SUNSHINE PREDICTION BY MACHINE LEARNING METHODS

Final Project Report

ISM6136.901F22.90231 - Data Mining

Submitted by

Manoj Kumar Goud Sowdari	U54666573
Sai Swetha Somi	U04880400
Prathyusha Hari	U43145266

Major in

BUSINESS ANALYTICS AND INFORMATION SYSTEM

Under the guidance of

Professor

BALAJI PADMANABHAN

MUMA COLLEGE OF BUSINESS

UNIVERSITY OF SOUTH FLORIDA

Table of Contents

1. INTRODUCTION TO PHOTOVOLTAIC TECHNOLOGY

A photovoltaic system also known as photovoltaic power system, solar PV system or casually solar array is a power system designed to supply usable solar power by means of photovoltaics. PV is a method of converting solar energy into direct current electricity using semiconducting materials that exhibit the photovoltaic effect. A photovoltaic system employs solar panels composed of several solar cells to supply usable solar power. Power generation from solar PV has long been seen as a clean sustainable energy technology which draws upon the planets most plentiful and widely distributed renewable energy source which is the sun.

Photovoltaic cells are made-up of at least two semiconductor layers. One layer containing a positive charge, the other a negative charge. Sunlight consists of little particles of solar energy called photons. As a PV cell is exposed to the sunlight, many of the photons are reflected, transmitted, or observed by the solar cell. When enough photons are absorbed by the negative layer of the photovoltaic cell, electrons are freed from the negative semiconductor material. the positive layer is manufactured in such a way that, these free electrons naturally migrate to the positive layer creating a voltage differential. When the two layers are connected to an external load the electrons flow through the circuit creating electricity. Each individual solar energy cell produces only one to two watts of electricity. To increase power output, cells are combined in a weather tight package called a solar module. These modules vary from one to several thousand are then wired up in serial or parallel connection in a solar array to create the desired voltage.

Photovoltaic powered lighting systems are reliable and low-cost alternative. These are widely used in commercial lightning purposes such as security, billboard sign, outdoor lighting etc. In consumer electronics, solar powered watches, calculators, and cameras are everyday applications of PV technologies. Photovoltaics can also be used in telecommunications. A residence located more than a mile from the electric grid can install a PV system more inexpensively than extending the electric grid. Over half a million homes worldwide use PV power as their only source of electricity. For water pumping, PV powered pumping systems are excellent, simple, reliable and has a life span of 20 years. Extensive research is being conducted for developing thin film solar cells using various materials and technologies. These are one of the best candidates for solar windows which can be installed on buildings, vehicle roofs etc., for solar generation and glare reduction.

The key to the successful solar energy installation is to use quality components that have long lifetimes and require minimum minimal maintenance. PV technology fills a significant need in supplying electricity creating local jobs and promoting economic development in rural areas, avoiding the external environmental costs associated with traditional electrical generation technologies. Major power policy reforms and tax incentives will play a major role in encouraging common people to opt for solar energy as an alternative sustainable energy.

2. PURPOSE OF SUNSHINE PREDICTION

Economically, PV system requires a high initial capital investment but in the long run, running costs are very low. The initial capital investment cost involves: the purchase of a PV system, balance of system, transportation and installation costs, custom design, and engineering. From a business point of view, it is important to estimate the time it will take to reach a break even in the savings of the electricity bill vs the investment & maintenance of a solar cell.

General Electricity Cost:

- The average residential electricity rate in Tampa, FL = 0.14 $/kWh.
- Average household electricity consumption = 914 KWh / month
- Monthly Electricity cost = 128 $ / month

PV electricity Cost:

- PV model assume t hours peak-sun equivalent per day = 30t hours / month
- LG solar panel rating = 385 W = 0.385 KW
- Electricity cost = 0.385 x 30t = 11.5t KWh / month
- Electricity saved per month = 11.5t x 0.14 = 1.6t $
- For 10 solar panels = 16t$ / month

- LG solar panel cost for 10 panels = 4800$
- Breakeven (no. of years) Z = $\frac{4800}{(16t-128)\times 30}$ => $\boxed{Z = \frac{10}{t-8}}$……………………..2.1
- Average lifespan of solar cells (L)= 25 years.
- For any profit margin,

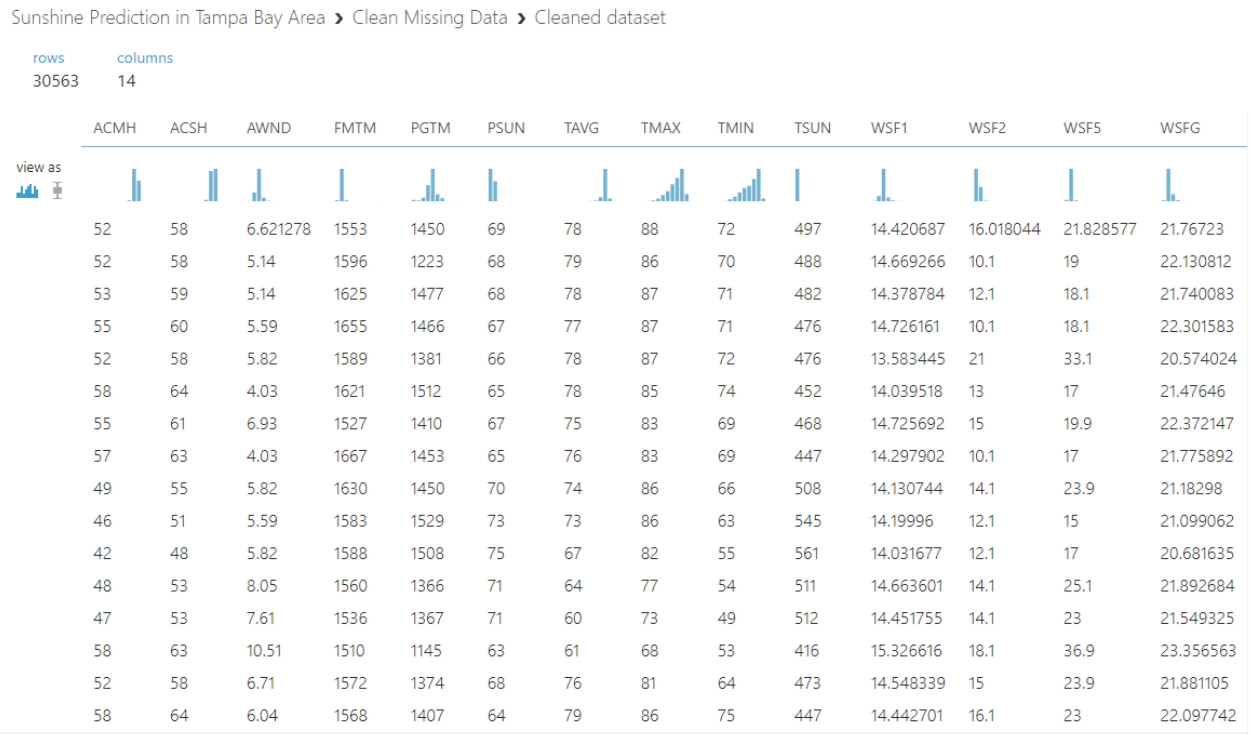
  Z < (0.6) L => Z < 15 => $\frac{10}{t-8} < 15$ => t ~ 9 hrs
- From the above calculations, the average daily sunshine hours in a month should be approximately 9 hours.
- In any geographical area, the amount of sunshine received (%) which is relative to the sunshine received without any clouds in the visibility (100%).
- Hence the challenge is to predict the percentage of average sunshine available per day from the data available by the observation of average cloudiness/day, wind speeds, air temperature, etc.
- We use machine learning techniques to build a model which gives accurate prediction of daily possible percentage of sunshine.

3. EXPLORATORY DATA ANALYSIS

A real time data set is taken from the daily summaries published by the NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION. In these daily summaries some of the important attributes are average cloudiness, minimum and maximum temperatures, average temperatures, gust time, wind speed, percentage of daily sunshine possible and that total sunshine in minutes.

Machine Learning platform Azure ML is chosen for the analysis. In the first step of exploratory data analysis 14 important attributes are chosen for prediction. Among these, the dependent variable is total daily sunshine in minutes. Upon visualization of the selected columns, it is found that there are several missing values in every attribute.

Using the probabilistic principal component analysis cleaning mode, we substitute the missing values estimated by the PCA. PCA is chosen because it is a natural approach to the estimation of the principal access in cases where data vectors exhibit one or more missing values in all the attributes. A snapshot of the clean data is shown in the figure 3.1. The total number of rows are 30563.



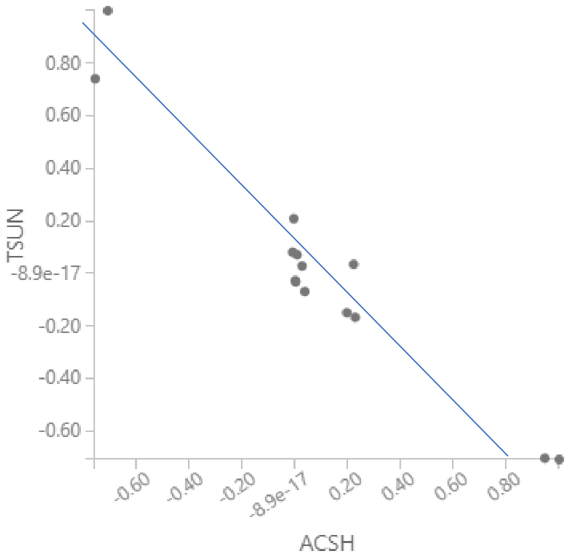
Sunshine Prediction in Tampa Bay Area › Clean Missing Data › Cleaned dataset

rows 30563

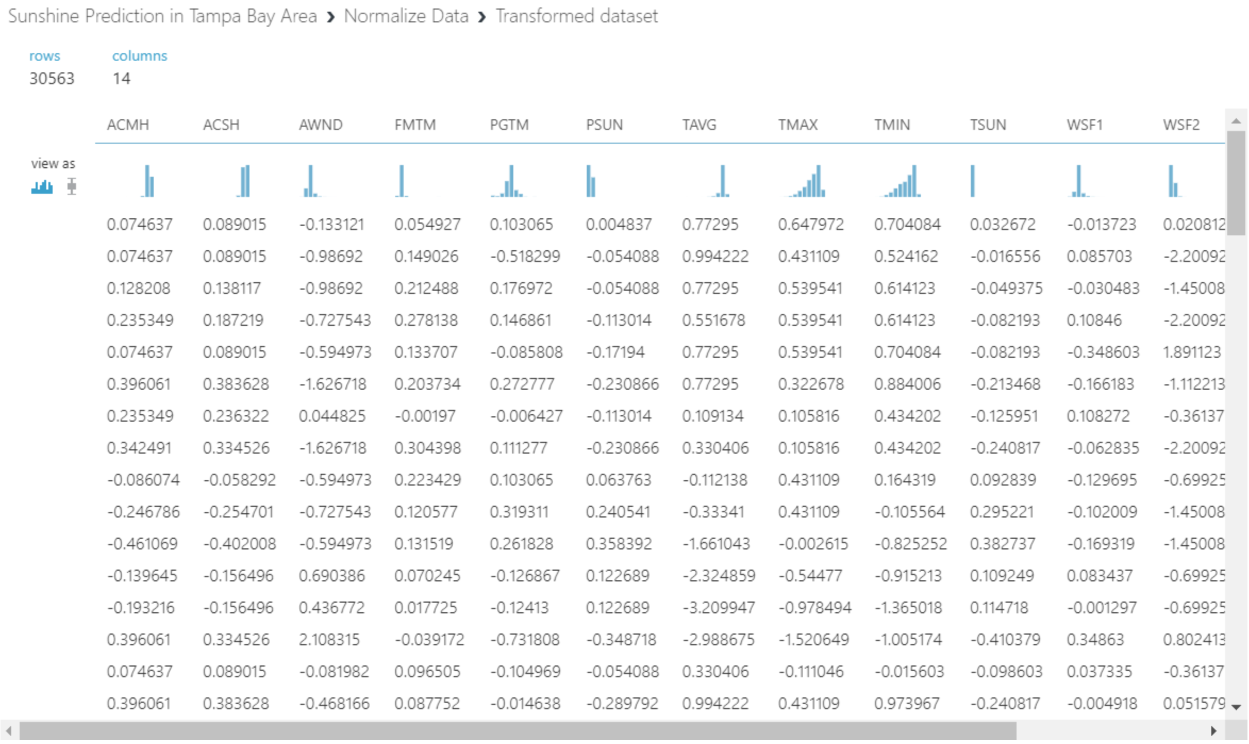
columns 14

ACMH	ACSH	AWND	FMTM	PGTM	PSUN	TAVG	TMAX	TMIN	TSUN	WSF1	WSF2	WSF5	WSFG
52	58	6.621278	1553	1450	69	78	88	72	497	14.420687	16.018044	21.828577	21.76723
52	58	5.14	1596	1223	68	79	86	70	488	14.669266	10.1	19	22.130812
53	59	5.14	1625	1477	68	78	87	71	482	14.378784	12.1	18.1	21.740083
55	60	5.59	1655	1466	67	77	87	71	476	14.726161	10.1	18.1	22.301583
52	58	5.82	1589	1381	66	78	87	72	476	13.583445	21	33.1	20.574024
58	64	4.03	1621	1512	65	78	85	74	452	14.039518	13	17	21.47646
55	61	6.93	1527	1410	67	75	83	69	468	14.725692	15	19.9	22.372147
57	63	4.03	1667	1453	65	76	83	69	447	14.297902	10.1	17	21.775892
49	55	5.82	1630	1450	70	74	86	66	508	14.130744	14.1	23.9	21.18298
46	51	5.59	1583	1529	73	73	86	63	545	14.19996	12.1	15	21.099062
42	48	5.82	1588	1508	75	67	82	55	561	14.031677	12.1	17	20.681635
48	53	8.05	1560	1366	71	64	77	54	511	14.663601	14.1	25.1	21.892684
47	53	7.61	1536	1367	71	60	73	49	512	14.451755	14.1	23	21.549325
58	63	10.51	1510	1145	63	61	68	53	416	15.326616	18.1	36.9	23.356563
52	58	6.71	1572	1374	68	76	81	64	473	14.548339	15	23.9	21.881105
58	64	6.04	1568	1407	64	79	86	75	447	14.442701	16.1	23	22.097742

Upon finding that correlations between the average cloudiness and average sunshine time, the scatter plot shows that there are small deviations from the linear plot. This is due to the influence of other attributes on the percentage of sunshine received. The scatter plot is shown in the figure 3.2.



Since different attributes have different fundamental dimensions, we choose to normalize the data by using z-score transformation method, to keep the values in between unity. This approach is suitable when the data set has continuous variables.


Sunshine Prediction in Tampa Bay Area › Normalize Data › Transformed dataset

rows 30563
columns 14

view as

ACMH	ACSH	AWND	FMTM	PGTM	PSUN	TAVG	TMAX	TMIN	TSUN	WSF1	WSF2
0.074637	0.089015	-0.133121	0.054927	0.103065	0.004837	0.77295	0.647972	0.704084	0.032672	-0.013723	0.020812
0.074637	0.089015	-0.98692	0.149026	-0.518299	-0.054088	0.994222	0.431109	0.524162	-0.016556	0.085703	-2.20092
0.128208	0.138117	-0.98692	0.212488	0.176972	-0.054088	0.77295	0.539541	0.614123	-0.049375	-0.030483	-1.45008
0.235349	0.187219	-0.727543	0.278138	0.146861	-0.113014	0.551678	0.539541	0.614123	-0.082193	0.10846	-2.20092
0.074637	0.089015	-0.594973	0.133707	-0.085808	-0.17194	0.77295	0.539541	0.704084	-0.082193	-0.348603	1.891123
0.396061	0.383628	-1.626718	0.203734	0.272777	-0.230866	0.77295	0.322678	0.884006	-0.213468	-0.166183	-1.112213
0.235349	0.236322	0.044825	-0.00197	-0.006427	-0.113014	0.109134	0.105816	0.434202	-0.125951	0.108272	-0.36137
0.342491	0.334526	-1.626718	0.304398	0.111277	-0.230866	0.330406	0.105816	0.434202	-0.240817	-0.062835	-2.20092
-0.086074	-0.058292	-0.594973	0.223429	0.103065	0.063763	-0.112138	0.431109	0.164319	0.092839	-0.129695	-0.69925
-0.246786	-0.254701	-0.727543	0.120577	0.319311	0.240541	-0.33341	0.431109	-0.105564	0.295221	-0.102009	-1.45008
-0.461069	-0.402008	-0.594973	0.131519	0.261828	0.358392	-1.661043	-0.002615	-0.825252	0.382737	-0.169319	-1.45008
-0.139645	-0.156496	0.690386	0.070245	-0.126867	0.122689	-2.324859	-0.54477	-0.915213	0.109249	0.083437	-0.69925
-0.193216	-0.156496	0.436772	0.017725	-0.12413	0.122689	-3.209947	-0.978494	-1.365018	0.114718	-0.001297	-0.69925
0.396061	0.334526	2.108315	-0.039172	-0.731808	-0.348718	-2.988675	-1.520649	-1.005174	-0.410379	0.34863	0.802413
0.074637	0.089015	-0.081982	0.096505	-0.104969	-0.054088	0.330406	-0.111046	-0.015603	-0.098603	0.037335	-0.36137
0.396061	0.383628	-0.468166	0.087752	-0.014638	-0.289792	0.994222	0.431109	0.973967	-0.240817	-0.004918	0.051579

4. DATA SET

Data Set: Tampa weather daily summaries data set

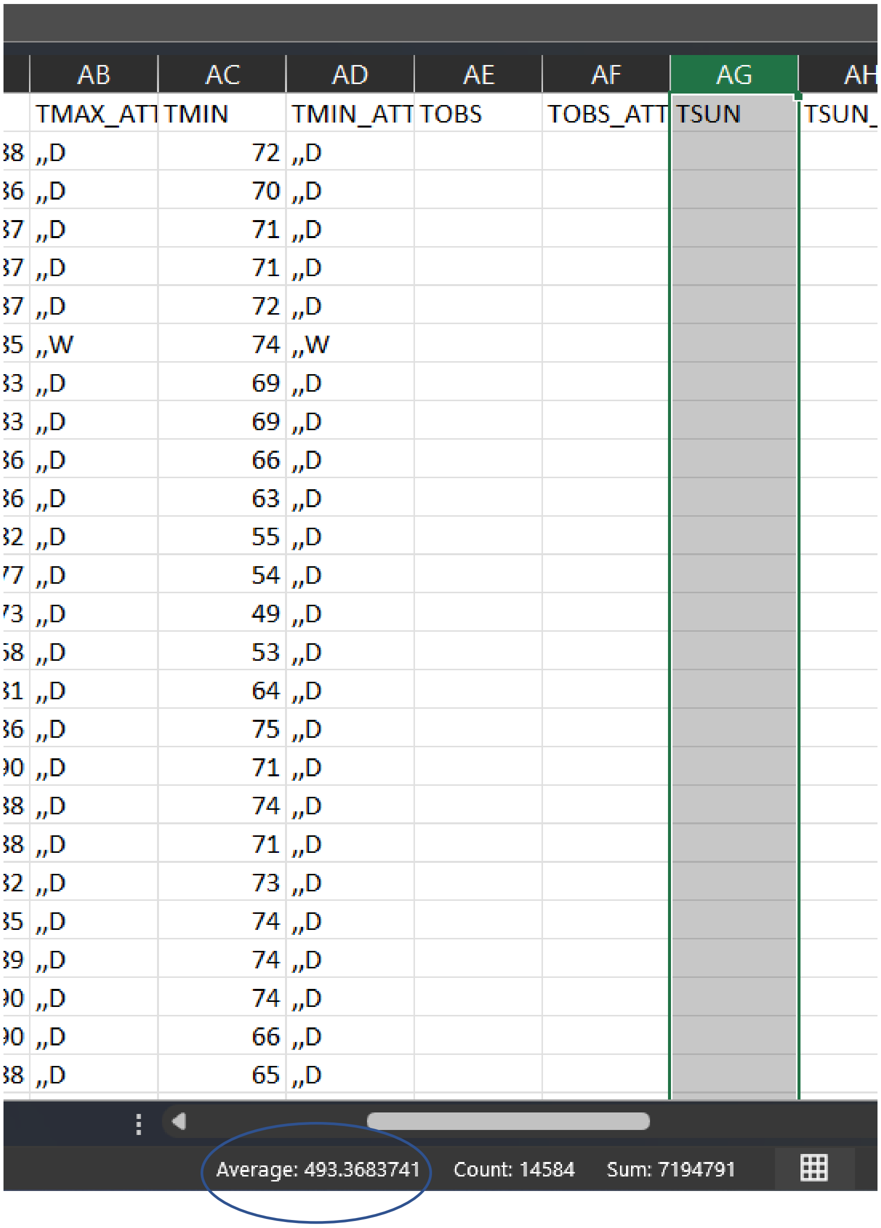
Independent variables:

ACMH	Average Cloudiness Midnight to midnight (percent)
ACSH	Average Cloudiness Sunrise to sunset (percent)
PSUN	Daily percent of Possible Sunshine (percent)
TAVG	Average temperature (Fahrenheit)
TMAX	Maximum temperature (Fahrenheit)
TMIN	Minimum temperature (Fahrenheit)
FMTM	Time of fastest mile or fastest 1-minute wind (hours and minutes)
PGTM	Peak Gust Time (hours and minutes,i.e,HMM)
AWND	Average daily wind speed (miles per hour)
WSF1	Fastest 1-minute Wind Speed (miles per hour)
WSF2	Fastest 2-minute Wind Speed (miles per hour)
WSF5	Fastest 5-second Wind speed (miles per hour)
WSFG	Peak Guest Wind speed (miles per hour)

Dependent variable/Target variable:

TSUN: Daily total sunshine in minutes.

In the real time data set, the daily average of the total sunshine for the past few decades is 493.368 minutes per day which implies, TSUN falls below 9 hours per day as shown in fig 5.1 satisfying the equation 2.1. Hence our dataset can be used for the prediction.



	AB	AC	AD	AE	AF	AG	AH
	TMAX_ATT	TMIN	TMIN_ATT	TOBS	TOBS_ATT	TSUN	TSUN_
38	,,D	72	,,D				
36	,,D	70	,,D				
37	,,D	71	,,D				
37	,,D	71	,,D				
37	,,D	72	,,D				
35	,,W	74	,,W				
33	,,D	69	,,D				
33	,,D	69	,,D				
36	,,D	66	,,D				
36	,,D	63	,,D				
32	,,D	55	,,D				
77	,,D	54	,,D				
73	,,D	49	,,D				
58	,,D	53	,,D				
31	,,D	64	,,D				
36	,,D	75	,,D				
90	,,D	71	,,D				
38	,,D	74	,,D				
38	,,D	71	,,D				
32	,,D	73	,,D				
35	,,D	74	,,D				
39	,,D	74	,,D				
90	,,D	74	,,D				
90	,,D	66	,,D				
38	,,D	65	,,D				

Average: 493.3683741 Count: 14584 Sum: 7194791

Fig 5.1

The output of the model describes the goodness of the fit and the error margin. For determining the goodness, we consider the coefficient of correlation and the mean absolute error.

5. DATA MODELLING

Model Building with Azure ML:

After the exploratory data analysis, the data is split into 70% (for training) and -30% (testing).. Next the dependent variable is selected in the train model node. The input for the train model is taken from the split data and regression models. here we choose three types of regression models they are boosted decision tree regression, neural network regression, Bayesian linear regression. Now the "score model" component is added to the process flow (to confirm the results obtained). Finally, the results are visualized using the "Evaluate model" component which gives the required output in the form of coefficient of correlation and mean absolute error. Figure 3.1 shows that complete structure of the model design in the Azure ML studio.

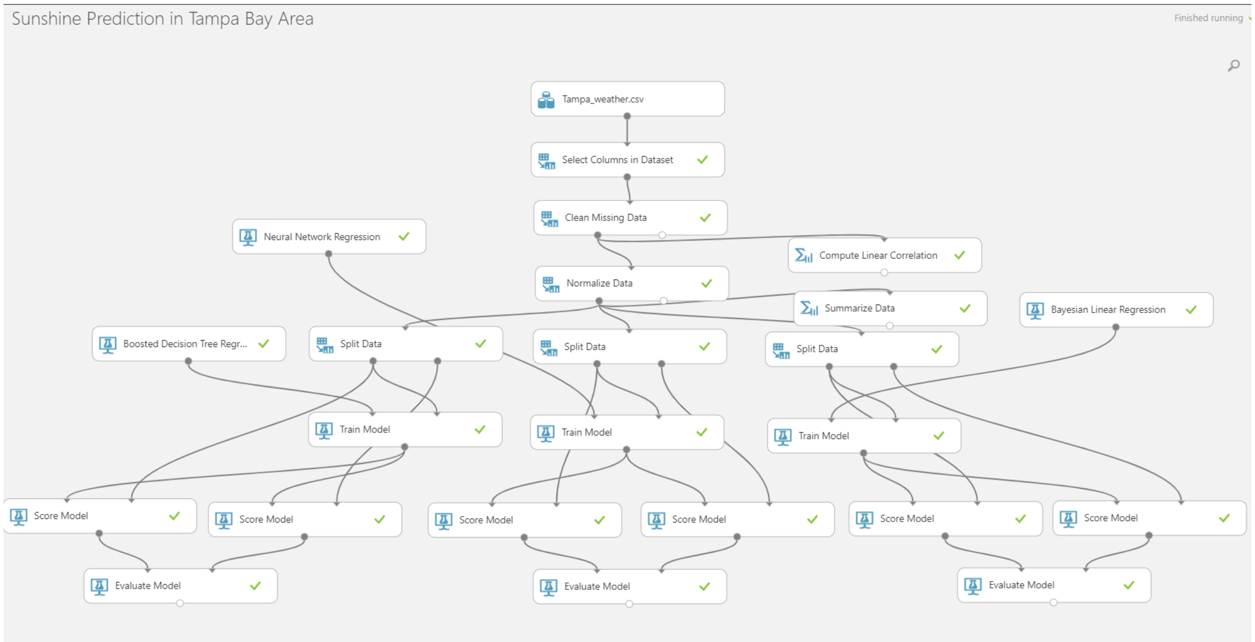
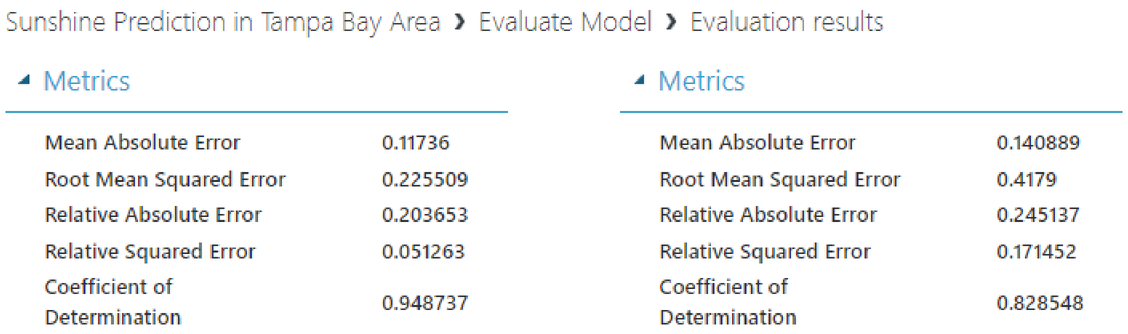


Fig 5.2 ML model in AZURE ML STUDIO

6. RESULTS

Upon visualization of the evaluate model node various parameters of evaluation of the predicted model is shown in the fig 6.1.

Boosted Decision Tree Regression:



Sunshine Prediction in Tampa Bay Area › Evaluate Model › Evaluation results

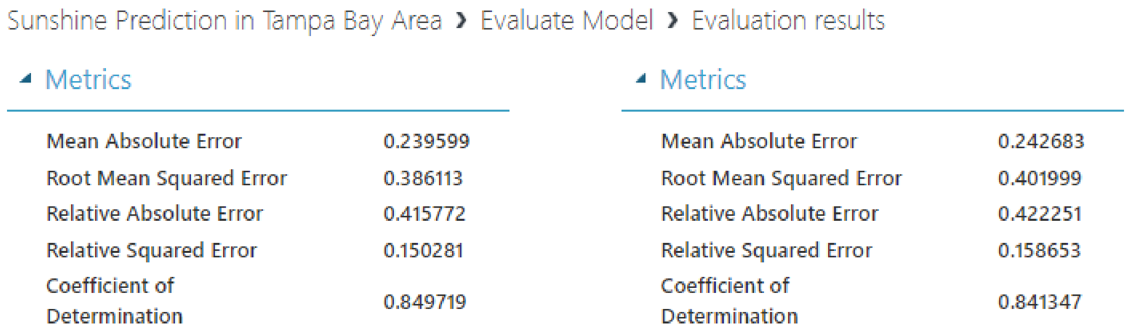
Metrics

Mean Absolute Error	0.11736
Root Mean Squared Error	0.225509
Relative Absolute Error	0.203653
Relative Squared Error	0.051263
Coefficient of Determination	0.948737

Metrics

Mean Absolute Error	0.140889
Root Mean Squared Error	0.4179
Relative Absolute Error	0.245137
Relative Squared Error	0.171452
Coefficient of Determination	0.828548

Neural Network Regression:



Sunshine Prediction in Tampa Bay Area › Evaluate Model › Evaluation results

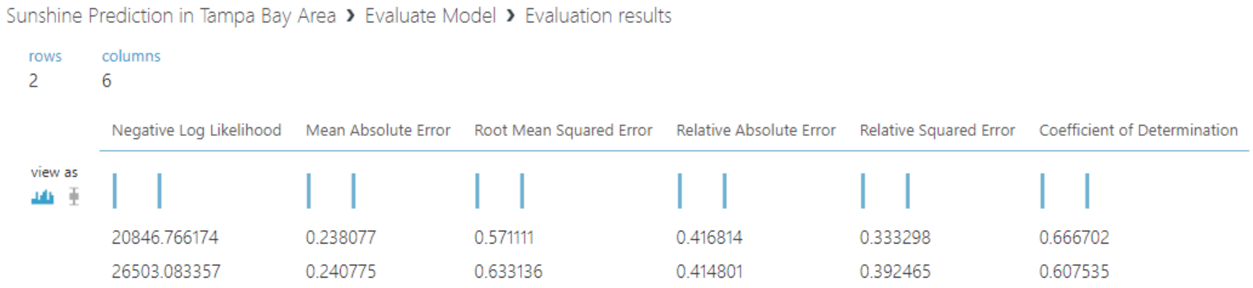
Metrics

Mean Absolute Error	0.239599
Root Mean Squared Error	0.386113
Relative Absolute Error	0.415772
Relative Squared Error	0.150281
Coefficient of Determination	0.849719

Metrics

Mean Absolute Error	0.242683
Root Mean Squared Error	0.401999
Relative Absolute Error	0.422251
Relative Squared Error	0.158653
Coefficient of Determination	0.841347

Bayesian Linear Regression:

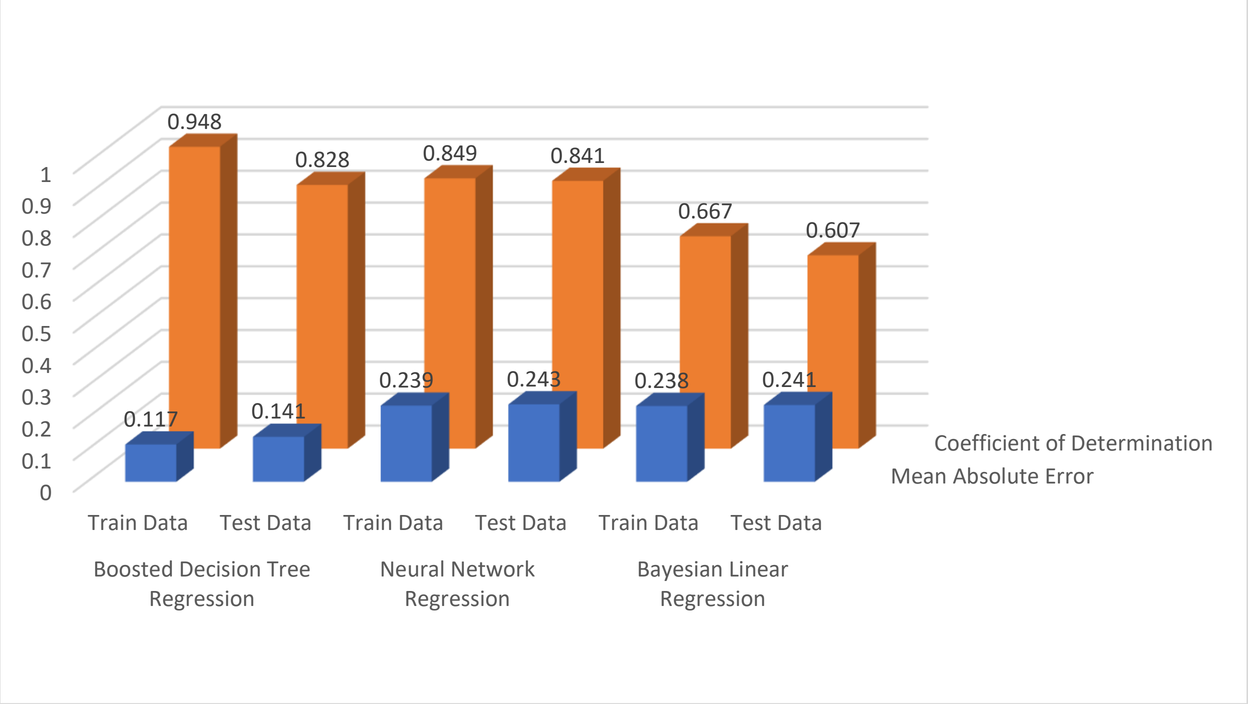


Sunshine Prediction in Tampa Bay Area › Evaluate Model › Evaluation results

rows 2 columns 6

Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
20846.766174	0.238077	0.571111	0.416814	0.333298	0.666702
26503.083357	0.240775	0.633136	0.414801	0.392465	0.607535

Model	Type of split Data	Mean Absolute Error	Coefficient of Determination
Boosted Decision Tree Regression	Train Data	0.115	0.96
	Test Data	0.138	0.837
Neural Network Regression	Train Data	0.225	0.784
	Test Data	0.229	0.674
Bayesian Linear Regression	Train Data	0.06	0.951
	Test Data	0.132	0.823



7. CONCLUSIONS:

From the above bar chart, it is evident that boosted decision tree regression has the highest coefficient of determination of 0.948 which indicates that the model is the perfect fit in the goodness level. Also, for the boosted decision tree regression, the train data and test data has the minimum of mean absolute error and there is no error difference between the train data and test data.

In the second model of Neural Network Regression the train data and test data has absolute error quite higher than the boosted decision tree regression but the coefficient of determination is around 85% which is a good fit.

In the third regression model of Bayesian linear regression, the coefficient of determination is around 65%. This is lower compared to the above two regression models. Even though the absolute mean error of train and test data is like that of Neural Network Regression, it fails in the accuracy based on the coefficient of determination. Hence, we can say that Bayesian linear regression is not a good fit for the prediction while boosted decision tree regression is the best fit for the prediction.