From Fog Nets to Neural Networks

Chitesh Tewani, Keerthi Nuthi, Siddharth Jain

4/29/2016

Project deals with predicting the yield of collection in fog nets using the historical data about meteorological conditions and the 600 square meters fog net installations in south-west Morocco by DSH.

Contents

1. Title

From Fog Nets to Neural Nets

2. Team Members

1. Chitesh Tewani
   ctewani@iu.edu

2. Keerthi Nuthi
   keenuthi@umail.iu.edu

3. Siddharth Jain
   jain8@umail.iu.edu

3. Objectives and Significance

Dar Si Hmad (DSH), a non-profit association based in Agadir and Sidi Ifni, Morocco manages a network of fog nets that collect and disseminate fresh water to landlocked communities in Southwest Morocco. DSH has assembled years of data about weather patterns and water yield. The goal of this project is to come up with a clever way to predict how much water they can expect in the future or put in other way, predict the 'yield' (normalized amount of water that the system yielded as measured at a particular time) of a collection of fog nets in the Anti-Atlas-mountains in Southwestern Morocco using data from the following two sources:

- Weather data from a sensor co-located with the fog nets (microclimate data)
- Data from weather stations in three nearby cities in Morocco (macroclimate data)

Significance

Accurate predictions will enable DSH to operate more effectively and the communities it serves to be have greater access to fresh water throughout the year. Also this can facilitate alternate water resource planning and modelling, which is an essential part of allocating resources for the growth of a region. The fog water harvesting system is an example of integrated water resource management (IWRM) which considers water resources as integral

to the ecosystem as well as social and economic goods. Hence this project has the potential to make positive impact to ecosystem as well as social and economic goods.

Motivation

We were excited to work on this project, because:

• Its inherently interdisciplinary

• Gave us an opportunity to make a positive contribution to the sum of human achievement, as this project will direct help the communities in that region

• The project required us to apply most of the concepts covered in our syllabus and also beyond it and thus proved to be a good exposure to practical data mining tasks

4. Background

Important concepts

Fog harvesting, fog catching, and fog milking are all names for a long-established and proven scientific technique called fog collection. The origins, inspiration and motivation of this technique can be traced back to nature, wherein plants and insects developed strategies to provide themselves the water to survive in scanty or no rainfall regions of world; our ancient civilization, wherein instances of man-made dew catching devices have been found, like medieval “dew ponds” in southern England or volcanic stone covers on the fields of Lanzarote.

Fog is composed of extremely small water droplets in range of 1 µm to 40 µm. This technique uses mesh like structure which is hung between two poles to trap the water droplets in the fog. The wind pushes the fog, which has a very low fall velocity, through the vertical nets. Because water-droplets within fog are extremely small, they tend to move horizontally as the wind pushes them, which leads them to get trapped. They eventually condense and fall in a container placed at the base of the unit. This technique can be used in landlocked regions, where there is limited to access to water resources, but ample of fog. Though not all fog are equally productive, the amount of droplets in the fog found in mountains is substantially higher than the one found valley or coastal regions.

Experimental projects conducted in Chile indicate that it is possible to harvest between 5.3 l/m2/day and 13.4 l/m2/day depending on the location, season, and type of collection system used. At El Tofo, Chile, during the period between 1987 and 1990, an average fog harvest of 3.0 l/n2/day was obtained using 50 fog collectors made with Raschel mesh netting. A minimum fog season duration of

half a year might serve as a guideline when considering the feasibility of using this technology for water supply purposes; however, a detailed economic analysis to determine the minimum duration of the fog season that would make this technology cost-effective should be made. In general, fog harvesting has been found more efficient and more cost-effective in arid regions than other conventional systems.

Our project deals with predicting the yield of collection in fog nets using the historical data about meteorological conditions and the 600 square meters fog net installations in south-west Morocco by DSH. South-West Morocco is a water-poor, abundant fog region. Scarce water, compromised wells and climate change-induced droughts have destabilized traditional Amazigh communities and have created heavy burdens on marginalized women. Delivery of fog water in southwest Morocco significantly reduces women's laborious water-gathering chores, and help foster stable communities, continuation of ancestral languages and ways of living in thriving local environments. Specifically, water-gathering chores took up to 3.5 hours/day and often interrupted, or prevented, girls from regularly attending school. And, water availability allows poor farmers to keep their livestock which they previously might have sold during increasingly frequent droughts that lowered the water table, forcing livestock sales and driving farmers into cycles of poverty.

Following are the factors affecting yield of collection in fog nets, some of which we plan to study/exploit to design our prediction model:

Frequency of fog occurrence: A function of atmospheric pressure and circulation, oceanic water temperature, and the presence of thermal inversions.

Fog water content: A function of altitude, seasons and terrain features.

Design of fog water collection system: A function of wind velocity and direction, topographic conditions, and the materials used in the construction of the fog collector.

Global Wind Patterns: Persistent winds from one direction are ideal for fog collection.

Topography: It is necessary to have sufficient topographic relief to intercept the fogs/clouds; examples, on a continental scale, include the coastal mountains of Chile, Peru, and Ecuador, and, on a local scale, isolated hills or coastal dunes.

Altitude: The thickness of the stratocumulus clouds and the height of their bases will vary with location. A desirable working altitude is at two-thirds of the cloud thickness above the base. This portion of the cloud will normally have the highest liquid water content.

Distance from the coastline: There are many high-elevation continental locations with frequent fog cover resulting from either the transport of upwind clouds or the formation of orographic clouds. In these cases, the distance to the coastline is irrelevant. However, areas of high relief near the coastline are generally preferred sites for fog harvesting.

Previous Work

As per best of our knowledge, no previous work has been done directly on building prediction model for yield of fog collectors. But work has been done towards predicting fog; studying the relationship of fog water deposition with climatic conditions; studying the feasibility of fog water harvesting. As per our understanding at this point, we can use the conclusions and pointers from these studies or work to at least get some pointers to establish which dimensions are relevant for building yield prediction model and which of them already have an established relationship. The papers/publications we read for this are listed in reference section. Following is the aggregate take away from all of these papers/publications:

- The meteorological variables which had the greatest influence on prediction of fog deposition were wind speed, wind direction, and the dew-point depression (difference between air temperature and dew point).
- Dew-point depression was a strong indicator for the presence and absence of fog deposition according to all of the studies
- There is a linear relationship between wind speed and the quantity of fog deposition. Though according to DSH, wind speed is favourable up to certain speed and the extreme speeds can actually hurt the yield.
- Elevation of the location and no. of rainy days can also play a factor in the amount of fog

Our Work

Our work aims towards building a model for efficient prediction of yield of the fog nets for everyday for an evaluation period from the noted historical climatic statistics. Constructing a model that accurately predicts the yield variable, will enable DSH to operate effectively and helps it in serving more communities to have access to fresh water.

5. Methods

Data

The data for this project is obtained from the data sets provided by drivendata.com, which hosts data science competitions to save the world, bringing cutting- edge predictive models to organizations tackling the world's toughest problems.

We have microclimate data which is the weather data from a sensor co-located with the fog nets and macroclimate data which is the data from weather stations in three nearby cities in Morocco. Data preprocessing as stated in the next sections will be done on this data. Various meteorological measures are recorded at these different weather stations. We will be using these data to predict yield of a collection of fog nets.

File	Description
Test set: Microclimate (5 min intervals)	The test set data for the weather station near the fog nets. 5 minute intervals
Test set: Microclimate (2 hour intervals)	The test set data for the weather station near the fog nets. 2 hour intervals. (Same as prediction intervals).
Macroclimate: Guelmim Airport	Weather station data from the Guelmim Airport throughout the entire competition period.
Macroclimate: Sidi Ifni Weather Station	Weather station data from Sidi Ifni throughout the entire competition period.
Macroclimate: Agadir Airport	Weather station data from the Agadir Airport throughout the entire competition period.
Training set: Microclimate (2 hour intervals)	The training set data for the weather station near the fog nets. 2 hour intervals. (Same as prediction intervals).
Training set: Microclimate (5 minute intervals)	The training set data for the weather station near the fog nets. 2 minute intervals.
Target Variable: Water Yield	The target variable values during the training periods (2 hr intervals).

Microclimate weather data:

Mata is collected by a sensor array co-located with the fog nets. Measurements of meteorological variables such as temperature, humidity and wind are created at 5 minute intervals. Both the 5-minute time scale, and an aggregated 2-hour time scale are provided. Yield predictions are made for every two hour interval during the test periods.

Name	Description
percip_mm	Precipitation (mm)
humidity	A measure of the humidity in the air
temp	The temperature
leafwet450_min	Leaf wetness (a measure of the presence of dew) sensor 1
leafwet460_min	Leaf wetness (a measure of the presence of dew) sensor 2
leafwet_lwscnt	Leaf wetness (a measure of the presence of dew) sensor 3
gusts_ms	A measure of the highest gust during the reporting interval
wind_dir	The dominant direction the wind is blowing in
wind_ms	A measure of the current wind speed

Dates and times available:

Microclimate data is provided on two time scales. The sensor array natively records measurements every 5 minutes. For convenience, we are provided with both the native

(5 minute interval) data and resampled (every 2 hours) microclimate measurements. There are missing values at different intervals for the microclimate data, which needs to be interpolated.

Macroclimate data:

The macroclimate data consists of measures from one weather station on the coast of Morocco in Sidi Ifni and two airport weather stations in the cities of Agadir and Guelmim. These weather stations collect many meteorological measurements. The variables recorded differ between the Sidi Ifni weather station and the two airports (Agadir and Guelmim). A description of these variables is below:

Weather Station Variables (Sidi Ifni)

Name	Description
Nh	Amount of all the CL cloud present or, if no CL cloud is present, the amount of all the CM cloud present
Tx	Maximum air temperature (degrees Celsius) during the past period (not exceeding 12 hours)
DD	Mean wind direction (compass points) at a height of 10-12 metres above the earth's surface over the 10-minute period immediately preceding the observation
tR	The period of time during which the specified amount of precipitation was accumulated
Tn	Minimum air temperature (degrees Celsius) during the past period (not exceeding 12 hours)
ff10	Maximum gust value at a height of 10-12 metres above the earth's surface over the 10-minute period immediately preceding the observation (meters per second)
Tg	The minimum soil surface temperature at night. (degrees Celsius)
Td	Dewpoint temperature at a height of 2 metres above the earth's surface (degrees Celsius)
Date / Loc	Local time in this location. Summer time (Daylight Saving Time) is taken into consideration
Po	Atmospheric pressure at weather station level (millimeters of mercury)
E'	State of the ground with snow or measurable ice cover
Ff	Mean wind speed at a height of 10-12 metres above the earth's surface over the 10-minute period immediately preceding the observation (meters per second)
RRR	Amount of precipitation (millimeters)
E	State of the ground without snow or measurable ice cover
H	Height of the base of the lowest clouds (m)
ff3	Maximum gust value at a height of 10-12 metres above the earth's surface between the periods of observations (meters per second)
sss	Snow depth (cm)
N	Total cloud cover
P	Atmospheric pressure reduced to mean sea level (millimeters of mercury)
U	Relative humidity (%) at a height of 2 metres above the earth
T	Air temperature (degrees Celsius) at 2 metre height above the earth's surface
VV	Horizontal visibility (km)
WW	Present weather reported from a weather station
Ch	Clouds of the genera Cirrus, Cirrocumulus and Cirrostratus
Cm	Clouds of the genera Altocumulus, Altostratus and Nimbostratus
Cl	Clouds of the genera Stratocumulus, Stratus, Cumulus and Cumulonimbus
Pa	Pressure tendency: changes in atmospheric pressure over the last three hours (millimeters of mercury)
W2	Past weather (weather between the periods of observation) 2
W1	Past weather (weather between the periods of observation) 1

Airport Weather Variables (Agadir, Guelmim)

Name	Description
W'W'	Recent weather phenomena of operational significance
c	Total cloud cover
VV	Horizontal visibility (km)
DD	Mean wind direction (compass points) at a height of 10-12 metres above the earth's surface over the 10-minute period immediately preceding the observation
WW	Special present weather phenomena observed at or near the aerodrome
P	Atmospheric pressure reduced to mean sea level (millimeters of mercury)
ff10	Maximum gust value at a height of 10-12 metres above the earth's surface over the 10-minute period immediately preceding the observation (meters per second)
U	Relative humidity (%) at a height of 2 metres above the earth
T	Air temperature (degrees Celsius) at 2 metre height above the earth's surface
Ff	Mean wind speed at a height of 10-12 metres above the earth's surface over the 10-minute period immediately preceding the observation (meters per second)
Td	Dewpoint temperature at a height of 2 metres above the earth's surface (degrees Celsius)
Date/Loca	Local time in this location. Summer time (Daylight Saving Time) is taken into consideration
Po	Atmospheric pressure at weather station level (millimeters of mercury)

Dates and times available:

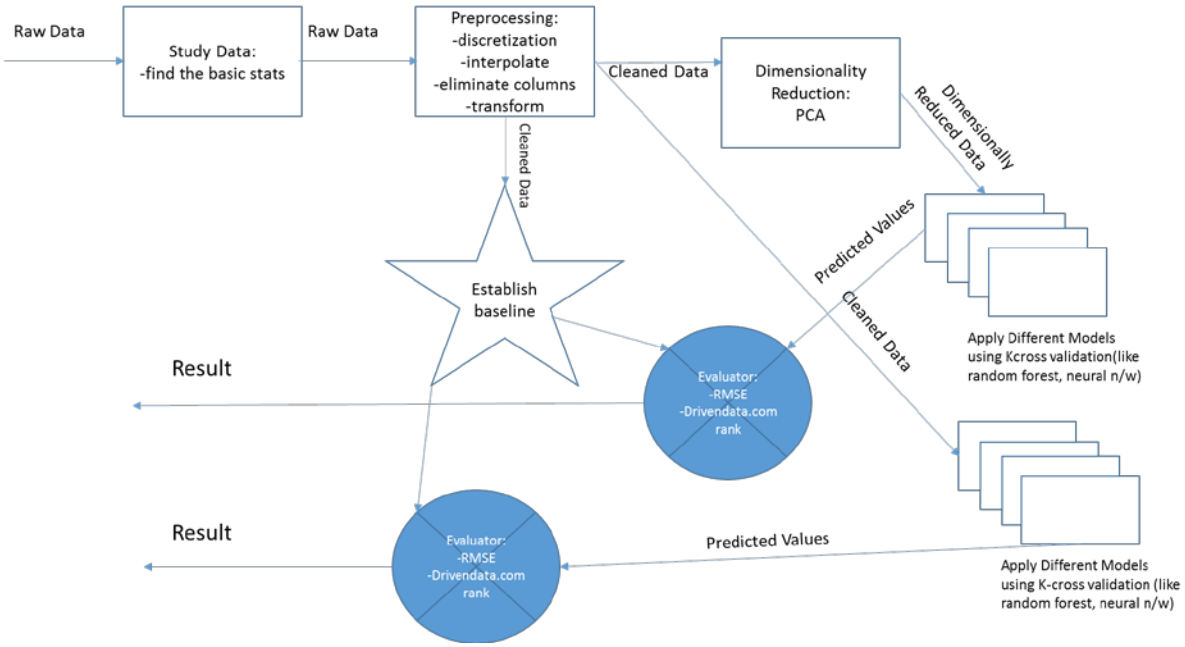
For the macroclimate data, historical measurements are provided for the length of the period of interest. The interval depends on the weather station and there are some brief missing time periods in the dataset.

Training set: Microclimate example

Attribute Name	Value
Percip_mm	0.81666666
Humidity	0.44363643
Temp	7.99583346
Leafwet450_min	0.33333333
Leafwet460_min	0
Leafwet_lwscnt	443.083333
Gusts_ms	5.34367144
Wind_dir	116.833333
Wind_ms	4.70662199

Methodology

Steps:



Study Data

- Studied the data and generated the basic stats of each data, consisting of mean, median, Std. dev, variance, min, max, max-min, rows having values, rows missing values, % values missing, 25 percentile (q1), 50 percentile (q2) and 75 percentile (q3)
- The basic stats data and the observations about it is available under appendix

Pre-processing/Preliminary Data Cleaning/Transformation

- All the temperatures were converted into K
- Discretization of DD,VV column in all the data
- Transformed complex attribute C/D with broken cloud data (values like "Few clouds (10-30%) 3300 m, broken clouds (60-90%) 480 m") into atomic attribute so as to enable comparison. There were many patterns and we build a regular expression to achieve this
- Eliminated columns WW, WW' because of extremely low % of data available, 4.4% and 0.15% in some data
- Estimated missing values for 2 hour training data by collating the 5 minute training data and averaging them

Establishment of base line

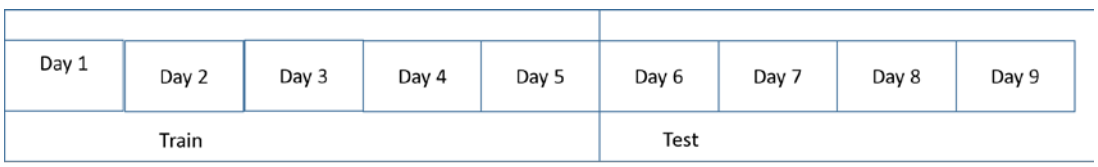
- Created the baseline by using linear regression and noted down the RMSE and drivendata.com score for future comparison

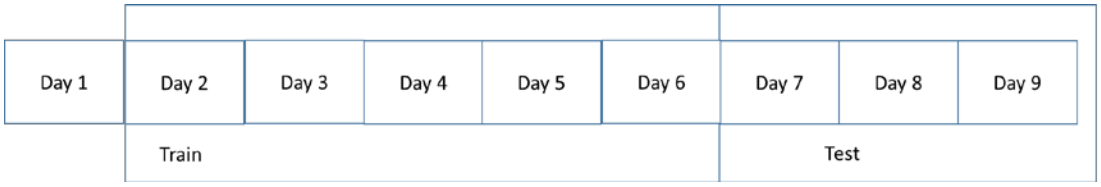
Dimensionality Reduction

- Used PCA to determine the most important features

Apply different Models and Evaluate

- Apply Multivariate Adaptive Regression Spline (MARS) model, Random Forest Model, Neural Network with Linear Regression
- Also we use a variant of K-cross validation for our time-series data, wherein we only use the data from past to predict the values in future with a windowing system. This variant of ours can be visualized for 3-cross validation in the below 9 day data:





Note from the above example, it might look the test data is drastically reduced, but if you apply on large data it works.

Evaluation Strategy

Accuracy

Accuracy of each of our model will be based on Root-mean square error

$$\sqrt{\frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y})^2}$$

1. Root-mean-square error is the square root of the sum of the squared differences between the predicted values and the actual values.
2. The goal is to minimize Root Mean Squared Error.

The predicted *yield*, target variable, which would be in float, and differences with the expected value is in float.

Along, with root-mean squared error, we can also evaluate on mean absolute error and mean squared error, but as the dataset challenge uses root-mean squared error, we plan to concentrate our results on this measure.

Apart from this, we also predict the values for given drivendata.com competition template (Note: the yield/result for the data is not provided to us, thus is a fair indicator of our model's performance) and use its score and ranking for self-evaluation.

Baseline

We use naïve linear regression model to compare all the accuracy of our refined models . The comparison is done on the basis of the RMSE of this model and drivendata.com score for the corresponding submission

6. Results

We started first with applying different models on only micro-climate data,

Baseline Establishment

Without PCA

Baseline RMSE= 2.695621

DrivenData.com score (DDS)= 4.1830

With PCA

Baseline RMSE= 2.69841

DrivenData.com score (DDS)= 4.0562

- The baseline RMSE with PCA is not better in comparison to the without PCA version, but is very close. Though we do observe with almost half the no. of columns we are approximately achieving the same results and this can immensely help us in handling cases of curse of dimensionality
- Our DDS improved with PCA

K-Fold Cross-Validation With PCA

The data for Micro-climate is spread over 484 days and performing 5-fold Cross validation technique for given time series, we achieved the following accuracy

Kth Fold	RMSE
1	2.021805
2	2.001484
3	2.107595
4	2.378778
5	1.307777

Average	1.96

DrivenData.com score (DDS) for RMSE 2.001484 = 4.1226

DrivenData.com score (DDS) for RMSE 1.307777 = 4.0637

- Though the last RMSE of 1.307777 is very low compared to other RMSE, it performed poorly on DDS, because the amount of records tested were low and so the RMSE was misleading.

Prediction With Other Models for Micro-climate data

Model	**RMSE w/ PCA**	**RMSE without PCA**	**Δ(RMSE$_{baseline}$ -RMSE) with PCA**	**Δ(RMSE$_{baseline}$ -RMSE) without PCA**	**DDS without PCA**	**DDS with PCA**
Multivariate Adaptive Regression Splines (MARS) Model	NA	1.844474	NA	0.85	5.7910	NA
Random Forest	1.194128	0.7733132	1.501	1.9923078	3.8816	3.9246
Neural Net	NA	2.280573	0.41	NA	NA	NA

- We used Multivariate Adaptive Regression Splines (MARS) from Salford System for prediction, as we thought the constraint of not creating a single linear regression line could help us improve our model. Though, MARS is usually used for data with high dimensions, we wanted to test the model and get important features/attributes.
- We used random forest with 2000 ensemble trees and it gave the best performance of all the other models
- Random forest without PCA version performs better than the PCA version
- We tried to run multi-layer neural net with Linear Regression as the activation function, unlike logistic regression as used in classification models. The threshold experimented with 0.01 and 0.02, step size of 1e5 & 1e6, but it did not converge for multi-layer network with 3 to 5 hidden nodes.

- We managed to build a single layer Neural Network using with three hidden nodes, which converged and gave decent results

Prediction For other model using Macroclimate and Combined Data

- After pre-processing steps on Macroclimate data, we developed Linear Regression and Random Forrest models.
- As the dimensionality of Macro-climate data was twice that was Micro-climate, we also performed Principal Component Analysis (PCA) for dimensionality reduction.
- The following are 10 Correlated pairs for Agadir and Guelmim Macro-climate data:
    - Guelmim Macroclimate Data:

      (P0,P), (U,Td), (T,DD), (DD,Td), (Overcast,cumulonimbus.clouds), (Few.clouds,Broken.clouds), (T,VV), (P0,VV), (Few.clouds,cumulonimbus.clouds), (Few.clouds,Scattered.clouds)

    - Agadir Macroclimate Data:

      (P0,P), (U,Td), (DD,T), (Ff,T), (DD,Td), (Td,T), (DD,Ff), (Few.clouds,Scattered.clouds), (Few.clouds,cumulonimbus.clouds), (Broken.clouds,cumulonimbus.clouds)

- Principal Component Analysis on Macroclimate Data reduced the dimensionality for both, Agadir and Guelmim. Applying linear regression models for both kind of data (with and without PCA), we were able to get better results for the prediction ones using PCA.
- Next, we built Linear Regression and Random Forrest models using the best discovered attributes for Microclimate and Macroclimate data, and predicted the yield.

The following are the results of our experiments, which has -

Train Data	**Model**	**RMSE** (with Train as Test)	**RMSE** (on drivendata.com)
Guelmim Macroclimate	Linear Regression	3.063	4.2537
Guelmim Macroclimate	Linear Regression + PCA	2.85	4.2470
Agadir Macroclimate	Linear Regression	3.72	4.1225
Agadir Macroclimate	Linear Regression + PCA	3.53	4.1349
Macroclimate (Agadir + Guelmim) + Microclimate	Linear Regression + PCA (individual)	2.216	4.2518
Macroclimate (Agadir + Guelmim) + Microclimate	Random Forrest (2000 trees) + PCA (individual)	0.622	4.3033

We could observe the following from the results -

- The linear regression models with PCA for both Macroclimate data performed better than their counterparts when Train data is used as test. The reason for this could be, as the entire training data is used, i.e., without cross validation for the time series, the model tends to over fit the data and gives much better results
- However, the same linear regression model with PCA do not perform as well on submission to drivendata.com, as most of the test set data is not in the training set
- Similar, inference can be drawn for Random Forrest with PCA for the MERGED data (Macroclimate [Agadir + Guelmim] + Microclimate), as it nearly fits the entire data resulting in lowest of RMSE
- Combining the data and using only the important features did **not** give a good result. Perhaps, using Neural Nets we would able to obtain better results

DrivenData.com Rank and Results

We are in the top 10% of the competitors on drivendata.com, as of 4/29/2015

7. Conclusions

We used following three types of data to predict the yield:

1. Micro-data only
2. Macro-data only
3. Micro-data and macro-data combined

Micro-data only gave the best results, as it's the data of the area which is nearest to the yield area. Next best yield prediction results were obtained with macro-data. We were expecting the combination of micro-data and macro-data to give the best predictions, because it essentially captures the whole picture, though it did not.

We think the reason for that is, there is a complex relationship between the two data and simple PCA analysis that we conducted was not enough to clear the noise, it requires further understanding of the domain and creation of complex transformations which captures this relationship.

In the future, we would like study more on Neural Network and Feature engineering to help us implement the inter-relation of complex attributes. As lot of research is done in the domain of time series and models are built differently for them, we would like to explore those areas as mentioned in the proposal.

We are satisfied with our ranking and performance on the leader-board of the competition, and encourage to take part in such competitions.

8. Individual Task

(a.)

The project was interesting in sense of understanding the attributes and the way different models can be built around it. As this was the first time to venture in domain of Linear Regression and Predictive Modelling, not just classification and rules. Understanding and researching on the domain of attributes was a one-of-a-kind experience. However, personally I would like to understand more the concept Neural Network and Feature Engineering.

(b)

Chitesh:

Data Preprocessing:

Transforming complex attributes to atomic attributes.

Cleaning Macroclimate data – Sidi Infi weather station

Random Forest implementation

With PCA

Without PCA

Neural Network implementation

Identification of correlation between attributes - Guelmim data and Agadir data combined with 'yield' attribute.

Keerthi:

Data Preprocessing:

Estimating missing values

Cleaning Macroclimate data – Guelmim weather station

Baseline model implementation

With PCA

Without PCA

Identification of correlation between attributes - Guelmim data

Siddharth:

Data Preprocessing:

Eliminating attributes

Cleaning Macroclimate data – Agadir weather station

Baseline model implementation:

K-fold cross validation with PCA

MARS model implementation

Identification of correlation between attributes - Agadir data

I am happy with the performance of my project partners.

(c).

This is a greenfield project.

The project undertaken is a part of on-going competition hosted by drivendata.com. We are competing for the grand prize of $15,000.

9. References

[1] - Hiatt, Cyrus, Daniel Fernandez, and Christopher Potter. *"Measurements of Fog Water Deposition on the California Central Coast."* Atmospheric and Climate Sciences (2012): 525-31. Link

[2] - J. Oliver, *"Fog Water Harvesting along the West Coast of South Africa: A Feasibility Study,"* Water SA,

Vol. 28, No. 34, 2002, pp. 349-360. Link

[3] - R. E. Newell, *"Comments on 'The Forecasting of Winter Fog: A Geographical Approach' "*, Journal of Applied Meteorology, Vol. 3, No. 3, 1964, pp. 342-343. Link

[4] - G. N. Pradhan and B. Prabhakaran, *"Association rule mining in multiple, multidimensional time series medical data"* Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on, New York, NY, 2009, pp. 1720-1723. Link

[5] - Liang-xi Qin and Zhong-zhi Shi, *"Efficiently mining association rules from time series Efficiently Mining Association Rules from Time Series"*. Link

[6] - How to get fresh water out of thin air Link [7] - Fog Collection Link

[8] - Fog Harvesting in Rural Southwest Morocco Link

[9] - Pang-Ning Tan, Michael Steinbach, Vipin Kumar, *Introduction to Data Mining, Pearson Addison Wesley*

[10] - DrivenData Challenge - Fog Net to Neural Net Link

[11] - DrivenData Challenge - Fog Net to Neural Net Data Link [12] - Fog Harvesting Link

[13] - Dimensionality Reduction Link

[14] - Fog Harvesting Link

[15] - DrivenData Challenge - Fog Net to Neural Net Problem Description Link

[16] - Clustering Bagging Link

[17] - MARS Link

[18] - ResearchGate - k-cross validation technique for Time series data

https://www.researchgate.net/post/How_to_apply_the_k-cross_validation_technique_when_the_data_is_in_the_form_time_series

[19] - R-bloggers – Fitting a Neural Network on R http://www.r-bloggers.com/fitting-a-neural-network-in-r-neuralnet-package/

[20] - R-bloggers – Computing and Visualizing PCA in R http://www.r-bloggers.com/computing-and-visualizing-pca-in-r/

Appendix

Basic Data Stats and Notes

Basic Data Stats

Microclimate training data (5min)

	percip_mm	humidity	temp	leafwet450_min	leafwet460_min	leafwet_lwscnt	gusts_ms	wind_dir	wind_ms
mean	0.07895	0.5551	15.56	0.9935	0.8014	457.5	3.382	135.2	2.827
median	0	0.4971	14.4	0	0	442	3.118	87	2.615
Std. dev	1.221	0.286	7.148	1.985	1.824	50.67	1.943	120.8	1.727
variance	1.491	0.08181	51.1	3.94	3.327	2568	3.777	14595	2.984
min	0	0	0	0	0	0	0	0	0
max	69.80000305	1.194626689	37	9	5	1023	13.8828	359	12.072
max-min	69.80000305	1.194626689	37	9	5	1023	13.8828	359	12.072
rows having values	138771	138770	138770	138771	110855	138771	139017	139017	139017
rows missing values	486	487	487	486	28402	486	240	240	240
% values missing	0.350217264	0.350940405	0.350940405	0.350217264	25.62085607	0.350217264	0.172640756	0.172640756	0.172640756
25 percentile (q1)	0	0.3	9.9	0	0	439	1.9	38	1.5
50 percentile (q2)	0	0.4971	14.4	0	0	442	3.118	87	2.615
75 percentile (q3)	0	0.8	20.9	0	0	447	4.5	281	3.9

Microclimate data Agadir Airport

	T	P0	P	U	Ff	ff10	Td
mean	19	755.9	762.5	67.1	3.1	12.1	11.7
median	19	762	755.3	72	3	12	13
Std. dev	6	3.77	3.85	20.96	1.97	3.1	5.13
variance	36	14.2	14.8	439.5	3.9	9.6	26.3
min	3	737.5	744	3	0	6	-16
max	46	769.4	776.2	100	14	21	27
max-min	43	31.9	32.2	97	14	15	43
rows having values	26403	26123	26285	26401	26270	123	26401
rows missing values	168	448	286	170	301	26448	170
% values missing	0.632268262	1.6860487	1.076361447	0.639795266	1.13281397	99.53708931	0.639795266
25 percentile (q1)	15	753.2	759.7	53	2	11	8
50 percentile (q2)	19	762	755.3	72	3	12	13
75 percentile (q3)	23	758.2	765	83	4	15	16

Microclimate 2 hour training data

	percip_mm	humidity	temp	leafwet450_min	leafwet460_min	leafwet_lwscnt	gusts_ms	wind_dir	wind_ms
count	5781	5781	5781	5781	4617	5781	5794	5794	5794
mean	0.078972	0.554852	15.566805	0.991033	0.799972	457.476362	3.381701	135.184925	2.827167
std	0.97397	0.282715	7.126274	1.903983	1.743959	48.172783	1.832613	96.18655	1.63749
min	0	0	0	0	0	297.625	0	0	0
25 percentile (q1)	0	0.319978	9.9375	0	0	438.583333	2.007544	60.760417	1.588485
50 percentile (q2)	0	0.496859	14.470833	0	0	441.625	3.143336	102.041667	2.619792
75 percentile (q3)	0	0.827473	20.9375	0	0	447	4.538977	209.65625	3.867351
max	24.25	1.072792	36.508334	5.173913	5	1023	11.5187	355	10.092204

Guelmim Airport Data

	T	P0	P	U	Ff	ff10	Td
count	9736	9714	9721	9731	8664	320	9731
mean	21.278554	735.972607	762.144718	53.394923	4.655471	15.1875	9.023944
std	6.841898	3.536293	3.896215	25.677071	2.677403	2.908613	7.052551
min	4	718.8	744.7	2	0	9	-22
25 percentile (q1)	17	733.5	759.7	33	2	13	5
50 percentile (q2)	21	735.1	761.2	53	5	15	11
75 percentile (q3)	26	738	764.3	73	7	17	14
max	46	749	776.2	100	21	31	25

Notes

- DD Values may be discretized, following are the possible values:
  - Null
  - Calm, no wind
  - variable wind direction
  - Wind blowing from the east
  - Wind blowing from the east-northeast
  - Wind blowing from the east-southeast
  - Wind blowing from the north
  - Wind blowing from the north-east
  - Wind blowing from the north-northeast
  - Wind blowing from the north-northwest
  - Wind blowing from the north-west
  - Wind blowing from the south
  - Wind blowing from the south-east
  - Wind blowing from the south-southeast
  - Wind blowing from the south-southwest
  - Wind blowing from the south-west
  - Wind blowing from the west
  - Wind blowing from the west-northwest
  - Wind blowing from the west-southwest
- VV
  - Few null values
  - Values range from 0 to “10.0 and more” (probably can be split into ranges and discretized)

        - Close to 80% of the values are “10.0 and more”
- W”W”
    - Only 0.15% of the rows have a value
    - Values:
        - Null
        - Drizzle
        - Fog
        - Mist
        - Rain
        - Shower(s), rain
        - Thunderstorm
        - Thunderstorm, hail
        - Thunderstorm, rain
        - Thunderstorm, shower(s)
- WW
    - Only 4.4% rows have any value
    - Values:
        - Null
        - Drizzle
        - Duststorm
        - Fog
        - Heavy rain
        - In the vicinity shower(s)
        - Light drizzle
        - Light rain
        - Light shower(s), rain
        - Light thunderstorm, rain
        - Mist
        - Mist, rain
        - Rain
        - Rain, mist
        - Shower(s), rain
        - Thunderstorm
        - Thunderstorm, hail, rain
        - Thunderstorm, light rain
        - Thunderstorm, light shower(s), rain
        - Thunderstorm, rain
        - Thunderstorm, shower(s), rain
    - C

- 25758 rows have values, i.e. 96.94% have values
- Will have to be split in like 3 columns or more, to be of any use
- Has 506 different values
- Example value “Scattered clouds (40-50%) 300 m, scattered clouds (40-50%) 480 m”
  => Can be split into something like: height-cloud coverage, height-cloud coverage