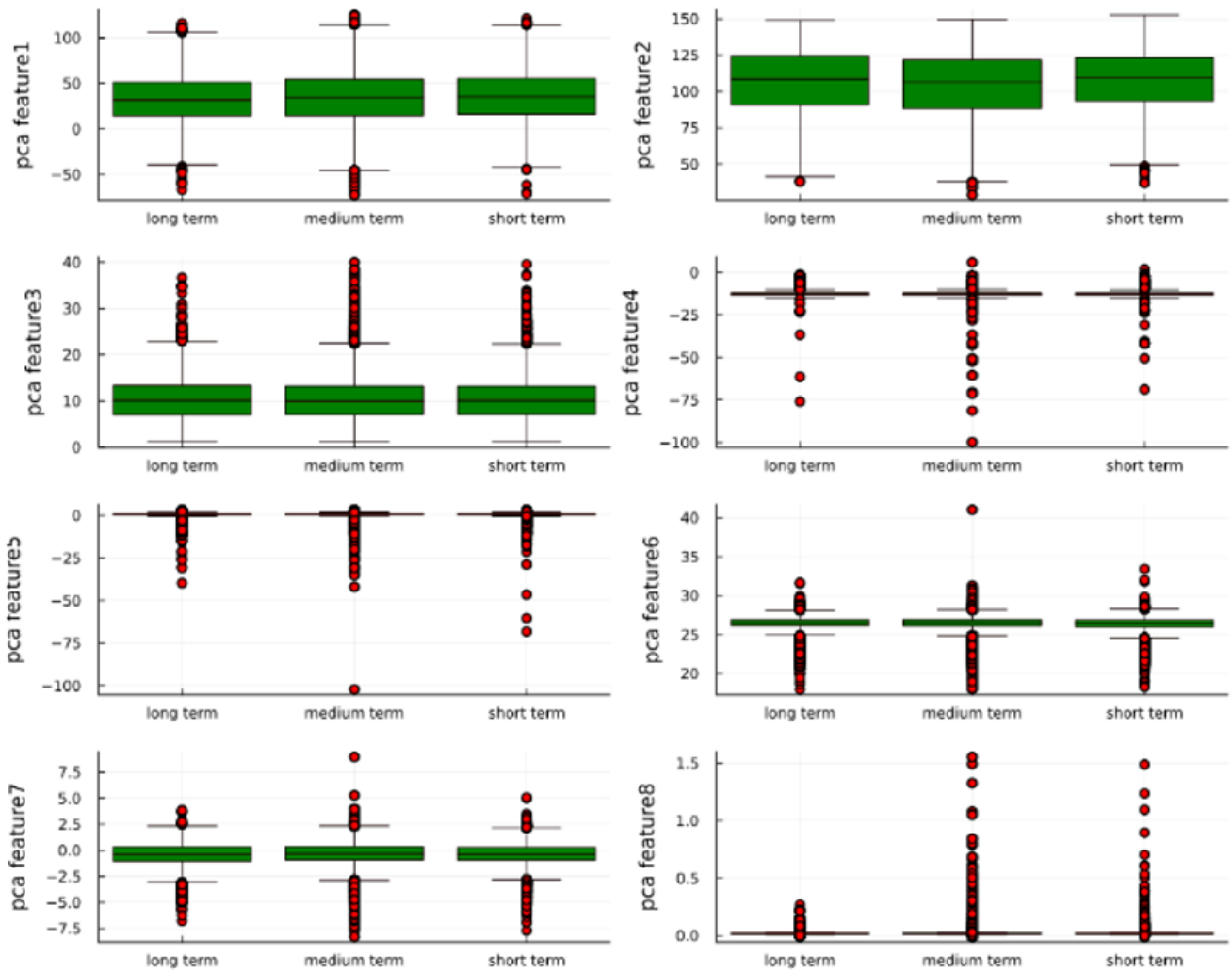*Fig. 7: Transformed data in PCA coordinate frame*

Fig. 8: Boxplots of transformed data in PCA coordinate frame

Although one could imagine that the first 2 principal components (that account for over 97% of the variance in the data) would be sufficient for the classification task, this is not the case. The plots in fig.7 reveal that the 4th, 5th, 6th, 7th, and 8th PCA coordinate frames also offer significant promise for the project goal. Additionally, the boxplots created for outlier detection will further guide in the decision tree creation. It can be seen from the plots in figures 7 and 8, that cuts can be made readily made to develop robust decision trees based on the Gini-impurity algorithm.

3.0. Predictive model development The predictive model development stage would leverage both the original feature space (with either Temperature of Wind chill dropped out) and the PCA feature space. Series of classification tree networks would be developed and validated with a confusion matrix plot. For the PCA-feature-space case, given a test data observation, it must be transformed to the PCA coordinate frame before evaluation on the model. As earlier stated, the model would be validated on four folds i.e.,4 sets of the original data would be created, and three would be used for training, while one (1) for testing, and this process would be repeated on all four sets of the data. This would potentially yield a robust and accurate model for the project task.

References:

1. Karimi Monsefi, Amin, Sobhan Moosavi, and Rajiv Ramnath. "Will there be a construction? Predicting road constructions based on heterogeneous spatiotemporal data.", *2022*

2. US Road Construction and Closures *2016 - 2021 from Kaggle*