INFO 7390
FINAL PROJECT_GROUP 8

--Hawaii Tourism Prediction

ZHAO WANG

LEZI WANG

QIAOYI HE

CONTENT

1. PROBLEM STATEMENT & REQUIREMENTS

As we all know, tourism industry is the largest capital source of Hawaii economy, tourism contributed to $1.5 billion in total state tax revenue in 2013, an incremental $30 million year over year.

And tourism is also the biggest generator of jobs among the major economic sectors, supporting 168,000 jobs in Hawaii in 2013.

During the past 9 years (from 2007-2015), millions of visitors come from domestic and international to enjoy their vacations in Hawaii, and predicting the visitor amount in future becomes a critical problem in Hawaii tourism industry.

Therefore, we are planning to explore the datasets, which contained visitors' information of Hawaii for past nine years, and build prediction models, in order to solve following problems:

1.Supply and demand balancing: forecast future tourism resources demand and highest leveling to avoid supply shortage.

2.Market making: help government develop tourism industry plan, and improve service quality.

3.Identify and diversify Hawaii's global markets.

2. ASSUMPTION

1. No unpredictable disasters, like earthquake, hurricane, seismic sea wave, infectious disease.

2. Stable political and macroeconomics environment, do not take the global financial crisis or rise of the oil price into account.

3. Ignoring the change of transport, like closure of airlines and cruise ships.

3. BUSINESS GOAL & DELIVERABLE MODELS

3.1 END USERS

The end user is the Hawaii government and related workers.

3.2 BUSINESS GOAL SOLUTIONS

Based on our problem statements and business goals—to guarantee the balance of tourism resources supply and demand, we decide to build following deliverable models for end users.

3.2.1 MODEL1

Goal: Predict monthly visitor amount in each island from multiple countries, diversify Hawaii's global and domestic major markets.

3.2.2 MODEL2

Goal: Predict monthly total visitor amount for one specific island in Hawaii area, in order to build more accurate prediction models for different islands.

3.2.3 MODEL3

Goal: Predict monthly total visitor amount in entire Hawaii area, enhance strategic plans to incorporate marketing programs that drive travel demand, visitor arrivals and spending.

3.2.4 MODEL4

Goal: Predict monthly total visitors' expenditures in entire Hawaii area, in order to enhance and promote the profits of Hawaii's tourism industry.

4. DATASET SOURCE & DESCRIPTION

1. Get Hawaii monthly visitor records from Hawaii government website.

http://dbedt.hawaii.gov/visitor/tourism/

2. Get Hawaii temperature records from US climate websites.

http://www.usclimatedata.com/climate/honolulu/hawaii/united-states/ushi0026

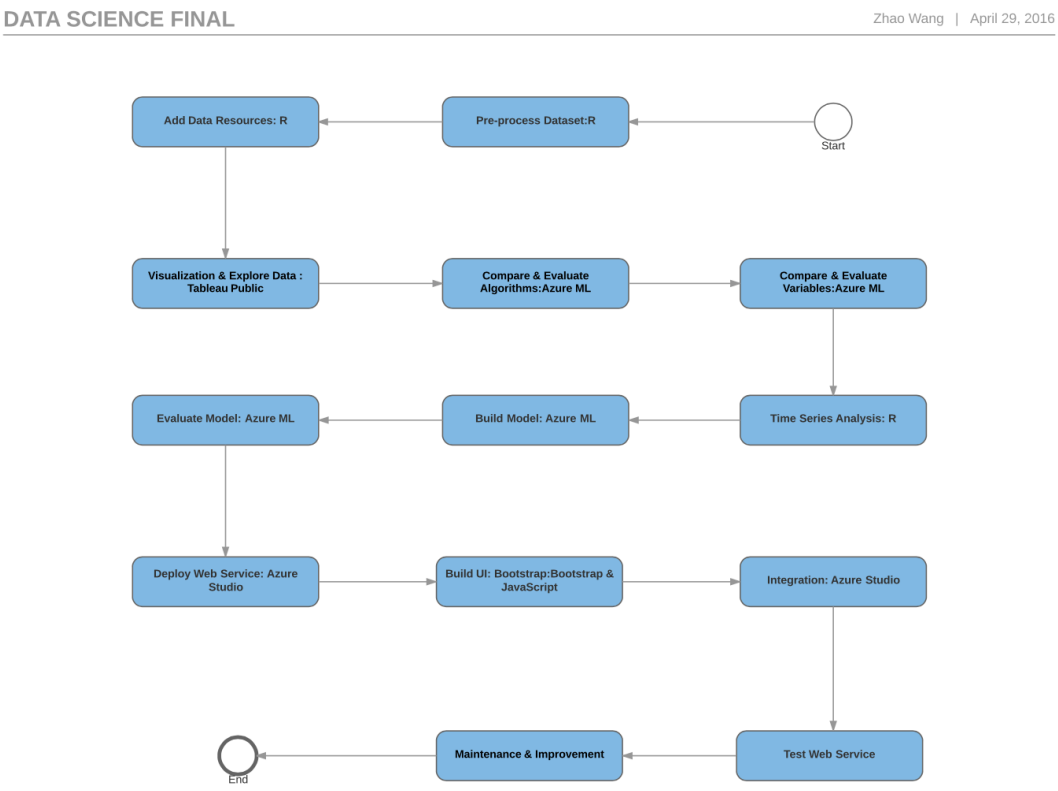
3. Get US monthly vacation days from timeanddate.com.

http://www.timeanddate.com/holidays/us/

4. Get Hawaii monthly tourism incomes (total visitors' expenditures) data from Hawaii Tourism website.

http://www.hawaiitourismauthority.org/research/reports/historical-visitor-statistics/

5. PROCESS STEPS & TOOLS



1) Pre-process Dataset→ R

2) Add data Resources→ R

3) Visualization &Explore Dataset →Tableau Public

4) Compare & Evaluate Algorithms →Azure Machine Learning

5) Compare & Evaluate Variables →Azure Machine Learning

6) Time Series Analysis→ R

7) Build Machine Learning Model→ Azure Machine Learning

8) Evaluate model→ Azure Machine Learning

9) Deploy web service→ Azure Studio

10) Build UI → Bootstrap & JavaScript

11) Integration → Azure Studio

12) Test web service

13) Maintenance & Improvement

6. PRE-PROCESS DATA

6.1 UNIFORM UNITS AND RESTRUCTURE

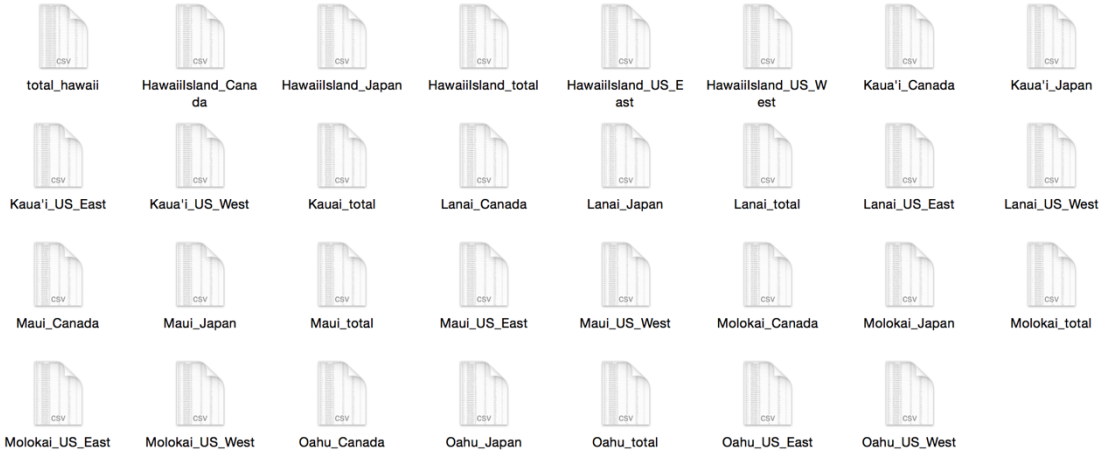
We get Hawaii monthly visitor statistics from Hawaii government website (http://dbedt.hawaii.gov/visitor/tourism/) and the monthly expenditure statistics from Hawaii tourism website(http://www.hawaiitourismauthority.org).

After down load all the data sets, we uniform units, the unit of expenditure is million and the unit of visitors' amount is ten thousand. After uniform units, we transpose the raw dataset and use the Hawaii island name as a new column named Island and use the source country as another new column named country to distinct different datasets. We only have four different countries – Canada, Japan, The U.S. East and The U.S. West, in our dataset, since those four countries are major source country of Hawaii tourism.

Then we import the data set into R for further processing. The following picture is the R code which we use to keep all the digits to ensure better accuracy.

```
data1.fulldigits <- format(data1, digits = 20)
```

The following picture is all data sets we have.



6.2 ADD RELATIVE VARIABLE

6.2.1 SEPARATION TIME INTO YEAR AND MONTH

The following picture is the R code which we use to add year and month into data set.

```
#import the time date
datatem<-read.csv("~/Desktop/time.csv")
#Add month and year data into dataset
tem<-datatem
data1<-data.frame(tem,data1.fulldigits)
```

6.2.2 ADD MORE VARIABLE.

After thorough explore the data, we think temperature and vacation day of month has a great impact on visitors' amount, so we add three more parameters relative the temperature– average maximum temperature, average minimum temperature and average temperature, and one parameter about the vacation day into dataset

We got the monthly Hawaii Temperature Record from US climate websites (http://www.usclimatedata.com/climate/honolulu/hawaii/united-states/ushi0026), and US monthly vacation days from this URL:

(http://www.timeanddate.com/holidays/us/).

The following picture is the R code which we use to add those parameters into dataset.

```
#import the temperature and vacation date
datatem2<-read.csv("~/Desktop/temperature&vacation.csv")
#Add temperature data into dataset
tem<-datatem2
data1<-data.frame(data2,tem)
```

6.3 CLEAN DATA SETS

We need to clean missing data for improving the usability and stability of data. The following picture is the R code which we use to clean the data set. We delete rows which contain missing value.

```
#clear missing data
data1<-na.omit(data1)
```

6.4 COMBINE DATA SETS:

Finally, we combine all data sets into a single data set. The following picture is the R code which we use to combine the data.

```
setwd("/Users/lucy/Desktop/cobming/")
temp<- list.files(pattern="*.csv")
hawaiiset <- do.call(rbind, lapply(temp, fread))
write.csv(hawaiiset,"/Users/lucy/Desktop/hawaiisetone.csv")
```

6.5 NEW STRUCTURE OF DATASETS

After combing data sets, now we have three different datasets for building different models.

The following picture shows the part of first data set which include 9 columns – Year, Month, Visitors' amount(visitors), average highest temperature, average lowest temperature, Average temperature, extra vacation day, island, Country. The dataset be used to predict monthly visitor amount in each island from multiple countries.

	A	B	C	D	E	F	G	H	I
1	Year	Month	Visitors	Average hig	Average low	Average ten	extra vacatic	Island	Country
2	2007	1	115535.638	80.9	68.8	74.85	2	HawaiiIslanc	Canada
3	2007	2	100557.7	80.2	66.7	73.45	1	HawaiiIslanc	Canada
4	2007	3	95819.4767	80.5	67.9	74.2	0	HawaiiIslanc	Canada
5	2007	4	52189.8731	83.6	69.7	76.65	1	HawaiiIslanc	Canada
6	2007	5	35086.8204	85	71.6	78.3	0	HawaiiIslanc	Canada
7	2007	6	26549.1139	87.3	74.1	80.7	1	HawaiiIslanc	Canada
8	2007	7	32061.7345	88	75.2	81.6	0	HawaiiIslanc	Canada
9	2007	8	45759.3657	88.3	75.8	82.05	0	HawaiiIslanc	Canada
10	2007	9	33578.4448	88.2	74.9	81.55	1	HawaiiIslanc	Canada
11	2007	10	49379.9716	86.3	74	80.15	1	HawaiiIslanc	Canada
12	2007	11	71148.8788	82.7	70.5	76.6	2	HawaiiIslanc	Canada
13	2007	12	99489.1121	80	71.1	75.55	2	HawaiiIslanc	Canada

The following picture shows the part of second data set which include 8 columns – Year, Month, Visitors' amount(visitors), average highest temperature, average lowest temperature, Average temperature, extra vacation day, island. The dataset be use to predict monthly total visitor amount for one specific island in Hawaii area.

	A	B	C	D	E	F	G	H
1	Year	Month	total_vistors	Average hig	Average low	Average ten	extra vacatic	Island
2	2007	1	195264.608	80.9	68.8	74.85	2	Maui
3	2007	2	196700.12	80.2	66.7	73.45	1	Maui
4	2007	3	227232.515	80.5	67.9	74.2	0	Maui
5	2007	4	202215.773	83.6	69.7	76.65	1	Maui
6	2007	5	198130.154	85	71.6	78.3	0	Maui
7	2007	6	241790.41	87.3	74.1	80.7	1	Maui
8	2007	7	247535.244	88	75.2	81.6	0	Maui
9	2007	8	237113.276	88.3	75.8	82.05	0	Maui
10	2007	9	186111.351	88.2	74.9	81.55	1	Maui
11	2007	10	190684.617	86.3	74	80.15	1	Maui
12	2007	11	184472.898	82.7	70.5	76.6	2	Maui
13	2007	12	214791.747	80	71.1	75.55	2	Maui

The following picture shows the part of third data set which include 8 columns – Year, Month, Total Visitors' amount, monthly expenditure, average highest temperature, average lowest temperature, Average temperature, extra vacation day. The dataset be use to predict monthly total visitor amount and monthly total visitors' expenditures in entire Hawaii area.

	A	B	C	D	E	F	G	H
1	Year	Month	Total_Visitor	expenditure	Average hig	Average low	Average ten	extra vacatio
2	2007	1	577231.793	1089.87397	80.9	68.8	74.85	2
3	2007	2	574762.708	996.76546	80.2	66.7	73.45	1
4	2007	3	674532.008	1028.06454	80.5	67.9	74.2	0
5	2007	4	597477.56	957.542315	83.6	69.7	76.65	1
6	2007	5	586545.552	922.25895	85	71.6	78.3	0
7	2007	6	672585.524	1135.55192	87.3	74.1	80.7	1
8	2007	7	711263.325	1191.92992	88	75.2	81.6	0
9	2007	8	733025.281	1177.58518	88.3	75.8	82.05	0
10	2007	9	558430.761	911.175938	88.2	74.9	81.55	1
11	2007	10	570646.621	969.321885	86.3	74	80.15	1
12	2007	11	576370.975	950.496904	82.7	70.5	76.6	2

7. VISUALIZATION DATA

For Visualization, we use Tableau Public tools. After data pre-processing, we have following 6 datasets in total, and we do visualization for each dataset separately.

1) Dataset1→Model1: Monthly total visitor numbers in each island from different countries, from 2007-2015.

2) Dataset2→Model2: Monthly total visitor numbers for each island, from 2007-2015.

3) Dataset3→Model3: Monthly total visitor numbers in Hawaii area, from 2007-2015.

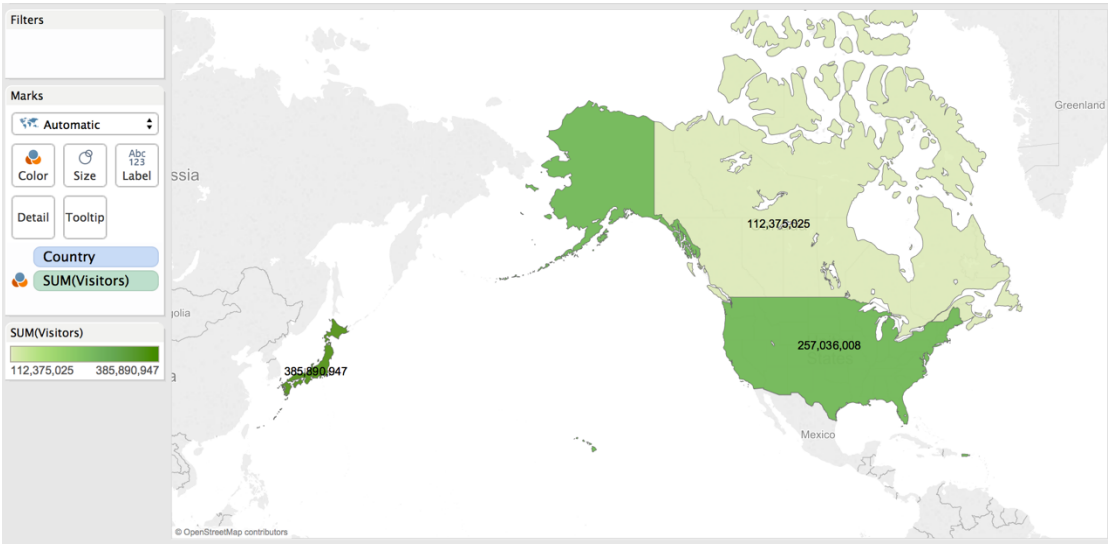
4) Dataset4→Model4: Monthly total expenditures in Hawaii area, from 2007-2015.

5) Dataset5→Model4: Predicted 2016 monthly total expenditures in Hawaii area.

6) Dataset6→Model4: Predicted 2017 monthly total expenditures in Hawaii area.

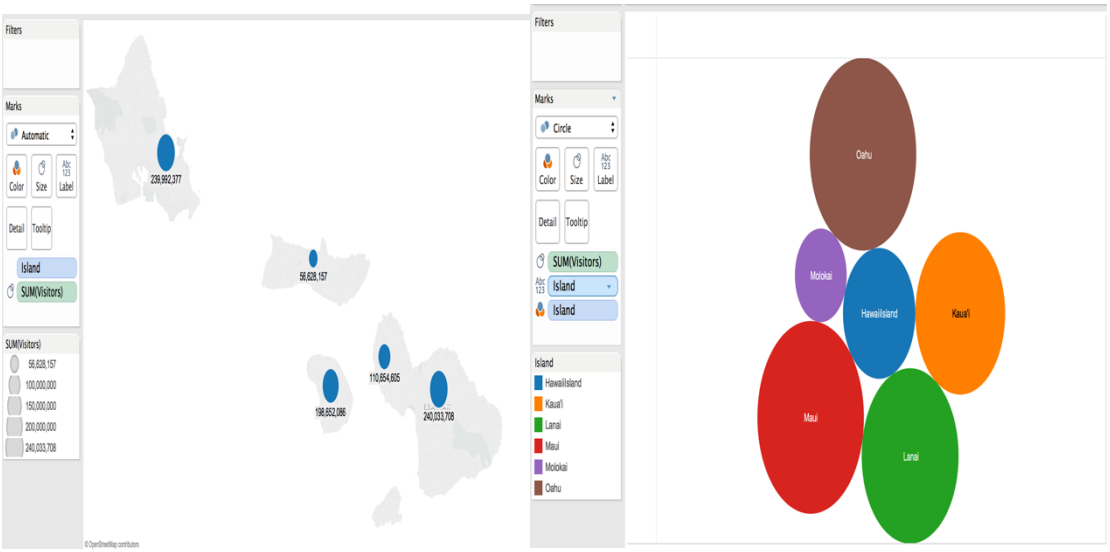
7.1 MODEL1 (DATASET1) VISUALIZATION

1. Explore total visitor (past 9 years) distribution by countries



As shown in above picture, visitors in Hawaii mainly come from Japan, US and Canada. Among them, about 50% visitors come from Japan, 30% visitors come from US, and 20% visitors come from Canada.

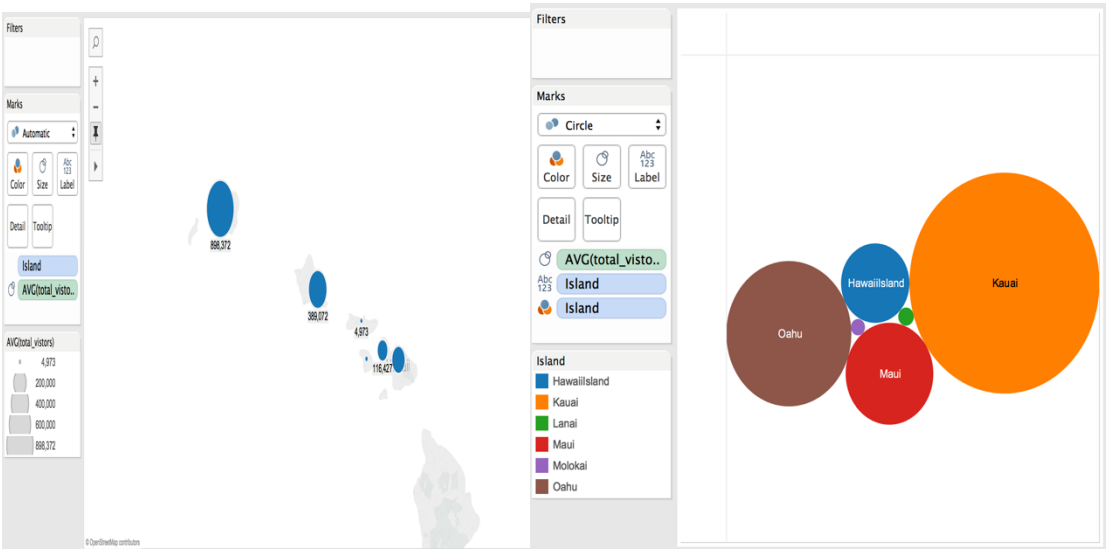
2. Explore total visitor (past 9 years) distribution by islands in Hawaii



As shown in above picture, total visitor numbers of different islands are really different. Maui Island and Oahu Island have the most visitor numbers -about 240 million in past 9 years, and then the Lanai Island-about 200 million visitors in past 9 years. And the Molokai Island has the least visitor numbers-about 5 million visitors in past 9 years.

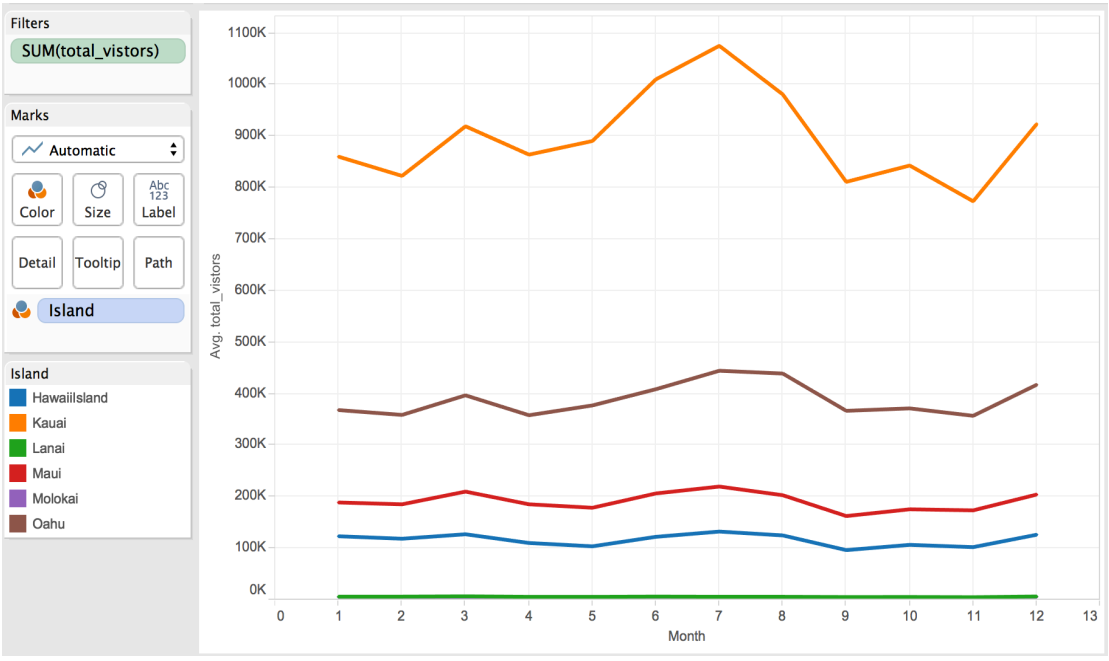
7.2 MODEL2 (DATASET2) VISUALIZATION

1. Explore monthly average visitor distribution by islands in Hawaii



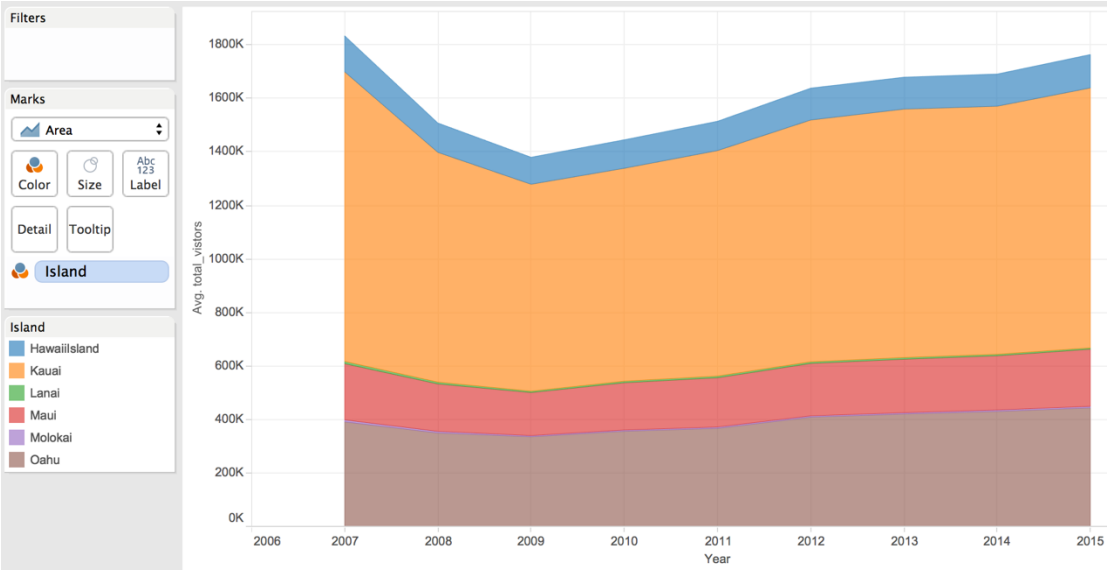
As shown in above pictures, monthly average visitor numbers of different islands are really different. Kauai Island has the most visitor numbers per month-about 898 thousands, and then the Oahu Island-about 389 thousands visitors per month. And the Molokai Island has the least visitor numbers-about 5 thousands per month.

2. Explore monthly average visitor numbers trends in different islands



As shown in above picture, monthly average visitor trends of different islands are really different. Kauai Island's visitor numbers are significantly affected by seasons, and it the visitor numbers increases in June and decreases in August. Visitor trends of other islands are pretty stable, which means those are not strongly affected by seasons.

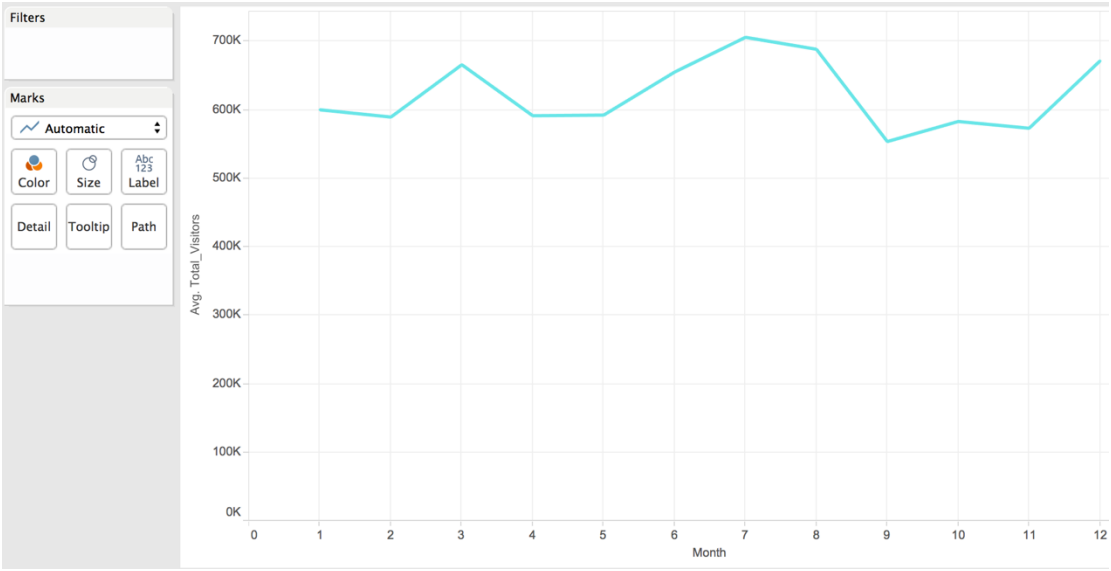
3. Explore yearly average visitor numbers trends in different islands



As shown in above picture, yearly average visitor trends of different islands are pretty similar. Almost all islands' visitor numbers decrease from 2006 to 2009, and then increase from 2009 to 2015, which means there may be some worldwide reasons that affects Hawaii's visitor number from 2006 to 2009.

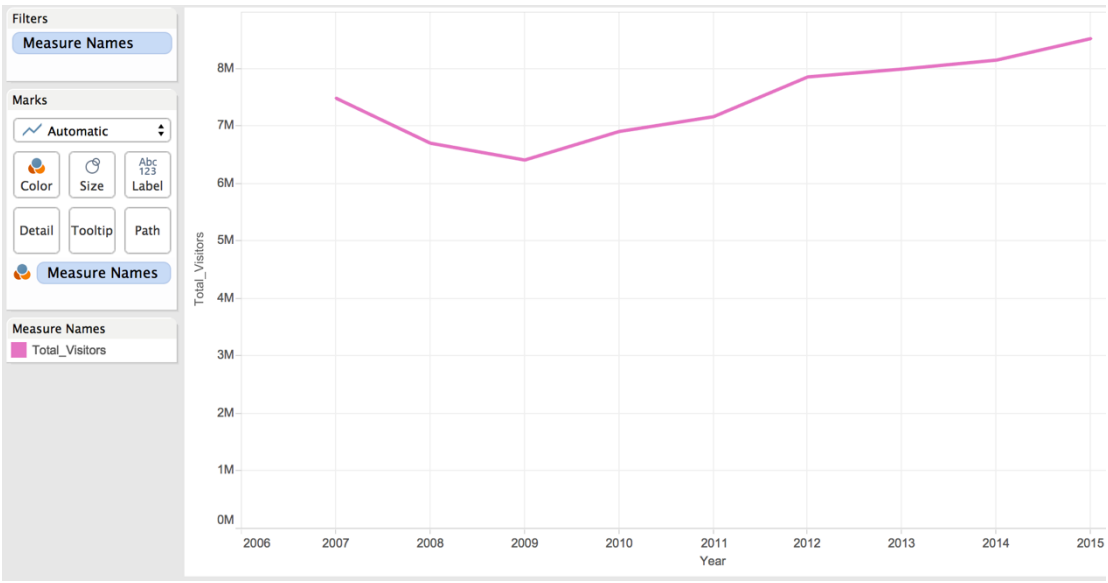
7.3 MODEL3 (DATASET3) VISUALIZATION

1. Explore monthly average total visitor trend



As shown in above picture, the total visitor numbers in Hawaii has relationship with month, as you can see, the total visitor numbers increase in March, June to August, and in December.

2. Explore yearly total visitor trend

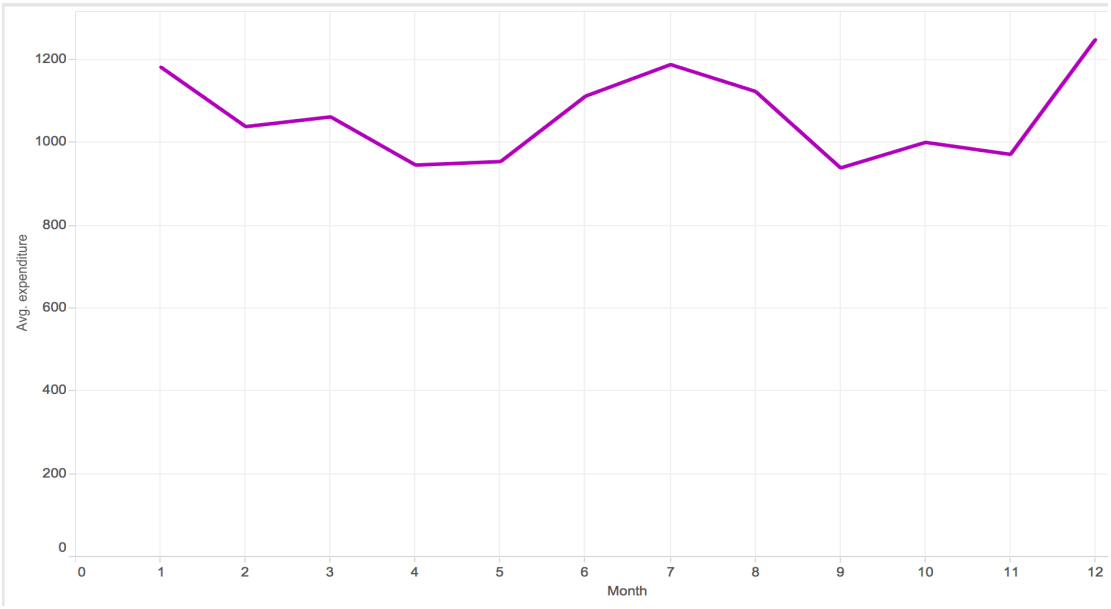


As shown in above picture, the total visitor number in Hawaii has significant yearly trend. It decreases from 2006 to 2009, and increases from 2009 to 2015, which means there may be some worldwide reasons that affects Hawaii's visitor number from 2006 to 2009.

7.4 MODEL4 DATASET4,5,6 VISUALIZATION

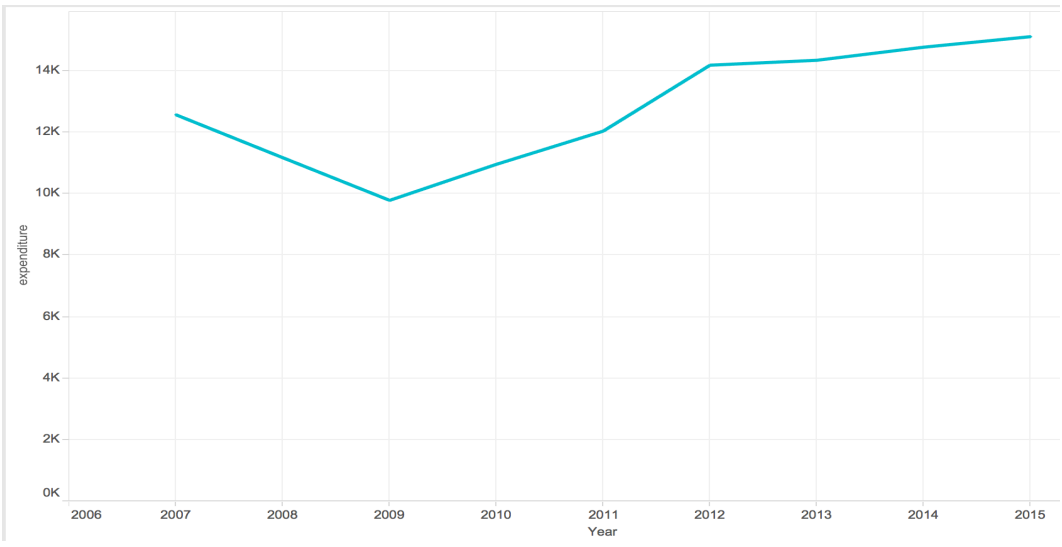
7.4.1 VISUALIZATION FOR VISITORS' EXPENDITURES FROM 2007-2015

1. Explore monthly average visitors' expenditures



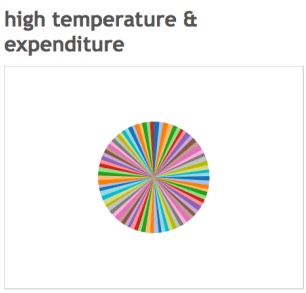
As shown in above picture, the monthly average visitors' expenditure has significant monthly trend. It decreases from Jan to May, and increases from May to July, and decreases again from July to November, and then increases again in December. Therefore, we can see that Hawaii has more visitors in Summer and Winter.

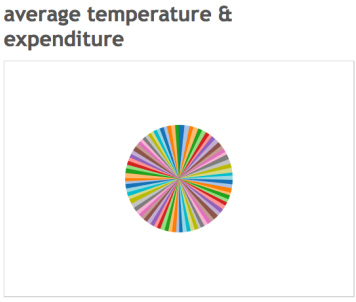
2. Explore yearly total visitors' expenditures

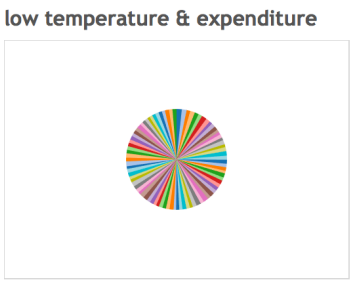


As shown in above picture, the yearly total visitors' expenditure in Hawaii has significant yearly trend. It decreases from 2006 to 2009, and increases from 2009 to 2015, which has similar trend with total yearly visitor numbers. It means total expenditure has strong correlation with total visitor number.

3. Explore average monthly expenditure and temperature

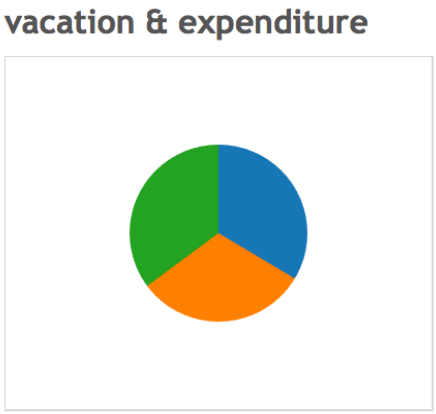






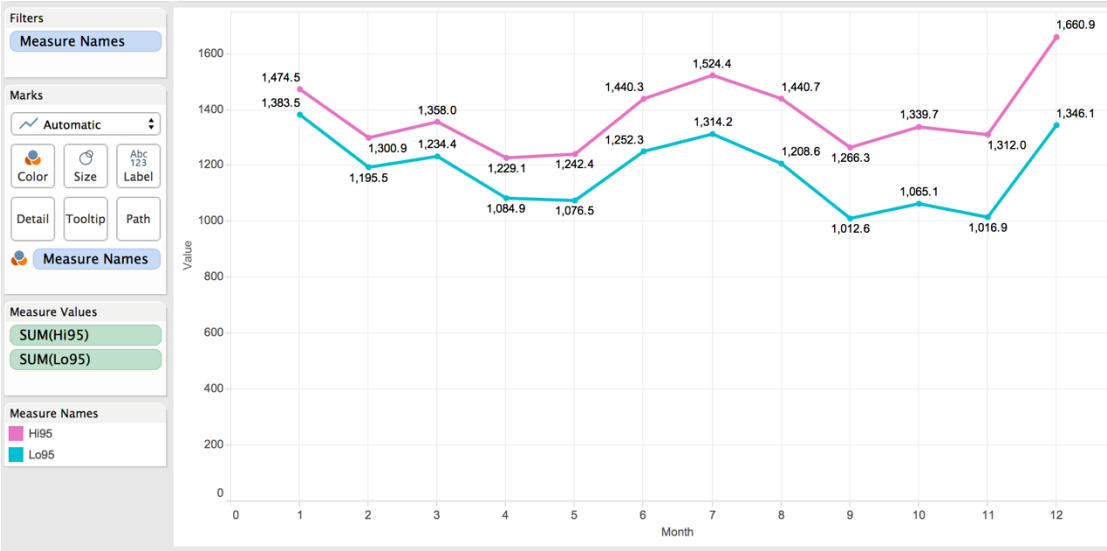
As shown in above pictures, it is hard to find a specific trend between temperature and expenditure; therefore, we can see average monthly expenditure does not have relationship with temperature.

4. Explore average monthly expenditure and monthly extra vacation day numbers



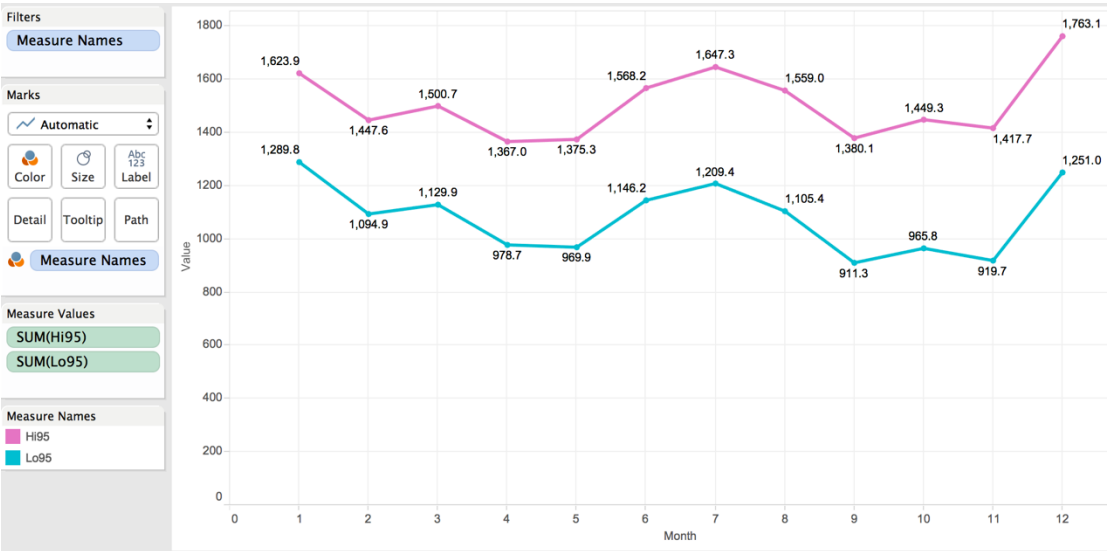
As shown in picture, average monthly expenditure does not have relationship with extra vacation days.

7.4.2 VISUALIZATION FOR PREDICTED MONTHLY TOTAL VISITORS' EXPENDITURES IN 2016



The above picture shows both highest and lowest predicted monthly total visitors' expenditures in 2016. As you can see, it decreases from Jan to May, and increase from May to July, and then decreases again from July to Nov, and then increases again in December.

7.4.3 VISUALIZATION FOR PREDICTED MONTHLY TOTAL VISITORS' EXPENDITURES IN 2017



The above picture shows both highest and lowest predicted monthly total visitors' expenditures in 2017. It has similar trends as 2016, but the amount increases a little.

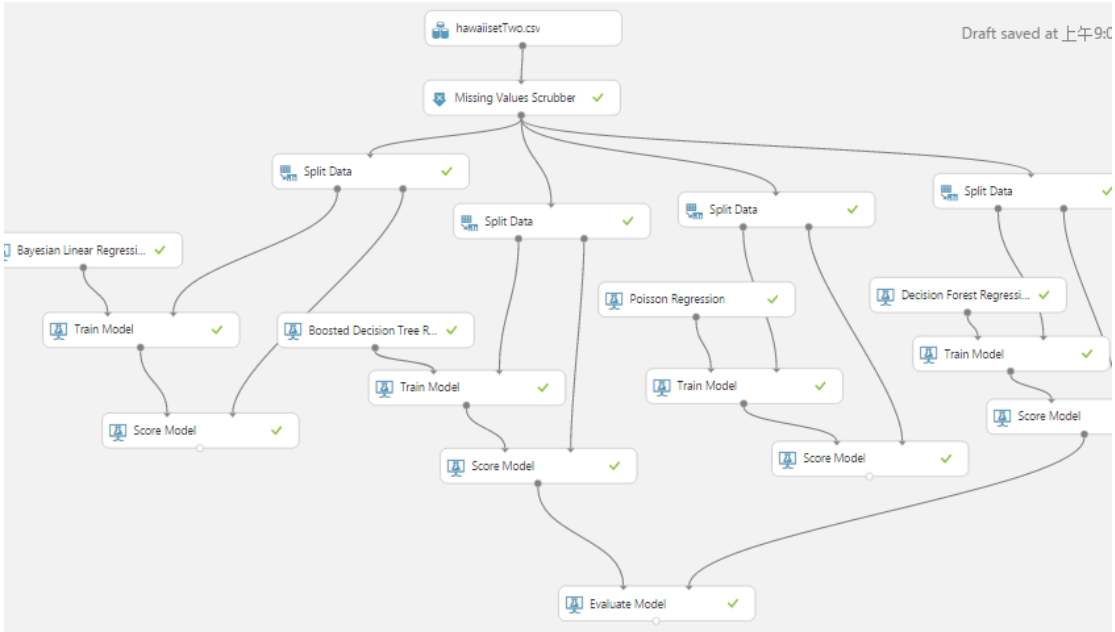
As you can see, it decreases from Jan to May, and increase from May to July, and then decreases again from July to Nov, and then increases again in December.

8. COMPARE & EVALUATE MODEL

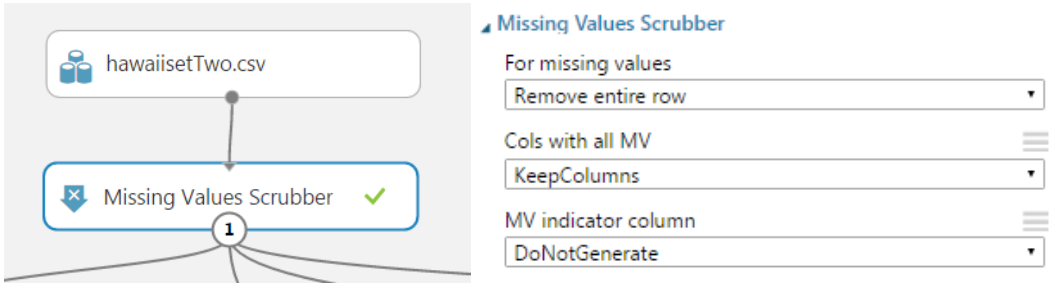
Because there are many algorithms for machine learning, so we used four algorithm in Azure Machine Leaning Studio to build prediction models, and compared their predict result, chosen the algorithm who has the best performance. In the Building Models Part, we used this best performance algorithm.

8.1 COMPARING ALGORITHM MODEL OVERVIEW

We post our model screen shot here, and in later several chapters, we will descript this model step by step very details.



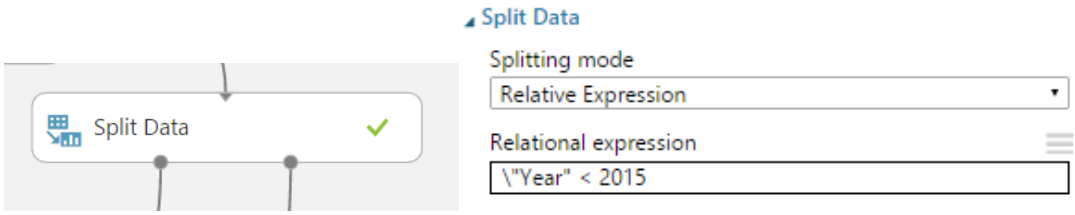
8.2 GET DATASET AND REMOVE MISSING DATA



We used the dataset which is saved in Azure Machine Learning Studio to build prediction models. We have three dataset for the whole project, and this dataset is used for the second prediction model which can predict the total visitor amount for each major island of Hawaii.

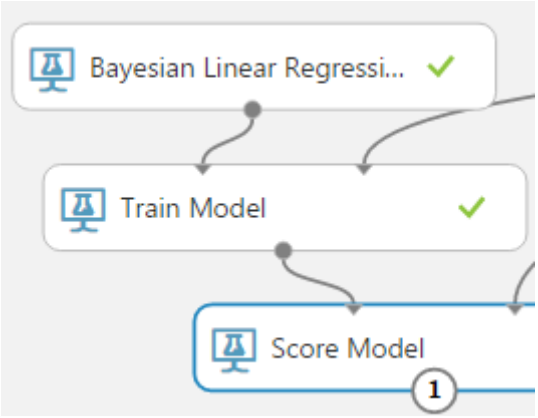
After importing the dataset, we used the Missing Values Scrubber to remove the entire row which contains missing data or N/A. Because missing data will affect the prediction result, we must deal with it before building models. You may replace missing data with median or average of that column, but for this situation we remove that row.

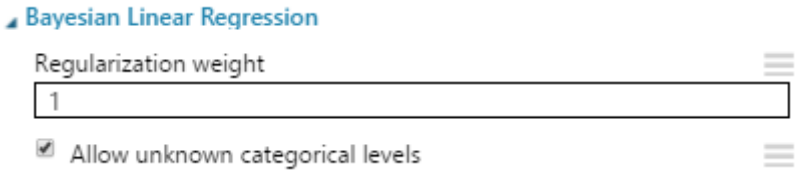
8.3 SPLIT DATA BASED ON YEAR



Because these dataset satisfy the requirements of Time Series Model, we followed the rule of Time Series Model for splitting data—splitting data based on Year. In our dataset, there are 9 years records (2007-2015) for each month. Thus, we split these data into 2007-2014's records for training data and 2015's recodes for validation data. As you can see in right picture, the Relation expression is "Year <2015", this is for split 2007-2014's records from all records.

8.4 TRAIN MODEL WITH BAYESIAN LINEAR REGRESSION AND SCORE





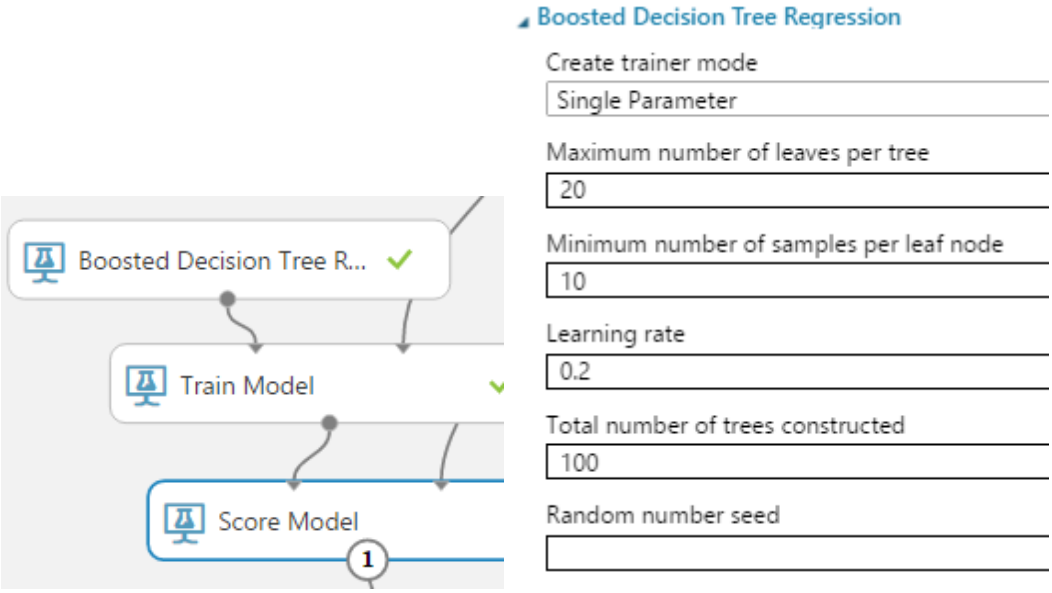
The first algorithm is Bayesian Linear Regression. In statistics, it is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference. When the regression model has errors that have a normal distribution, and if a particular form of prior distribution is assumed, explicit results are available for the posterior probability distributions of the model's parameters.

In our model, we choose 1 as the regulation weight. Usually we set the weight as 1, so there is no need to change it.

Year	Month	total_vistors	Average high temperature	Average low temperature	Average temperature	extra vacation	Island	Scored Label Mean	Scored Label Standard Deviation
2015	1	214819.9659	62.9	48.2	55.55	2	Maui	162386.747054	100049.021087
2015	2	198999.3332	68.8	53.9	61.35	1	Maui	182150.765689	99619.54011
2015	3	234905.7754	68	52.5	60.25	0	Maui	195160.433049	99743.255487
2015	4	205070.8646	72.2	55	63.6	1	Maui	183249.254028	99605.268362
2015	5	209508.309	65.6	51.5	58.55	0	Maui	195413.574995	100135.139456
2015	6	233046.9678	73.1	55.5	64.3	1	Maui	184360.600935	99600.431636
2015	7	245896.4667	75.5	60.3	67.9	0	Maui	204068.453916	99579.930467

The Score of this algorithm model posted above, you can see the difference between the real value and the mean of predicted value is not small. The differences are almost 50000, and the unit is myriad. Of course, we cannot evaluate an algorithm only based on the score, but for scoring part, this algorithm does not have a good performance.

8.5 TRAIN MODEL WITH BOOSTED DECISION TREE REGRESSION AND SCORE



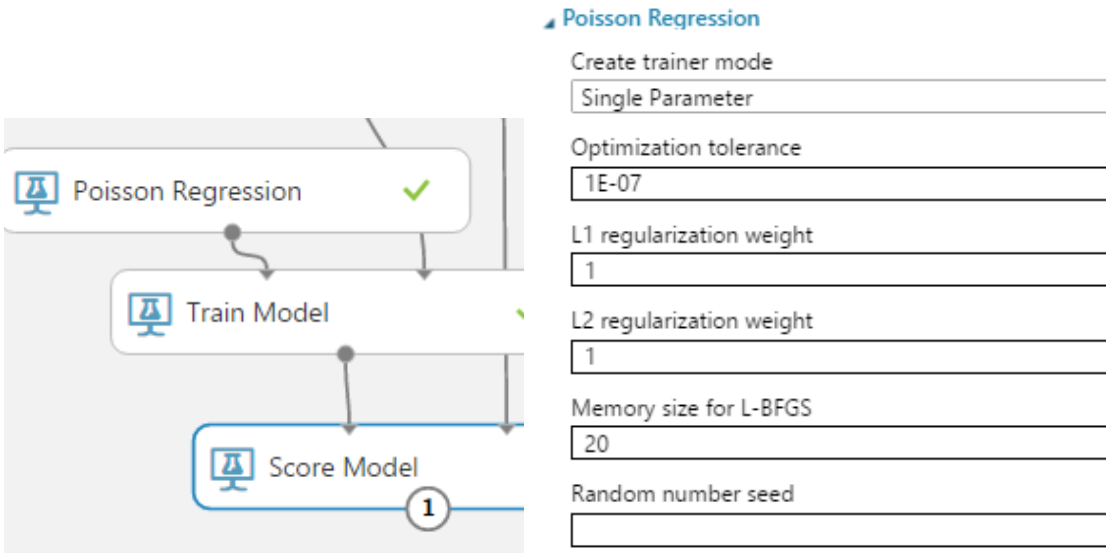
The second algorithm is Boosted Decision Tree Regression. It enables to create an ensemble of regression trees using boosting. Boosting means that each tree is dependent on prior trees, and learns by fitting the residual of the trees that preceded it. Thus, boosting in a decision tree ensemble tends to improve accuracy with some small risk of less coverage. This regression method is a supervised learning method.

In our model, we only use a single parameter. And we set the maximum and minimum leaves number per tree as normal. And the leaning rate is 0.2 has the best performance. We do not want this model spend too much time, so we set the trees number as 100.

Year	Month	total_vistors	Average high temperature	Average low temperature	Average temperature	extra vacation	Island	Scored Labels
2015	1	214819.9659	62.9	48.2	55.55	2	Maui	196203.671875
2015	2	198999.3332	68.8	53.9	61.35	1	Maui	201752.734375
2015	3	234905.7754	68	52.5	60.25	0	Maui	216471.078125
2015	4	205070.8646	72.2	55	63.6	1	Maui	192462.0625
2015	5	209508.309	65.6	51.5	58.55	0	Maui	192117.578125
2015	6	233046.9678	73.1	55.5	64.3	1	Maui	214409.453125

The Score of this algorithm model posted above, you can see the difference between the real value and the predicted value is not that much big. The differences are roughly 20000, and the unit is myriad, it smaller than the first algorithm's. Of course, we cannot evaluate an algorithm only based on the score, but for scoring part, this algorithm has a normal performance.

8.6 TRAIN MODEL WITH POISSON REGRESSION AND SCORE



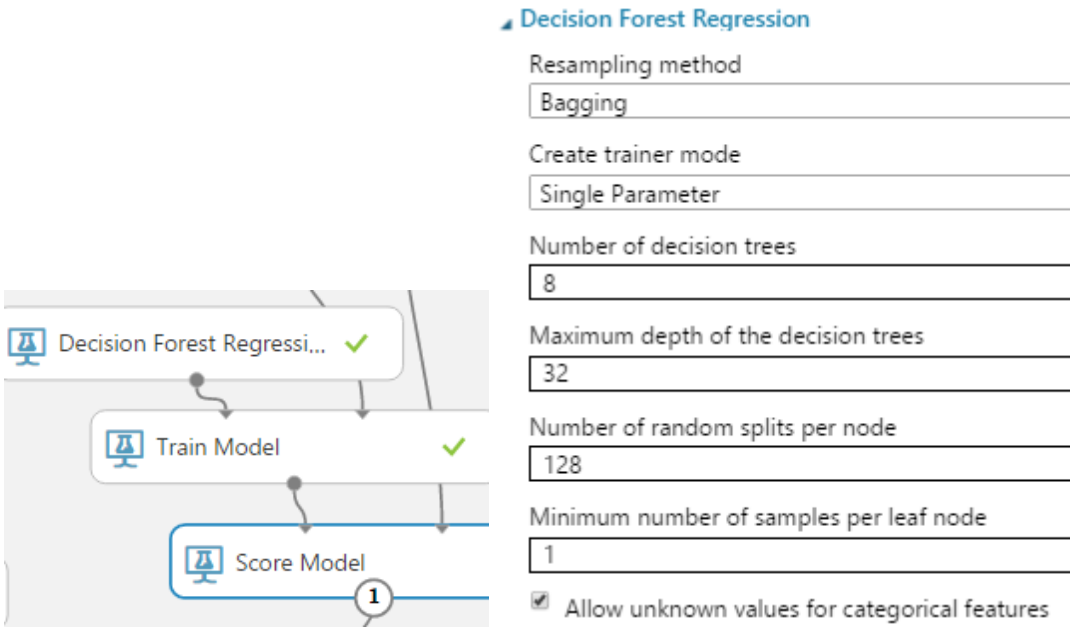
The third algorithm is Poisson Regression. In statistics, it is a form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable *Y* has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. A Poisson regression model is sometimes known as alog-linear model, especially when used to model contingency tables.

Poisson regression models are generalized linear models with the logarithm as the (canonical) link function, and the Poisson distribution function as the assumed probability distribution of the response.

Year	Month	total_vistors	Average high temperature	Average low temperature	Average temperature	extra vacation	Island	Scored Labels
2015	1	214819.9659	62.9	48.2	55.55	2	Maui	166043.622711
2015	2	198999.3332	68.8	53.9	61.35	1	Maui	170234.342101
2015	3	234905.7754	68	52.5	60.25	0	Maui	174468.382847
2015	4	205070.8646	72.2	55	63.6	1	Maui	170243.712007
2015	5	209508.309	65.6	51.5	58.55	0	Maui	174459.649169
2015	6	233046.9678	73.1	55.5	64.3	1	Maui	170248.011367

The Score of this algorithm model posted above, you can see the difference between the real value and the predicted value is not that much big. The differences are roughly greater than 30000, and the unit is myriad, it smaller than the first algorithm's but bigger than second's. Of course, we cannot evaluate an algorithm only based on the score, but for scoring part, this algorithm has a not that bad performance. So far, the second algorithm has the best performance in scoring part.

8.7 TRAIN MODEL WITH DECISION FOREST REGRESSION AND SCORE



The fourth algorithm is Decision Forest Regression. It is used to create a regression model using an ensemble of decision trees. Decision trees are non-parametric models that perform a sequence of simple tests for each instance, traversing a binary tree data structure until a leaf node (decision) is reached.

In our model, we use Bagging as the re-sampling method. And we used a single parameter, and 8 decision trees.

Year	Month	total_vistors	Average high temperature	Average low temperature	Average temperature	extra vacation	Island	Scored Label Mean	Scored Label Standard Deviation
2015	1	214819.9659	62.9	48.2	55.55	2	Maui	193362.504663	6618.615888
2015	2	198999.3332	68.8	53.9	61.35	1	Maui	195639.658442	5854.769274
2015	3	234905.7754	68	52.5	60.25	0	Maui	214312.728277	16912.30164
2015	4	205070.8646	72.2	55	63.6	1	Maui	188456.257519	14009.340933
2015	5	209508.309	65.6	51.5	58.55	0	Maui	209307.254381	20483.745813
2015	6	233046.9678	73.1	55.5	64.3	1	Maui	206927.208723	11699.270409

The Score of this algorithm model posted above, you can see the difference between the real value and the mean of predicted value is not that much big. The differences are roughly 20000, and the unit is myriad, it is similar with second algorithm. Of course, we cannot evaluate an algorithm only based on the score, but for scoring part, this algorithm has a normal performance similar with second algorithm.

8.8 COMPARING ALGORITHM RESULT

After comparing the scoring of each algorithm, we found the second algorithm (Boosted Decision Tree Regression) and the fourth algorithm (Decision Forest Regression) have better performance on prediction values.

Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
909.851096	40545.591246	64267.14665	0.147051	0.035679	0.964321
Infinity	17420.976855	30573.238088	0.063182	0.009166	0.990834

Compared the first and second models, we can see the mean absolute error of second is less than the first a lot, that means the error rate of first is bigger than second's, and the average difference of first model between predictive values and real values are greater than second's. Also, the RMS of second is smaller than first's, RMS stand for same situation as absolute error, but just squared it. For Negative Log Likelihood value, the smaller the better, so infinity negative is better than 909.

Maximum likelihood estimation optimizes likelihood function (no negative sign). The log function is monotonic and makes it easy to calculate. Some software do it in minimum that is why there a negative sign (-). Because a regression problem is defined to minimize sum of error square.

For this result, we can say that second model which used Boosted Decision Tree Algorithm is better than first used Bayesian Linear Regression Algorithm.

Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
Infinity	54790.387504	77334.407763	0.198714	0.051663	0.948337
914.027199	16723.019723	31575.335785	0.060651	0.008613	0.991387

Compared the third and fourth models, we can see the mean absolute error of fourth is less than the third, that means the error rate of third is bigger than fourth's, and the average difference of third model between predictive values and real values are greater than fourth's. Also, the RMS of fourth is smaller than third's, RMS stand for same situation as absolute error, but just squared it. For Negative Log Likelihood value, the smaller the better, so infinity negative is better than 914.

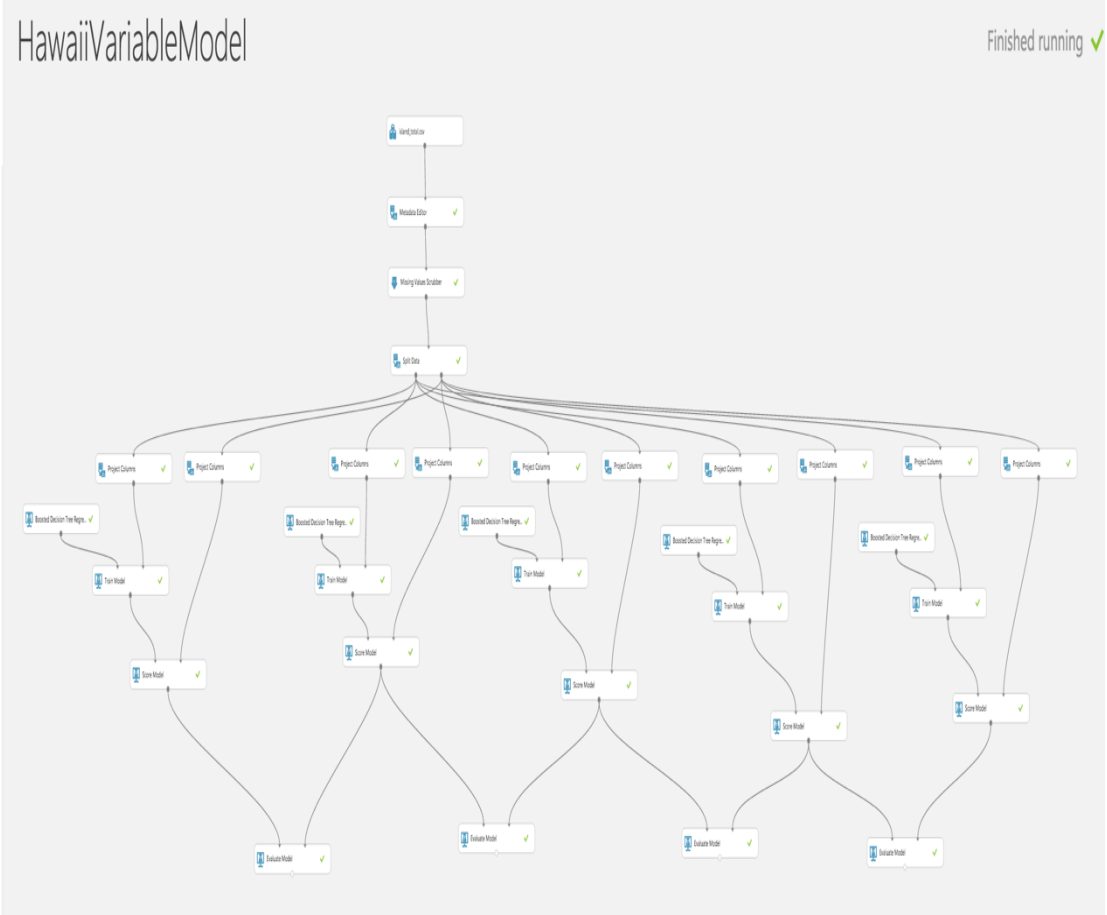
For this result, we can say that fourth model which used Decision Forest Regression Algorithm is better than third used Poisson Regression Algorithm though third model has a better Negative Log Likelihood.

The second and fourth algorithm has similar performance on Mean Absolute Error and RMS and scoring, but the second has a better Negative Log Likelihood. Thus, we choose second algorithm—Boosted Decision Tree Regression as our training model algorithm for all prediction models.

9. COMPARE & EVALUATE VARIABLE

Here we build prediction models with different variables in Azure ML. We choose dataset2 (monthly total visitors in different islands) to build four kinds of monthly total visitor prediction models with different variable subsets, and then compare models' performance and choose the best variable subsets for further prediction models.

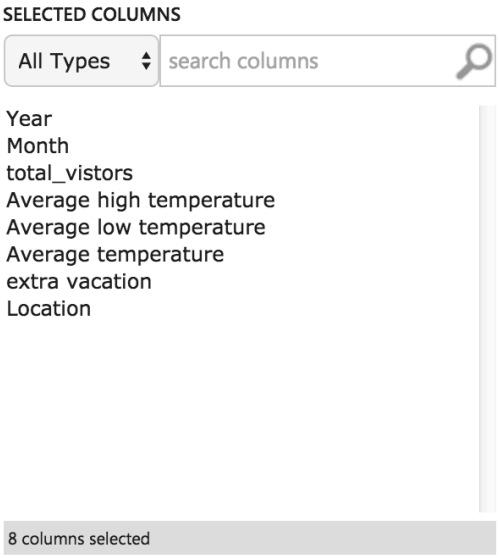
Below is the whole model, and we will explain it in details later.



9.1 MONTHLY TOTAL VISITOR PREDICTION MODELS WITH ALL VARIABLES

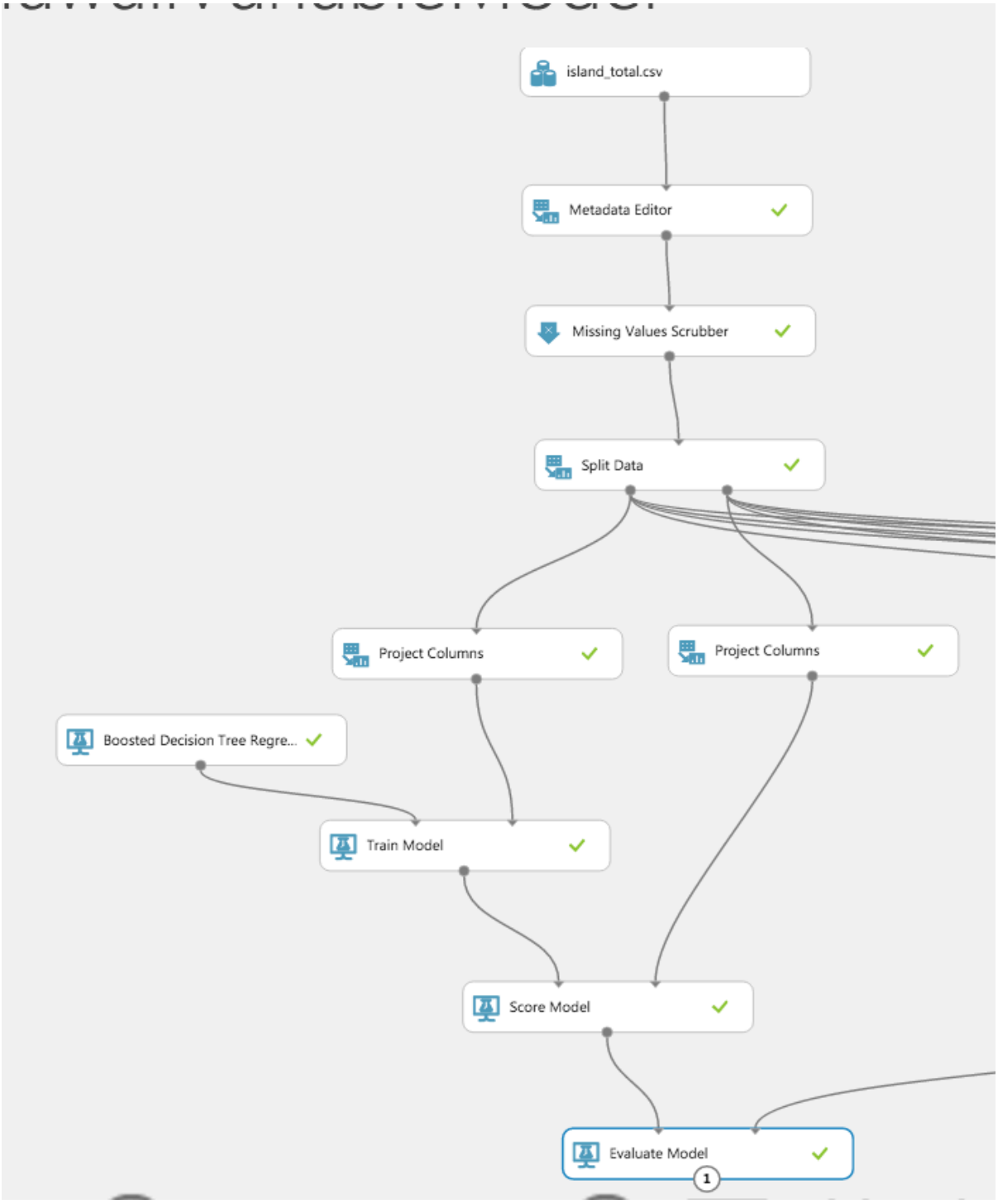
9.1.1 SELECT VARIABLES (PROJECT COLUMNS)

We use all 8 variables to build the prediction model, and the model performance is shown below.



9.1.2 BUILD MODEL

First we change Location column from String to Categorical type. And then we replace missing data with median, and divide data into training data (2007-2014) and validation data (2015). Here we use Boosted Decision Tree Regression Algorithms to train model. After that, we use validation dataset to evaluate model performance.


island_total.csv
Metadata Editor
Missing Values Scrubber
Split Data
Project Columns
Project Columns
Boosted Decision Tree Regre...
Train Model
Score Model
Evaluate Model
1

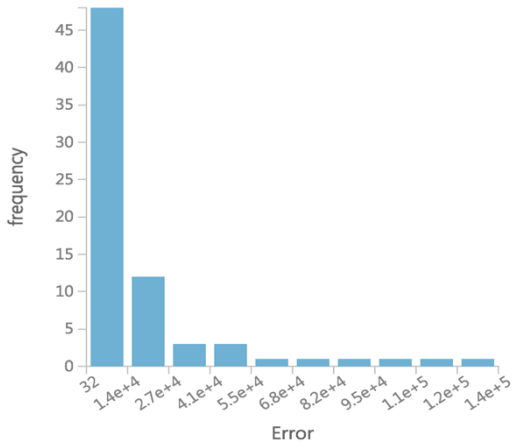
9.1.3 EVALUATE MODEL

As shown in picture below, the RMS of this prediction model is 32573, which means the predict value has less than +/-32573 error from real value.

Metrics

Mean Absolute Error	17420.976855
Root Mean Squared Error	32573.238088
Relative Absolute Error	0.063182
Relative Squared Error	0.009166
Coefficient of Determination	0.990834

Error Histogram



9.2 MONTHLY TOTAL VISITOR PREDICTION MODELS WITHOUT AVERAGE TEMPERATURE

9.2.1 SELECT VARIABLES (PROJECT COLUMNS)

We remove the average temperature variable to build the prediction model, and the model performance is shown below.

SELECTED COLUMNS

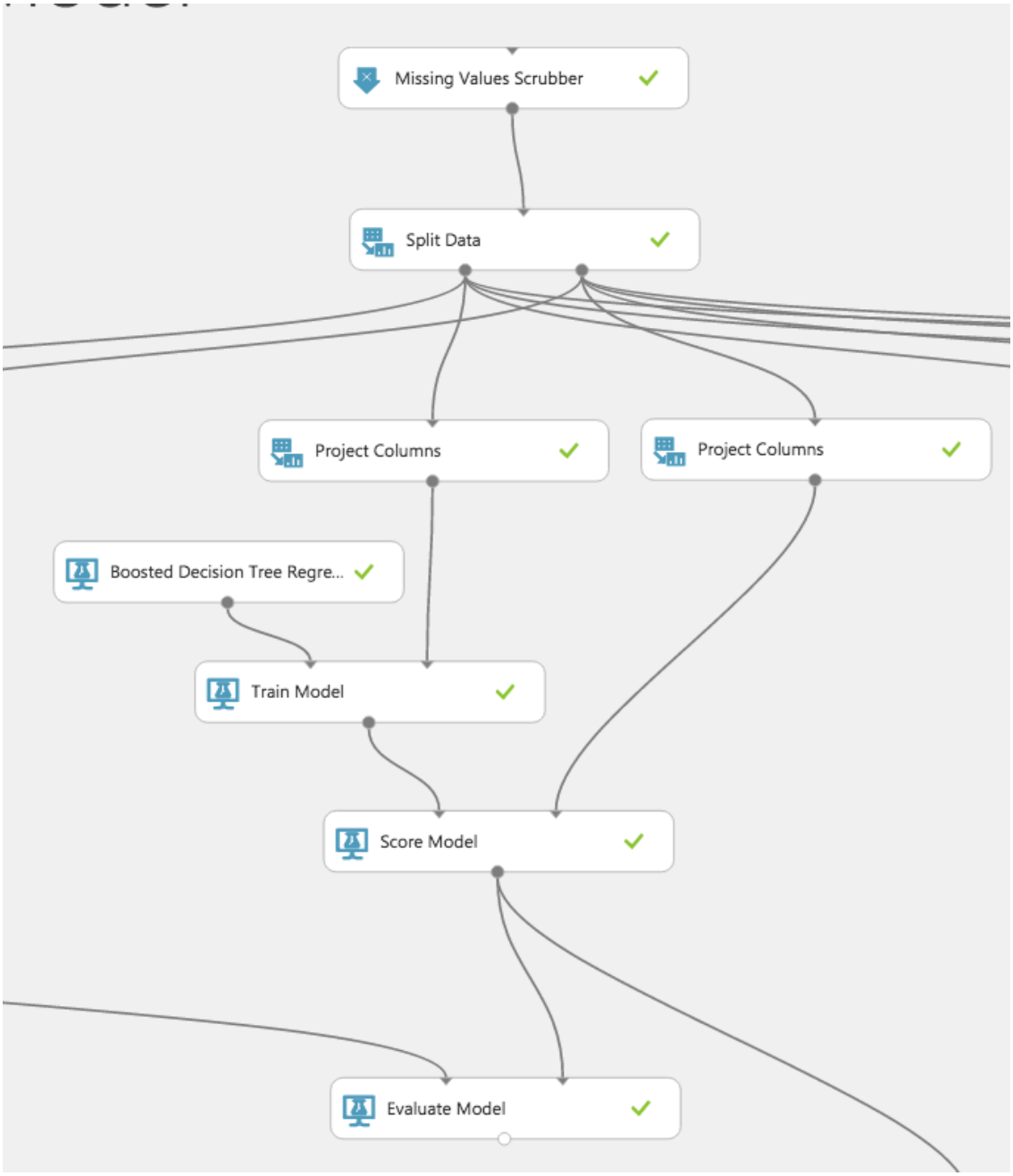
All Types | search columns

Year
Month
total_vistors
Average high temperature
Average low temperature
Location
extra vacation

7 columns selected

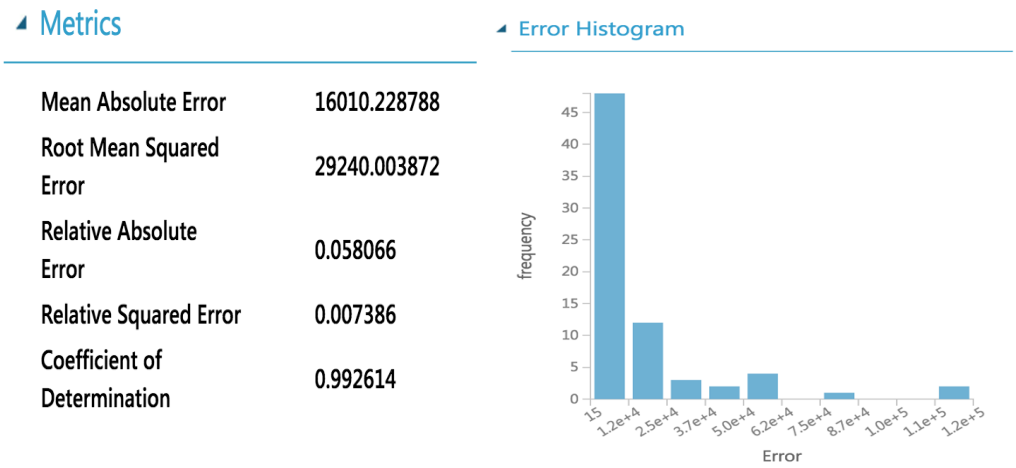
9.2.2 BUILD MODEL

First we change Location column from String to Categorical type. And then we replace missing data with median, and divide data into training data (2007-2014) and validation data (2015). Here we use Boosted Decision Tree Regression Algorithms to train model. After that, we use validation dataset to evaluate model performance.



9.2.3 EVALUATE MODEL

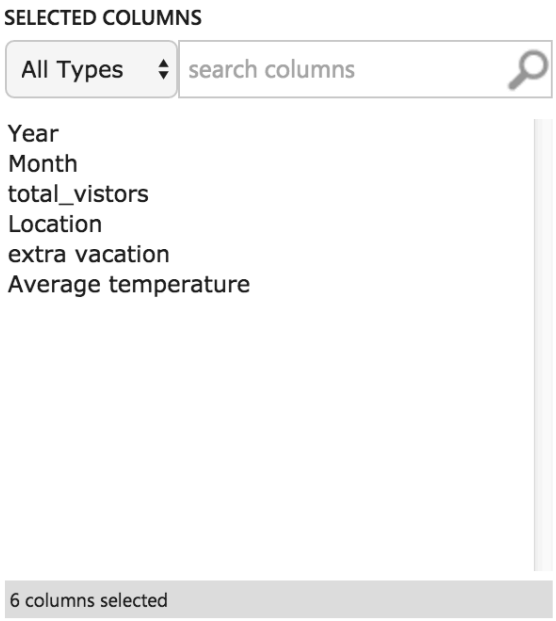
As shown in picture below, the RMS of this prediction model is 29240, which means the predict value has less than +/-29240 error from real value.



9.3 MONTHLY TOTAL VISITOR PREDICTION MODELS WITHOUT HIGH/LOW TEMPERATURE

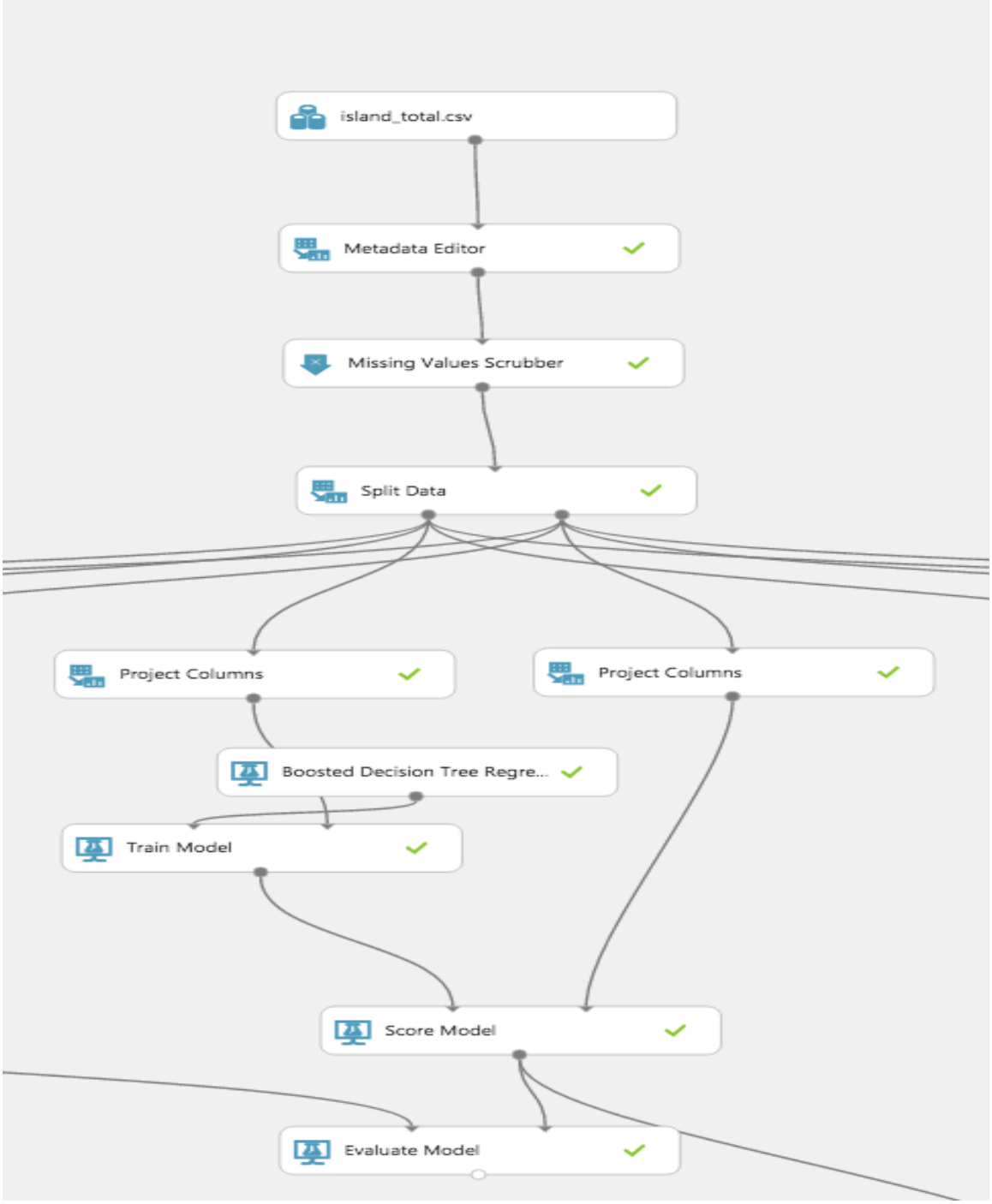
9.3.1 SELECT VARIABLES (PROJECT COLUMNS)

We remove the high and low temperature variables to build the prediction model, and the model performance is shown below.



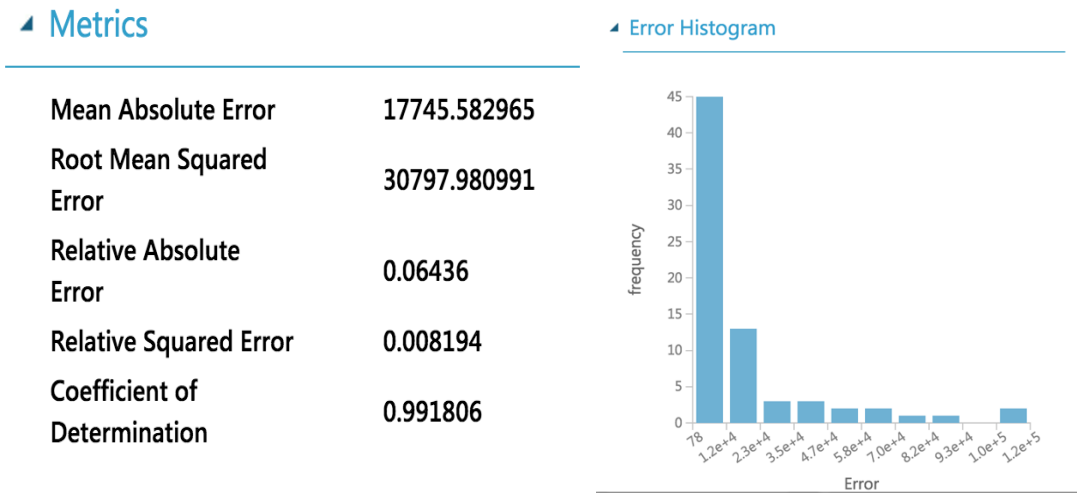
9.3.2 BUILD MODEL

First we change Location column from String to Categorical type. And then we replace missing data with median, and divide data into training data (2007-2014) and validation data (2015). Here we use Boosted Decision Tree Regression Algorithms to train model. After that, we use validation dataset to evaluate model performance.



9.3.3 EVALUATE MODEL

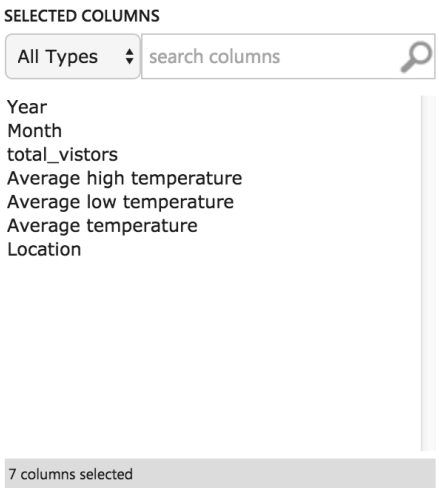
As shown in picture below, the RMS of this prediction model is 30797, which means the predict value has less than +/-30797 error from real value.



9.4 MONTHLY TOTAL VISITOR PREDICTION MODELS WITHOUT EXTRA VACATION DAYS

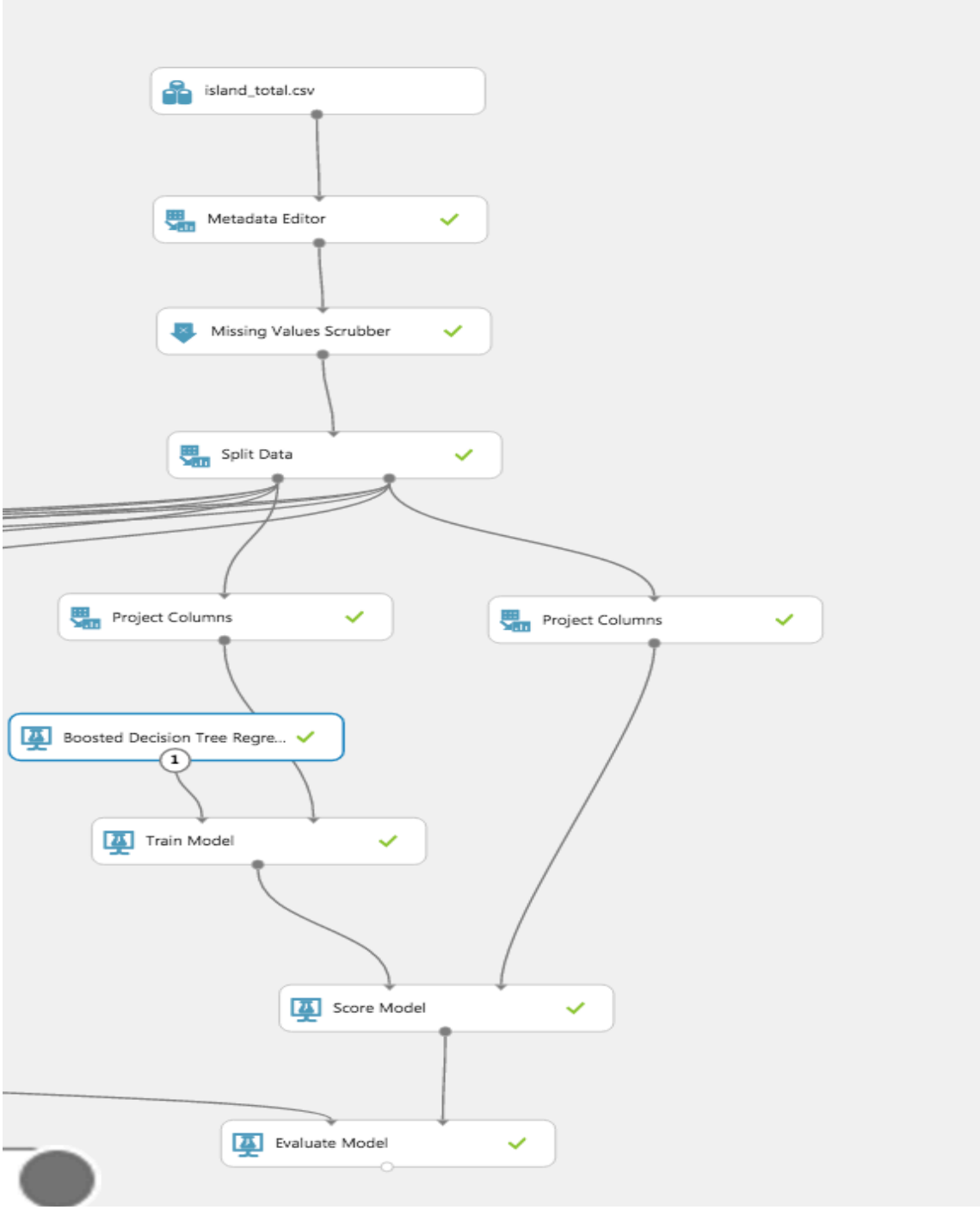
9.4.1 SELECT VARIABLES (PROJECT COLUMNS)

We remove the extra vacation day variable to build the prediction model, and the model performance is shown below.



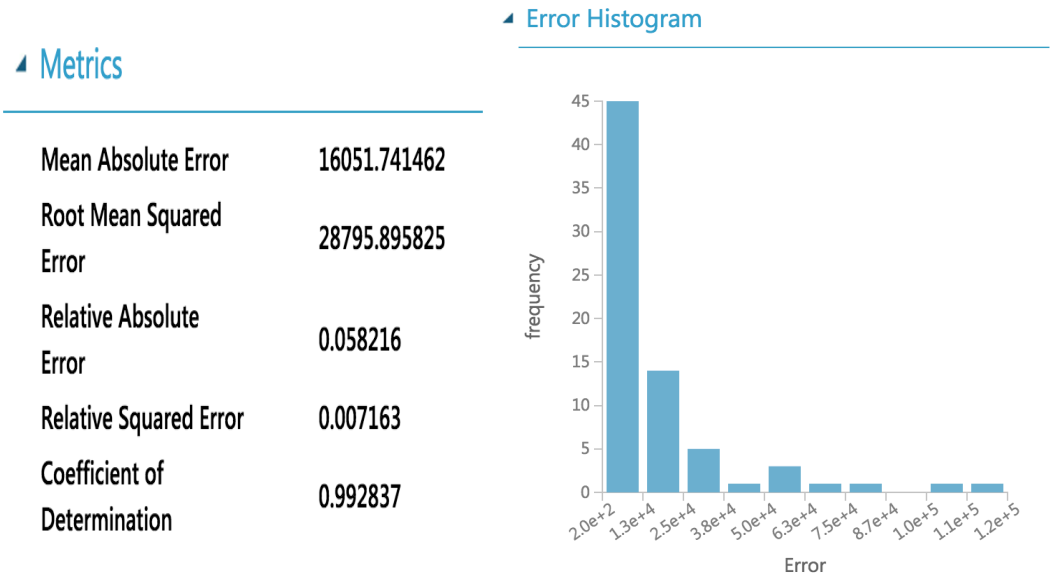
9.4.2 BUILD MODEL

First we change Location column from String to Categorical type. And then we replace missing data with median, and divide data into training data (2007-2014) and validation data (2015). Here we use Boosted Decision Tree Regression Algorithms to train model. After that, we use validation dataset to evaluate model performance.



9.4.3 EVALUATE MODEL

As shown in picture below, the RMS of this prediction model is 28795, which means the predict value has less than +/-28795 error from real value.



9.5 VARIABLE SELECTION

After comparing those models, we can see that the model without average temperature has smallest RMS, which means it has best performance. Therefore, we decide to we decide to remove extra vacation day variable.

Choose following variables to build prediction models: Year, Month, Average_high temperature, Average_low temperature, Average temperature, Location (island name).

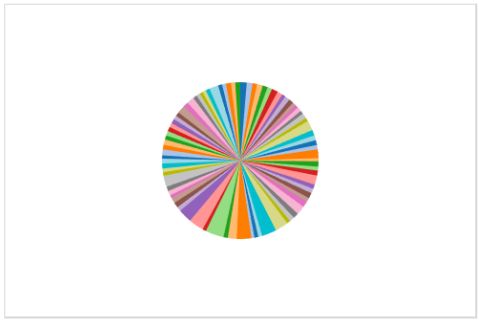
	Model1	Model2	Model3	Model4
RMS	32573	29240	30797	28795

10. TIME SERIES:

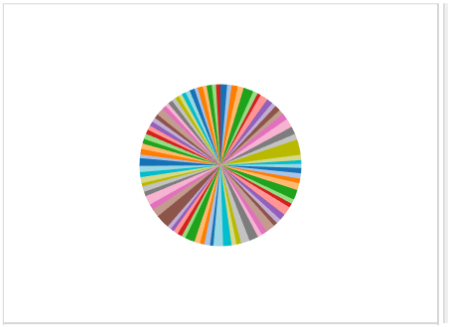
10.1 EXPLORE DATA BY TABLEAU

The following picture shows the relationship between the expenditure and three parameters relative the temperature, and the relationship between the expenditure. As you can see, all the picture is uniform distribution. In other words, expenditure is independence of those parameters.

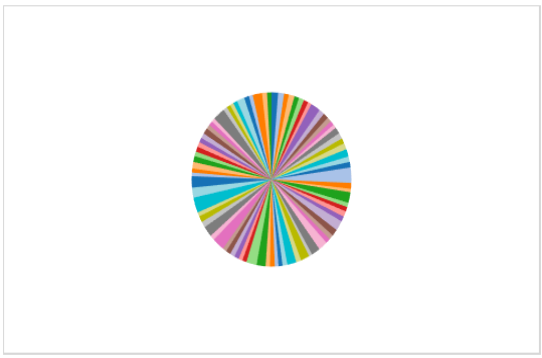
low temperature & expenditure



high temperature & expenditure



average temperature & expenditure



vacation & expenditure

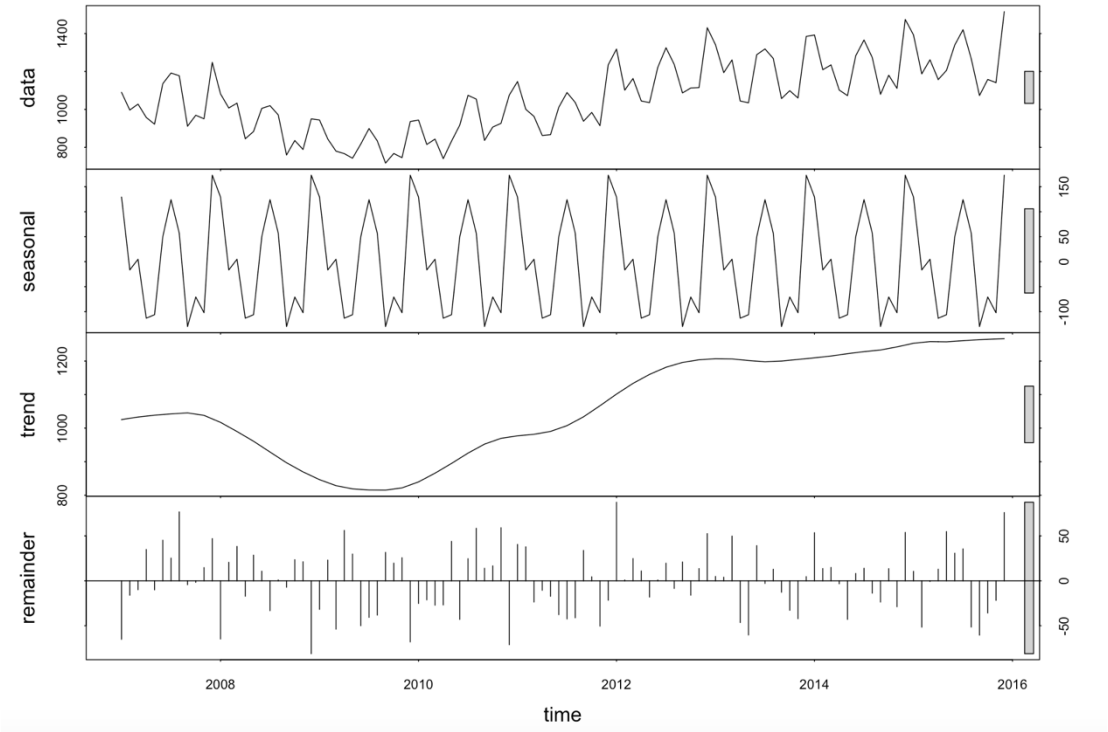
10.2 EXPLORE DATA BY R

(1) We import the data set into R, then transform data to time series, unit is month, start time point is January 2007. The following picture is the R code which we use to do that.

```
data1<-read.csv("~/Desktop/total_hawaii_expenditure.csv")
#transform data to time series, unit is month, start time point is jan 2007
tdata<- ts(data1[[3]],start=c(2007,1), frequency = 12)
```

(2) We use STL function to decompose season of time series by loess for observing the trend. The following picture are the R code and result.

```
#Seasonal Decomposition of Time Series by Loess
plot(stl(tdata,s.window="periodic"))
```



The first picture is the line chart of raw data set. The second picture is the seasonal trend which is periodic reversal. And the third picture is the trend of the data set. As you can see, the trend is mainly slow rise after a descend, And the last picture is the residual term.

To sum up, the expenditure shows clear season trend and be less influence by other variables. So we decide use the time series to building the model of expenditure prediction.

10.3 PARTITION DATA

We separate the data set into training dataset and testing dataset according to time. We use the data from 2007 to 2014 for building model, and use 2015 data to validate the result. The following picture is the R code which we use to separate the data set.

```
#2007-2014 as traindata, 2015 as valitation data
traindata<-window(tdata,start=2007,end=2014+11/12)
testdata<-window(tdata,start=2015)
```

10.4 COMPARE THE ALGORITHMS

we total use 9 different methods to build the time series model.

10.4.1 EXPONENTIAL SMOOTHING STATE SPACE MODEL

We use four different function - SES, holt, HW and ETS, to build the Exponential smoothing state space model. In fact, SES, holt and HW are simply convenient wrapper functions for forecast(ETS(...)). And for the fit1, since we not specify the model, R returns the best model automatically. The following picture is the R code and the note of accuracy is the RMS of the model.

```
pred_holt<-holt(traindata,h=12,damped=F,initial="simple",beta=0.65)
accuracy(pred_holt)#166
plot(pred_holt)

pred_ses <- ses(traindata,h=12,initial='simple',alpha=0.2)
accuracy(pred_ses)#123
plot(pred_ses)

pred_hw<-hw(traindata,h=12,seasonal='multiplicative')
accuracy(pred_hw)#42.8
plot(pred_hw)

fit1<-ets(traindata)
accuracy(predict(fit1,12),testdata) #42.43
plot(fit1)
```

The following picture is the accuracy of the model which build by ETS function.

```
> accuracy(predict(fit1,12),testdata) #42.43
                     ME     RMSE      MAE          MPE     MAPE      MASE
Training set   1.224696 42.43927 32.16605  0.08735142 3.139882 0.3239598
Test set     -48.025355 68.14964 57.95313 -3.92560033 4.756453 0.5836740
                    ACF1 Theil's U
Training set 0.004679037        NA
Test set     0.624958536 0.4194878
```

10.4.2 ARIMA

We use four different function-naive, Snaive, arima and auto.arima , to build the ARIMA model. Naive() returns forecasts and prediction intervals for an ARIMA(0,1,0) random walk model applied to x. Snaive() returns forecasts and prediction intervals from an ARIMA(0,0,0)(0,1,0)m model where m is the seasonal period. The different between the auto.arima function and the rest is the auto.arima returns best ARIMA model according to either AIC, AICc or BIC value. So the only argument auto,arima need is dataset. The following picture is the R code and the note beside accuracy is the RMS of the model.

```
pred_naive<-naive(traindata,h=12)
accuracy(pred_naive)#132

pred_snaive<-snaive(traindata,h=12)
accuracy(pred_snaive)#123
plot(pred_snaive)

fit2<-auto.arima(traindata)
accuracy(forecast(fit2,h=12),testdata) #45
plot(fit2)

ma = arima(traindata, order = c(0, 1, 3),   seasonal=list(order=c(0,1,3), period=12))
p<-predict(ma,12)
accuracy(p$pred,testdata)  #48
```

The following picture is the accuracy of the model which build by auto.arima function.

```
> accuracy(forecast(fit2,h=12),testdata) #45
                    ME     RMSE      MAE        MPE     MAPE      MASE
Training set   2.510318 45.80851 34.83629  0.1975113 3.498749 0.3508531
Test set     -52.219994 67.69040 59.87994 -4.3156069 4.950865 0.6030799
                   ACF1 Theil's U
Training set -0.009153377        NA
Test set      0.358379796 0.4143118
```

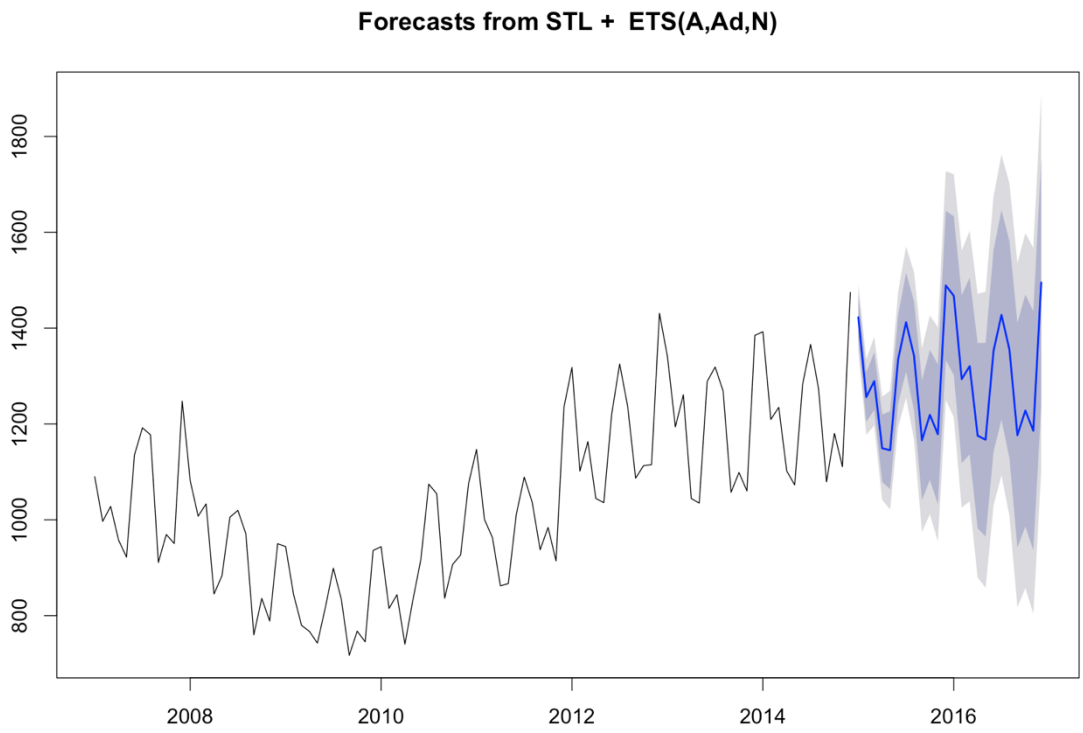
10.4.3 STLF

The ets models is a better choose if the data are non-seasonal or the seasonal period is 12 or less and if the seasonal period is 13 or more stlf is a better option. Stlf combines STL decomposition and ETS model. It takes a ts argument, applies an STL decomposition, models the seasonally adjusted data, reseasonalizes, and returns the forecasts. The following picture is the R code and the accuracy of the model.

```
#stl+ets(AAN)
pred_stlf<-stlf(traindata)
accuracy(pred_stlf)#34.99
plot(pred_stlf)
```

```
> accuracy(pred_stlf)#34.99
                    ME     RMSE      MAE        MPE     MAPE      MASE
Training set 2.014956 34.99689 28.13921 0.1217105 2.798023 0.2834036
                    ACF1
Training set 0.002465656
```

The following picture show the prediction result of STLF. The deep gray area represents the 80% forecast period, and the light gray area represents the 95% forecast period.



10. 5 CONCLUSION

The fist picture is the accuracy of model fit1 which is built by ETS function, the second picture is the accuracy of model fit2 which is built by Arima function, the third picture is the accuracy of model pred_stlf which is built by STLF function. We choose STLF function to build the time series model. As you can see, the third model has lower ME, RMSE, MAE, MAPE and MASE than ARIMA model. What's more, compare to ETS functions, STLF function can avoid seasonality being ignored. So we choose STLF function to build the time series model.

```
> accuracy(predict(fit1,12),testdata) #42.43
                     ME     RMSE      MAE          MPE     MAPE      MASE
Training set   1.224696 42.43927 32.16605  0.08735142 3.139882 0.3239598
Test set     -48.025355 68.14964 57.95313 -3.92560033 4.756453 0.5836740
                   ACF1 Theil's U
Training set 0.004679037        NA
Test set     0.624958536 0.4194878
```

```
> accuracy(forecast(fit2,h=12),testdata) #45
                    ME      RMSE      MAE         MPE     MAPE      MASE
Training set   2.510318 45.80851 34.83629  0.1975113 3.498749 0.3508531
Test set     -52.219994 67.69040 59.87994 -4.3156069 4.950865 0.6030799
                   ACF1 Theil's U
Training set -0.009153377        NA
Test set      0.358379796 0.4143118
```

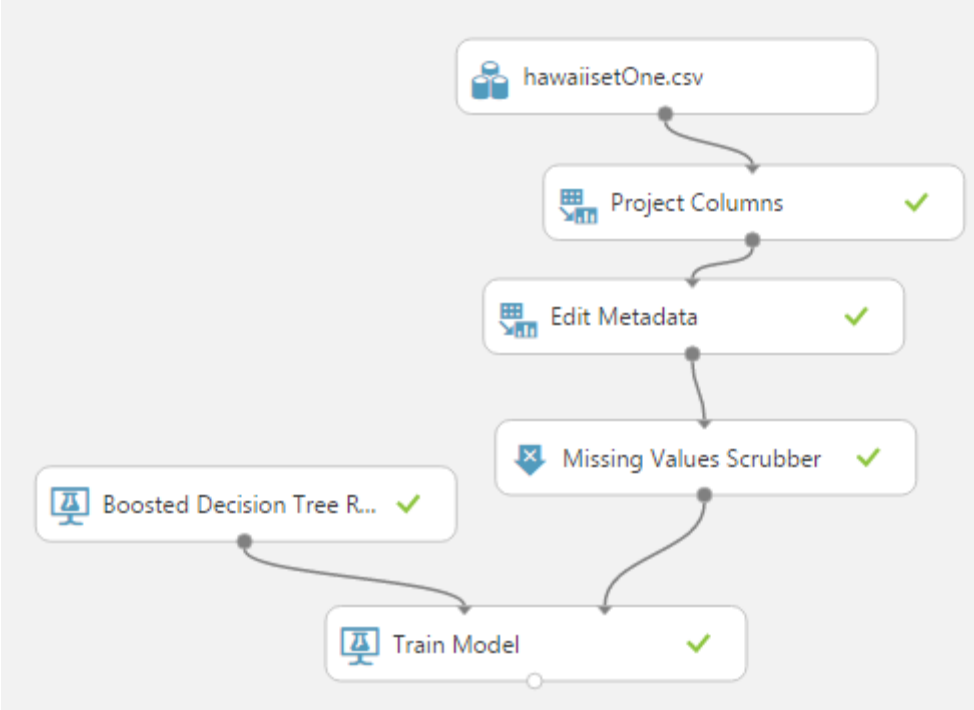
```
> accuracy(pred_stlf)#34.99
                    ME      RMSE      MAE        MPE     MAPE      MASE
Training set 2.014956 34.99689 28.13921 0.1217105 2.798023 0.2834036
                   ACF1
Training set 0.002465656
```

11. AZURE MACHINE LEARNING

11.1 MODEL ONE (PREDICT VISITOR AMOUNT OF EACH ISLAND FROM EACH COUNTRY)

11.1.1 MODELS ONE OVERVIEW

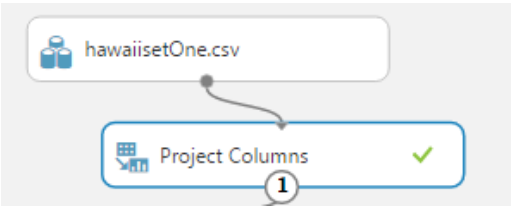
We post our model screen shot here, and in later several chapters, we will descript this model step by step very details.



11.1.2 IMPORTING DATASET AND PRE-PROCESS

We import the dataset into Azure Machine Learning Studio, as you can see in the under picture, there still is Extra Vacation Feature. In our Selecting Variables Part, we have decided remove this feature because it has little effect on our prediction result. Thus, we used Project Column to remove this column.

Year	Month	Visitors	Average high temperature	Average low temperature	Average temperature	extra vacation	Island	Country
2007	1	115535.638	80.9	68.8	74.85	2	Hawaiilsland	Canada
2007	2	100557.6995	80.2	66.7	73.45	1	Hawaiilsland	Canada
2007	3	95819.4767	80.5	67.9	74.2	0	Hawaiilsland	Canada
2007	4	52189.87305	83.6	69.7	76.65	1	Hawaiilsland	Canada



Project Columns

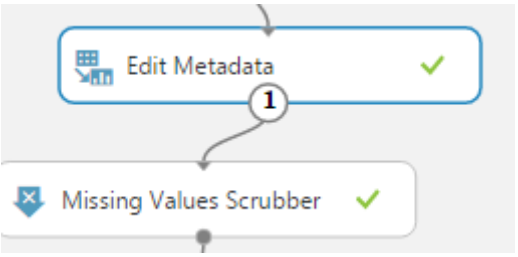
Select columns

Selected columns:
Column names: Year,Month,Visitors,Average high temperature,Average low temperature,Average temperature,Island,Country

Launch column selector

You can see in the right picture, all Selected Column includes Year, Month, Visitors, High Temperature, Low Temperature, Average Temperature, Island and Country, but except Extra Vacation.

11.1.3 EDIT METADATA



Edit Metadata

Column

Selected columns:
Column names: Island,Country

Launch column selector

Data type

Unchanged

Categorical

Make categorical

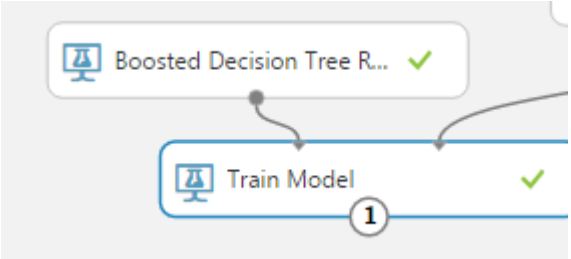
Fields

Unchanged

Because the Island and Country variables in our dataset are String type, we need make them categorical if we want them have a influence on the final result. Thus, we used Edit Metadata Module, choose Island and Country variable, and then made them categorical.

Before building the model, we need deal with the missing data. We used Missing Values Scrubber to remove the entire row which contains Missing data.

11.1.4 BUILDING TRAINING MODEL

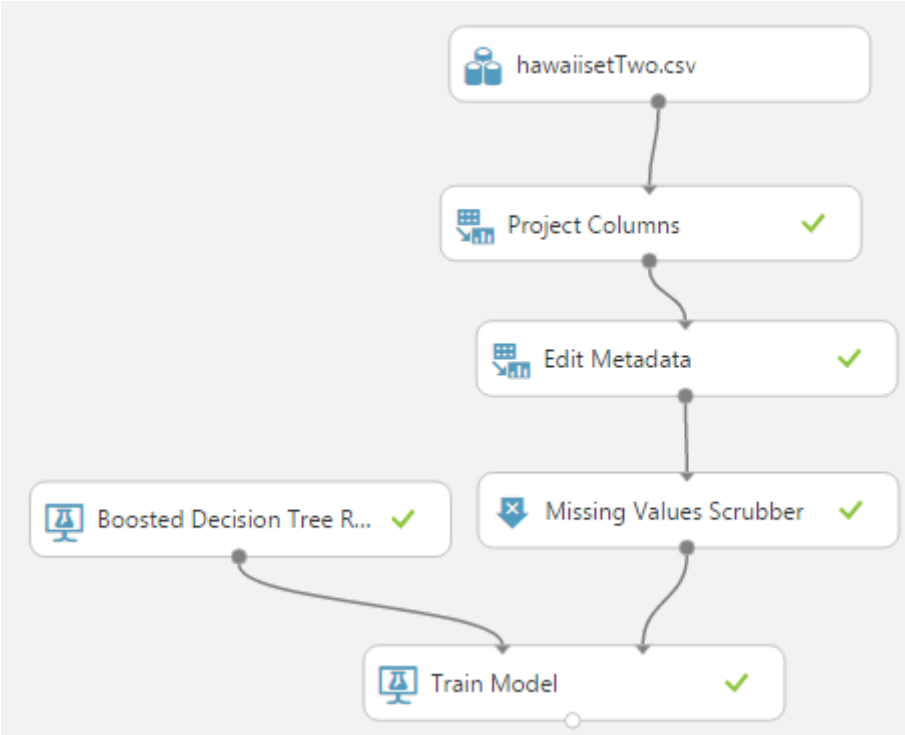


We have already compared four algorithms' performance in our Comparing Algorithm Part. And we though the most suitable algorithm for our models is Boosted Decision Tree. And we have already validated it in Comparing Algorithm Part, so we use all records instead of splitting it into two dataset. Besides, we have scored and evaluated it in Comparing part, so in this part, we directly build model.

11.2 MODEL TWO (PREDICT VISITOR AMOUNT OF EACH ISLAND)

11.2.1 MODELS TWO OVERVIEW

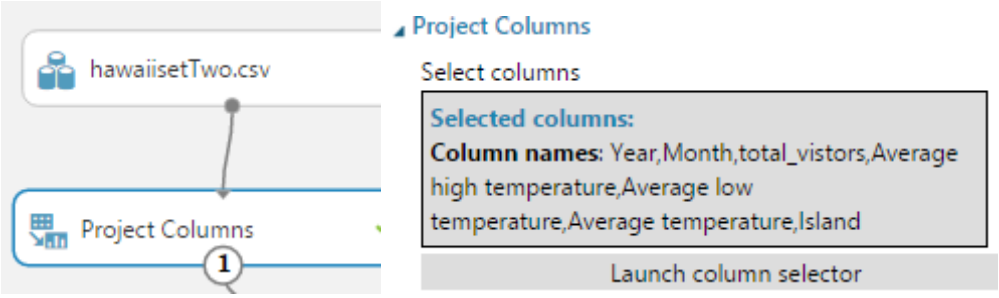
We post our model screen shot here, and in later several chapters, we will descript this model step by step very details.



11.2.2 IMPORTING DATASET AND PRE-PROCESS

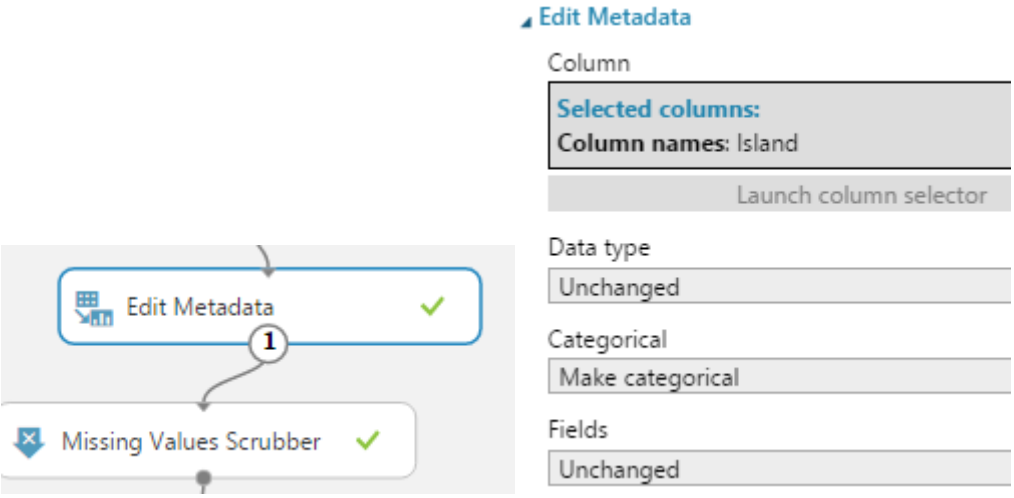
This Model is similar with the first model. At the first, we import the dataset into Azure Machine Learning Studio, as you can see in the under picture, there still is Extra Vacation Feature. There is no Country variable because this model needs not that variable. This model is used for predicting the total visitor who from multiple country amount for each island. In our Selecting Variables Part, we decided remove Extra Vacation feature because it has little effect on our prediction result. Thus, we used Project Column to remove this column.

Year	Month	total_vistors	Average high temperature	Average low temperature	Average temperature	extra vacation	Island
2007	1	195264.6077	80.9	68.8	74.85	2	Maui
2007	2	196700.1198	80.2	66.7	73.45	1	Maui
2007	3	227232.5152	80.5	67.9	74.2	0	Maui
2007	4	202215.7726	83.6	69.7	76.65	1	Maui
2007	5	198130.1536	85	71.6	78.3	0	Maui



You can see in the right picture, all Selected Column includes Year, Month, total_visitors, High Temperature, Low Temperature, Average Temperature, and Island, but except Extra Vacation.

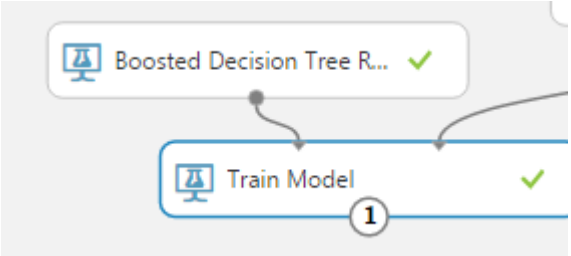
11.2.3 EDIT METADATA



Because the Island variable in our dataset is String type, we need make it categorical if we want it have an influence on the final result. Thus, we used Edit Metadata Module, choose Island variable, and then made it categorical.

Before Building the model, we need deal with the missing data. We used Missing Values Scrubber to remove the entire row which contains Missing data.

11.2.4 BUILDING TRAINING MODEL

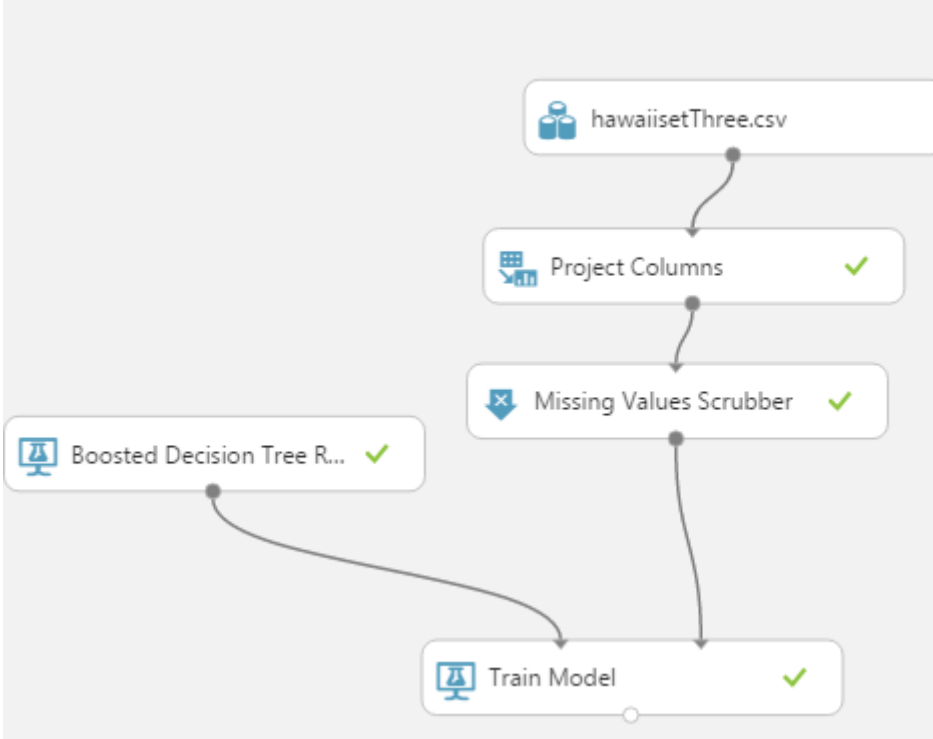


We have already compared four algorithms' performance in our Comparing Algorithm Part. And we though the most suitable algorithm for our models is Boosted Decision Tree. And we have already validated it in Comparing Algorithm Part, so we use all records instead of splitting it into two dataset. Besides, we have scored and evaluated it in Comparing part, so in this part, we directly build model.

11.3 MODEL THREE (PREDICT TOTAL VISITOR AMOUNT OF HAWAII)

11.3.1 MODELS THREE OVERVIEW

We post our model screen shot here, and in later several chapters, we will descript this model step by step very details.

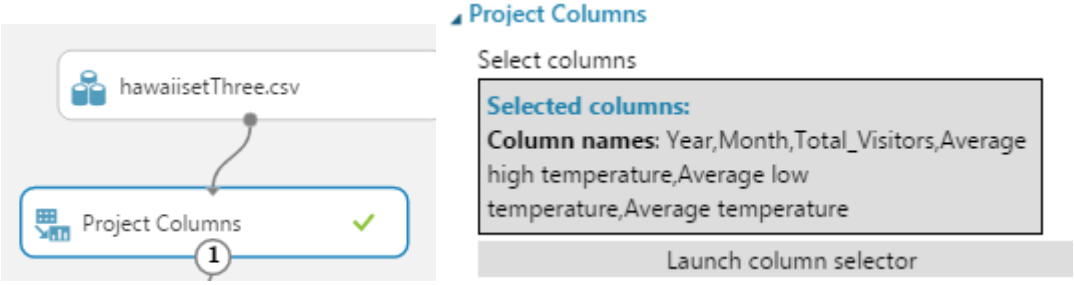


11.3.2 IMPORTING DATASET AND PRE-PROCESS

This Model is similar with the first and second model. At the first, we import the dataset into Azure Machine Learning Studio, as you can see in the under picture, there still is Extra Vacation Feature and expenditure variable (for model four). This model is used for predicting the total visitor of the entire Hawaii, so we do not use Expenditure, Island, and Country variables. In our Selecting Variables Part, we decided remove Extra

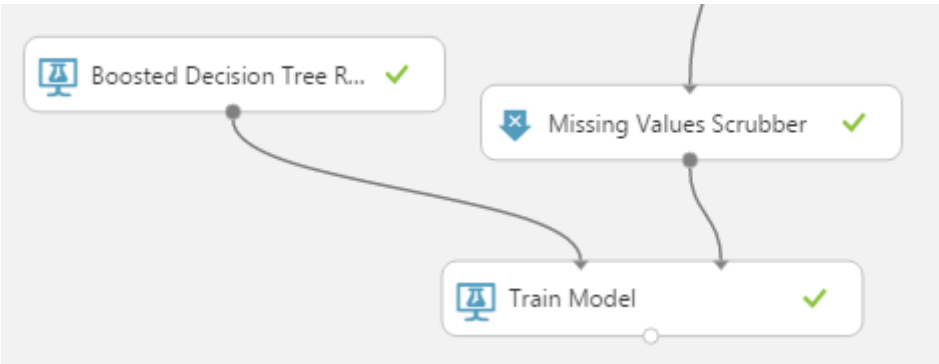
Vacation feature because it has little effect on our prediction result. Thus, we used Project Column to remove this column.

Year	Month	Total_Visitors	expenditure	Average high temperature	Average low temperature	Average temperature	extra vacation
2007	1	577231.7925	1089.873966	80.9	68.8	74.85	2
2007	2	574762.7083	996.76546	80.2	66.7	73.45	1
2007	3	674532.0081	1028.064539	80.5	67.9	74.2	0
2007	4	597477.5602	957.542314	83.6	69.7	76.65	1



You can see in the right picture, all Selected Column includes Year, Month, total_visitors, High Temperature, Low Temperature, Average Temperature, but except Extra Vacation and Expenditure.

11.3.3 BUILDING TRAINING MODEL



Because there is no String type variable, we do not edit metadata and make it categorical. But, before Building the model, we need deal with the missing data. We used Missing Values Scrubber to remove the entire row which contains Missing data.

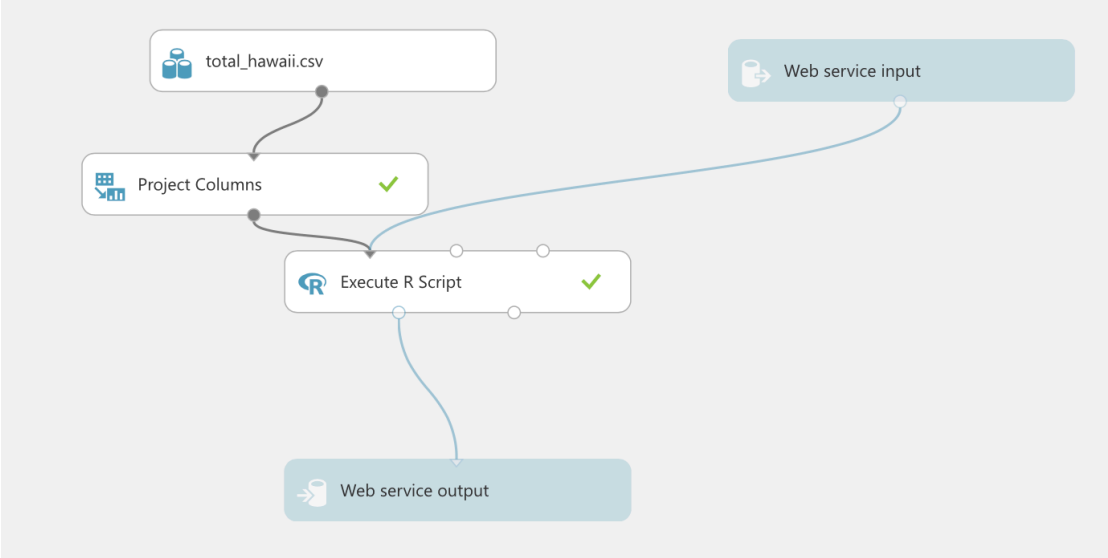
We have already compared four algorithms' performance in our Comparing Algorithm Part. And we though the most suitable algorithm for our models is Boosted Decision Tree. And we have already validated it in Comparing Algorithm Part, so we use all

records instead of splitting it into two dataset. Besides, we have scored and evaluated it in Comparing part, so in this part, we directly build model.

11.4 MODEL FOUR (PREDICT TOTAL EXPENDITURE OF HAWAII)

11.4.1 MODELS THREE OVERVIEW

(1) We post our model screen shot here, and in later several chapters, we will descript this model step by step very details.



11.4.2 IMPORTING DATASET AND PRE-PROCESS

At the first, we import the dataset into Azure Machine Learning Studio, as you can see in the under picture. This model is used for predicting the total expenditure of the entire Hawaii. Building the time series only need the time as input, so we use project columns to delete other columns.

Year	Month	Total_Visitors	expenditure	Average high temperature	Average low temperature	Average temperature	extra vacation
2007	1	577231.7925	1089.873966	80.9	68.8	74.85	2
2007	2	574762.7083	996.76546	80.2	66.7	73.45	1
2007	3	674532.0081	1028.064539	80.5	67.9	74.2	0
2007	4	597477.5602	957.542314	83.6	69.7	76.65	1

Project Columns

Select columns

Selected columns:
Column names:
Year,Month,expenditure

Launch column selector

START TIME	4/28/2016 2:56:41 PM
END TIME	4/28/2016 2:56:41 PM
ELAPSED TIME	0:00:00.000
STATUS CODE	Finished
STATUS DETAILS	Task output was present in output cache

11.4.2 BUILDING THE TIME SERIES MODEL AND OUTPUT PREDICT RESULT

Then we use the Execute R Script to build the time series model and out put the result. We use all the data and STLF function to build the time series model and the output of this step is the predict result. We use the mean as the forecast value and the Lo95 and Hi95 is the boundary of the 95% forecast period.

```
# Map 1-based optional input ports to variables
data1 <- maml.mapInputPort(1) # class: data.frame
library(forecast)
#transform data to time series
tdata<- ts(data1[[3]],start=c(2007,1), frequency = 12)
#use all data to building model
pred_stlf<-stlf(tdata)
#predict the mean and 95% forecast period.
Forecast <- pred_stlf$mean
Lo95 <- pred_stlf$lower[,1]
Hi95 <- pred_stlf$upper[,1]
# Select data.frame to be sent to the output Dataset port
output<-data.frame(cbind(Forecast, Hi95, Lo95),Month=1:12,Year=20
maml.mapOutputPort("output");
```

The out put of the model is the predict result for the next two years, and the Hi95 and Lo95 represents the boundary of 95% forecast period.

rows 24 columns 5

view as 

Forecast	Hi95	Lo95	Month	Year
1429.025321	1474.506843	1383.543798	1	2016
1248.16645	1300.850846	1195.482054	2	2016
1296.184481	1357.993297	1234.375664	3	2016
1156.996789	1229.07144	1084.922137	4	2016
1159.435679	1242.36237	1076.508989	5	2016
1346.297722	1440.306006	1252.289438	6	2016
1419.317028	1524.414627	1314.21943	7	2016
1324.675684	1440.734174	1208.617193	8	2016
1139.452167	1266.261048	1012.643287	9	2016
1202.3756	1339.676736	1065.074464	10	2016
1164.448152	1311.957992	1016.938311	11	2016
1503.50419	1660.928214	1346.080167	12	2016
1456.85083	1623.892951	1289.808709	1	2017
1271.238436	1447.607085	1094.869787	2	2017
1315.315003	1500.726972	1129.903033	3	2017
1172.85918	1367.04197	978.676391	4	2017
1172.588246	1375.281373	969.895119	5	2017
1357.203392	1568.15901	1146.247775	6	2017
1428.359647	1647.342686	1209.376608	7	2017
1332.173522	1558.961522	1105.385522	8	2017
1145.669125	1380.051853	911.286397	9	2017
1207.530495	1449.30944	965.751549	10	2017
1168.722419	1417.710212	919.734625	11	2017
1507.04827	1763.068063	1251.028476	12	2017

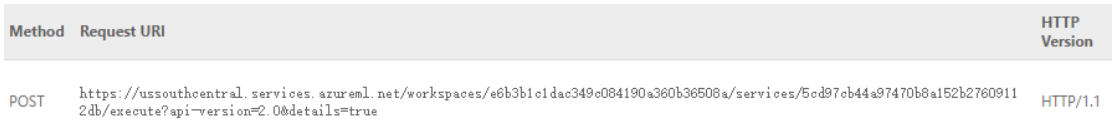
12. DEPLOY WEB SERVICE & CONFIGURATION

12.1 DEPLOY WEB SERVICE

After creating three training models, Azure Machine Leaning Workspace can create three predictive models automatically. These three predictive models are similar with that three training models we created before.

After predictive models were created, we deployed Web Service in Azure to get models' APIs and URIs. These two values are very important for developing an integration of model. You can see the details of APIs and URI in under picture.

API stands for application programming interface. It can be helpful to think of the API as a way for different apps to talk to one another. For many users, the main interaction with the API will be through API keys, which allow other apps to access your account without you giving out your password.



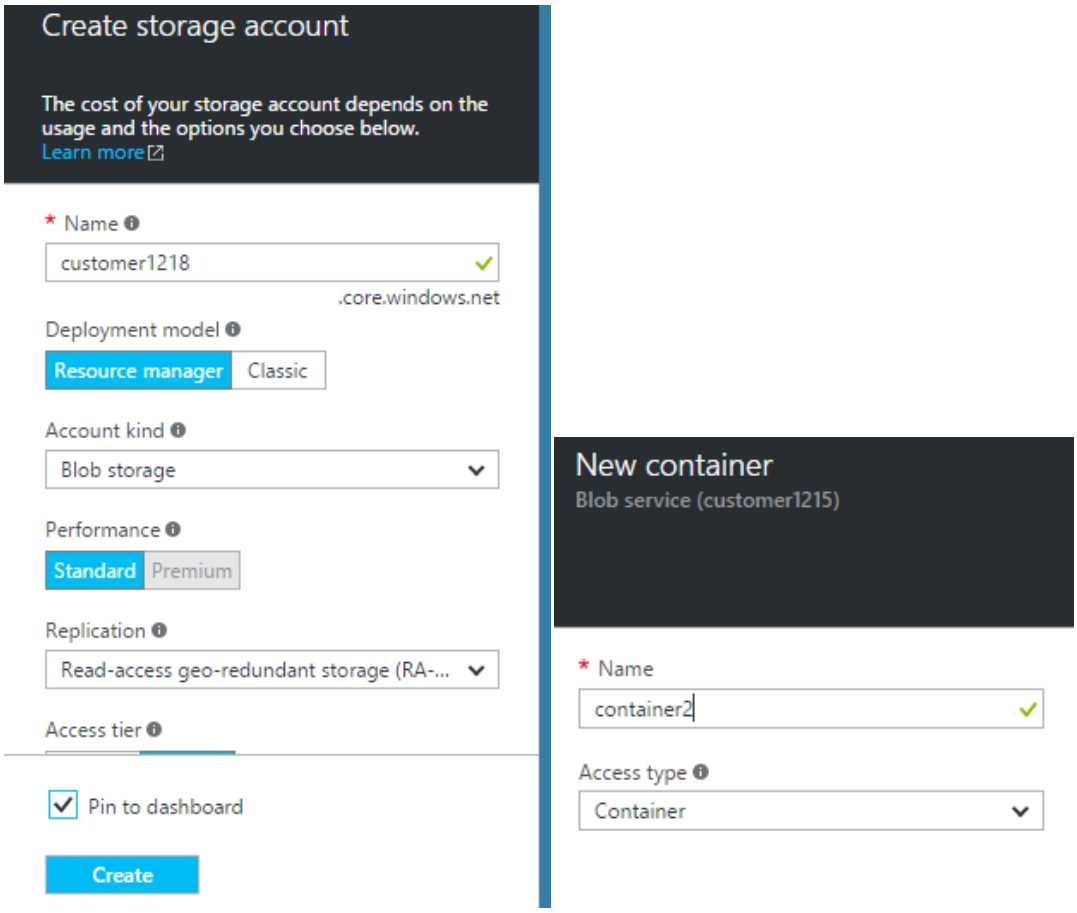
To paraphrase the World Wide Web Consortium, Internet space is inhabited by many points of content. A URI (Uniform Resource Identifier; pronounced YEW-AHR-EYE) is the way you identify any of those points of content, whether it be a page of text, a video or sound clip, a still or animated image, or a program. The most common form of URI is the Web page address, which is a particular form or subset of URI called a Uniform Resource Locator (URL).

12.2 BLOB STORAGE

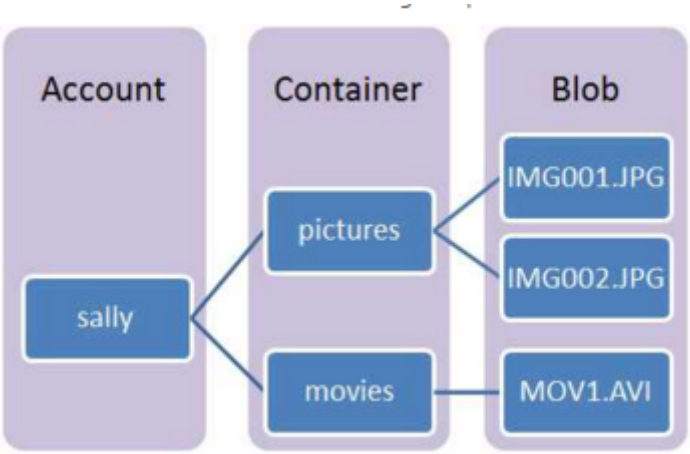
We assume uses who want to use our webpage to predict visitor amount should have an Azure Storage account. In this way, they can used their own dataset as the input of the prediction models. Also, our prediction model can store the prediction results to their own storage account. That improve the information security of this prediction model, and it can guarantee customers confidentiality.

12.2.1 STORAGE ACCOUNT

Now, we will show the process to create a storage account in Azure. After Logging in the Azure account, click Create Storage Account, and give the basic information for this new storage account.



After creating an account, we can create a container inside this storage account. Each file we upload to storage account is a blob, and it must be stored inside a container. The relationship among storage account, container, and blob will be shown below.



Storage Account: All access to Azure Storage is done through a storage account. This storage account can be a **General Purpose Storage Account** or a **Blob Storage Account** which is specialized for storing objects/blobs.

Container: A container provides a grouping of a set of blobs. All blobs must be in a container. An account can contain an unlimited number of containers. A container can store an unlimited number of blobs. Note that the container name must be lowercase.
Blob: A file of any type and size. Azure Storage offers three types of blobs: block blobs, page blobs, and append blobs.

12.2.2 UPLOAD FILES

Azure Storage Account is similar with GitHub, we cannot directly upload files to this account. We must use the third party tools like Azure Powershell or .Net studio to upload files to Blob Storage Account.

Now, we will show the process of uploading files to storage account using Azure Powershell. All screen shot came from Powershell Command Window.

1. Log-in to Azure

```
PS C:\Users\CANDICEHO> Login-AzureRmAccount


Environment           : AzureCloud
Account               : candiceho1215@gmail.com
TenantId              : 96ca9cb2-8460-4c50-8b45-8facc9c832cc
SubscriptionId        : eeba4fbe-f754-4282-aa24-12442c1211b5
CurrentStorageAccount :
```

A pop-up log-in window will show, and customer should log in to their Azure account.

2. Check Azure Subscription

```
PS C:\Users\CANDICEHO> Get-AzureRmSubscription


SubscriptionName : Free Trial
SubscriptionId   : eeba4fbe-f754-4282-aa24-12442c1211b5
TenantId         : 96ca9cb2-8460-4c50-8b45-8facc9c832cc
State            : Enabled
```

It will give you the subscription information about you Azure account, such as Subscription Name, Id and State.

3. Check Azure Context

```
PS C:\Users\CANDICEHO> Get-AzureRmContext


Environment           : AzureCloud
Account               : candiceho1215@gmail.com
TenantId              : 96ca9cb2-8460-4c50-8b45-8facc9c832cc
SubscriptionId        : eeba4fbe-f754-4282-aa24-12442c1211b5
CurrentStorageAccount :
```

It will tell you where you are, and which Azure account you are connecting to. But we have not yet set which Storage Account you want to connect, so that line is empty.

4. Set Storage Account

```
PS C:\Users\CANDICEHO> Set-AzureRmCurrentStorageAccount -ResourceGroupName “test” -StorageAccountName “customer1215”
customer1215
```

We should give the Storage Account name, and the Group your account belongs to.

Now, you can see the Current Storage Account is “customer1215”.

```
PS C:\Users\CANDICEHO> Get-AzureRmContext


Environment           : AzureCloud
Account               : candiceho1215@gmail.com
TenantId              : 96ca9cb2-8460-4c50-8b45-8facc9c832cc
SubscriptionId        : eeba4fbe-f754-4282-aa24-12442c1211b5
CurrentStorageAccount : customer1215
```

5. Set Account Parameters

```
PS C:\Users\CANDICEHO> $storageAccountName = "customer1215"
PS C:\Users\CANDICEHO> $containerName = "container1"
PS C:\Users\CANDICEHO> $storageAccountKey = "6lcGoWsDu9wTEZki0RsNHAV3fttcQwcRYKxBE7pBTvZ26T5z8N5Y9fnNGHKFXqGW8qu4smyPK+0
OcAAYJ9w4Zw=="
PS C:\Users\CANDICEHO> $blobContext = New-AzureStorageContext -StorageAccountName $storageAccountName -StorageAccountKey
 $storageAccountKey
```

You can create several parameters, such as “$accountName”, “$containerName”, “$storage AccessKey”, and “blobContext”. It is convenient to give a uploading files command.

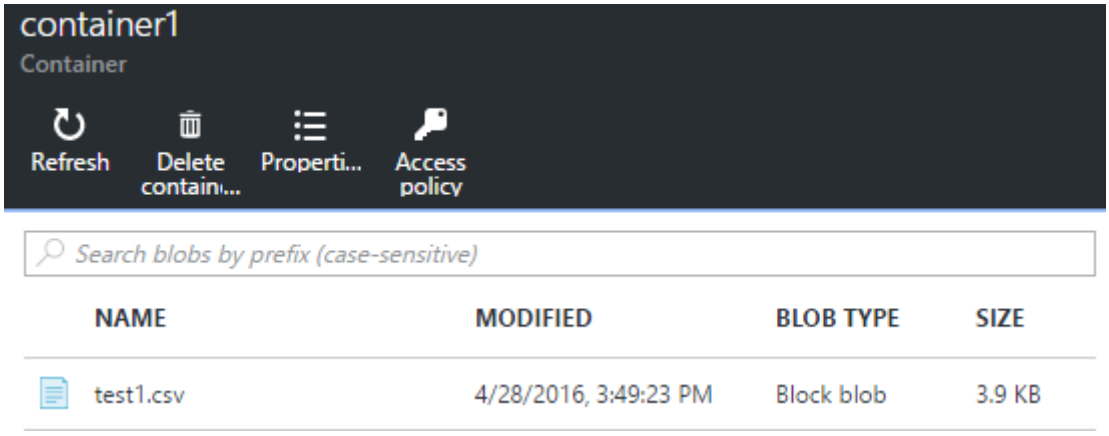
6. Upload Files

```
PS C:\Users\CANDICEHO>  Set-AzureStorageBlobContent -File d:\data\test1.csv -Container $containerName -Context $blobCont
ext -Force


ICloudBlob        : Microsoft.WindowsAzure.Storage.Blob.CloudBlockBlob
BlobType          : BlockBlob
Length            : 3992
ContentType       : application/octet-stream
LastModified      : 2016/4/28 22:49:23 +00:00
SnapshotTime      :
ContinuationToken :
Context           : Microsoft.WindowsAzure.Commands.Common.Storage.AzureStorageContext
Name              : test1.csv
```

Use parameters created before, and give the address of file which you want to upload.

7. Check in Azure Storage Account



You can see, there is a new file in your container which came from your local drive.

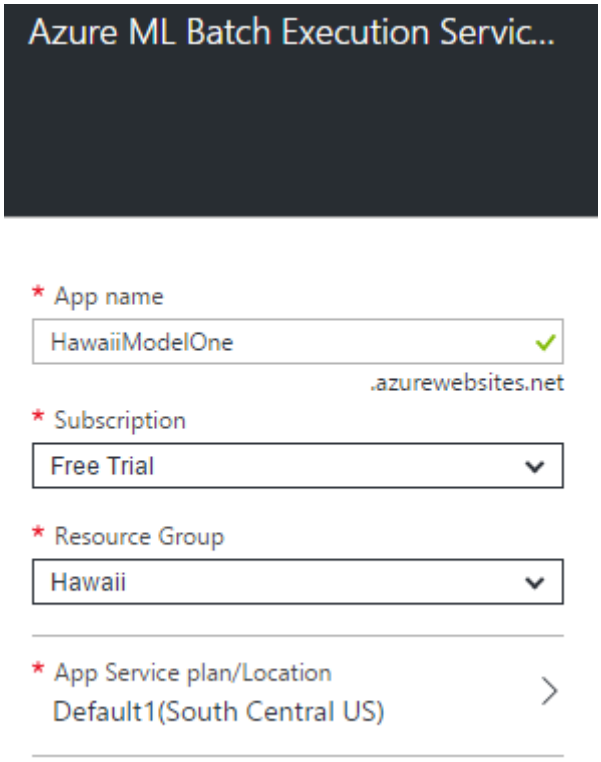
12.3 MODELS INTEGRATION

We used Azure Web Service to create Web App. For the first model, we allow users to upload their own CSV files as that model's input, and the model will generate a CSV file as the output of prediction. Thus, we used Batch Execution Web App for first model.

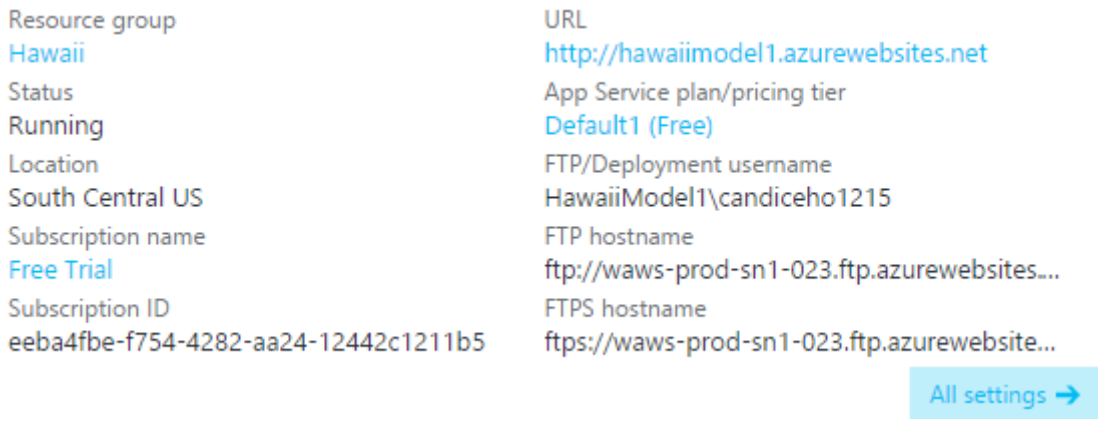
For second and third model, we allow users to input each variable value, and models will give them one prediction value for the certain input. Thus, we used Request-Response Web App.

12.3.1 BATCH EXECUTION INTEGRATION

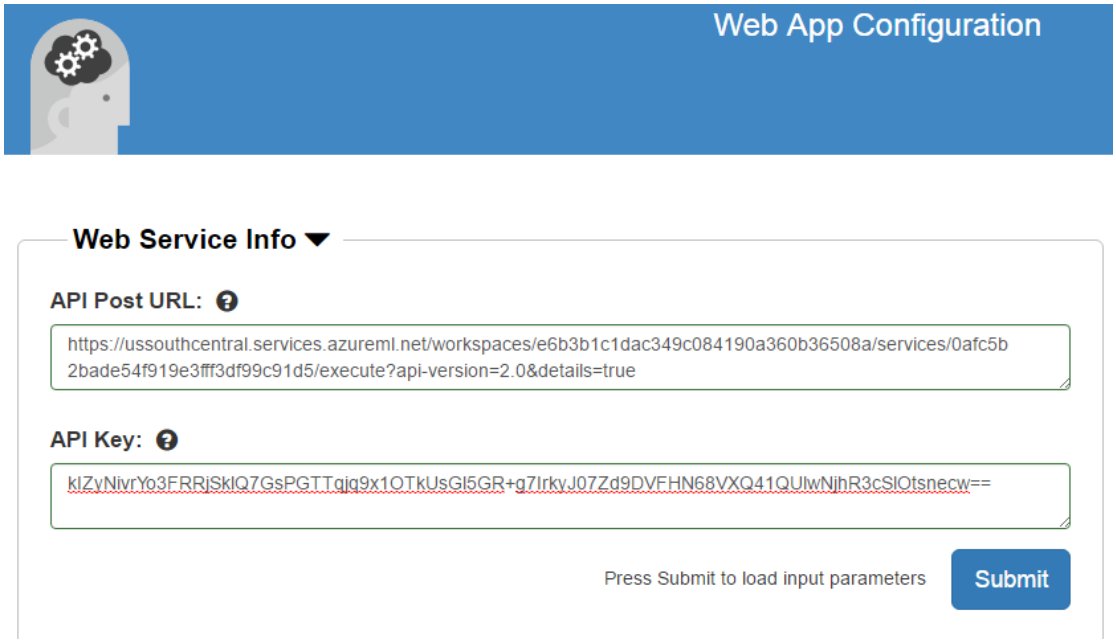
In Azure studio, we create an Azure ML Batch Execution Service Web App, and input the API key and URL of models.



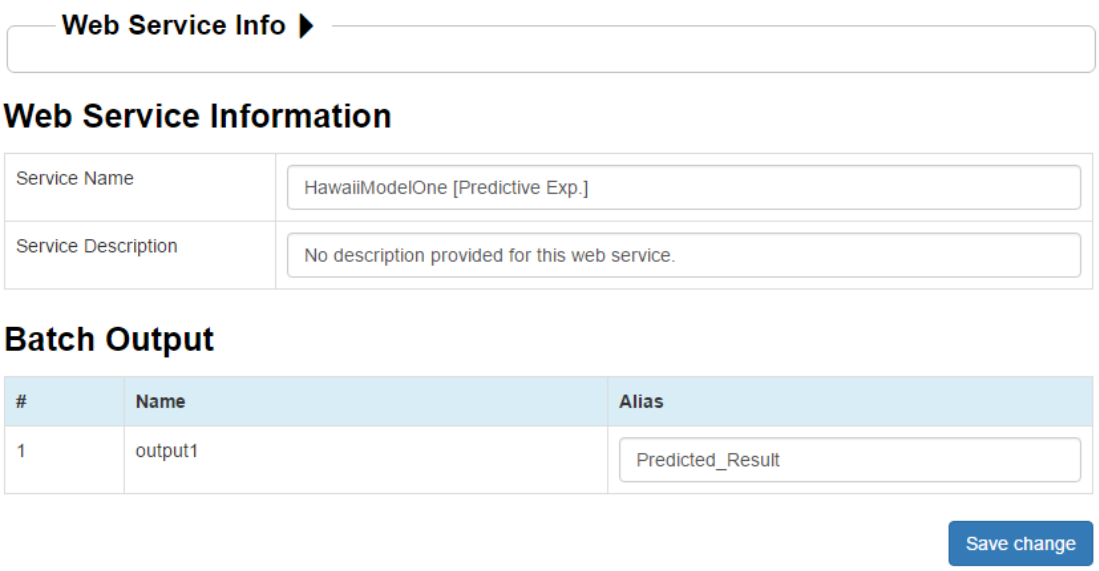
This is the process to create a Web App in Azure Studio. After inputting name of App name and subscription, Azure will create a Web app for us. And it will give us a URL of this Web app.



Now, this page is just an empty page, and it does not have any function. We need click the URL and configure this page.



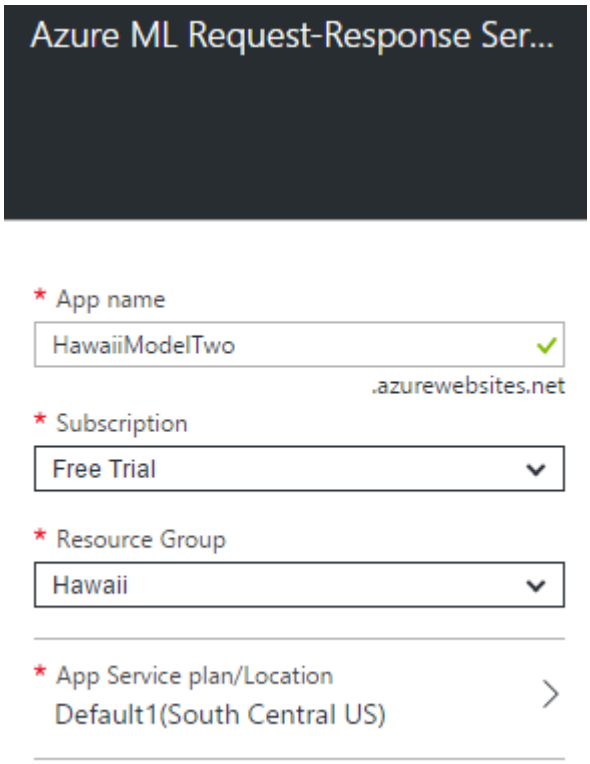
Now, we can copy the API key and URL of one of model to here to implements that model to this page.



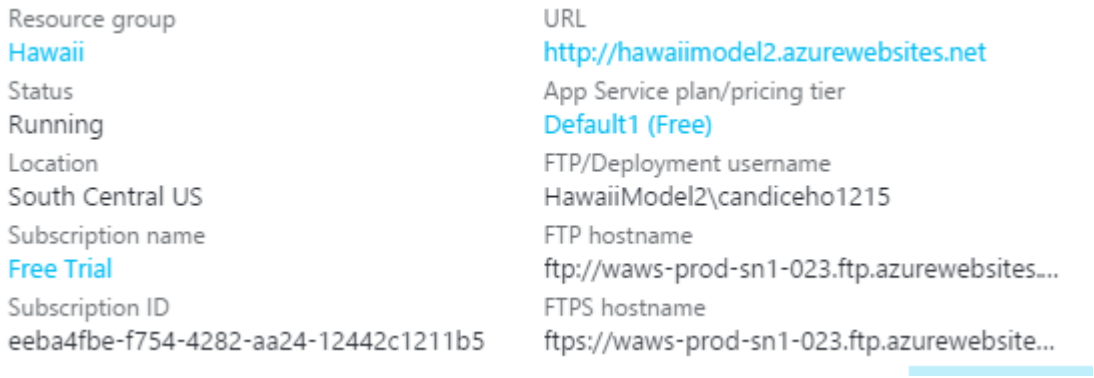
After implementing models, we can set the default value to this page. This model will generate a CSV file as the prediction result and store it in customer's Storage Account, or customers can download this CSV file. Thus, we can set a default name for this result CSV files.

12.3.2 REQUEST/ RESPONSE INTEGRATION

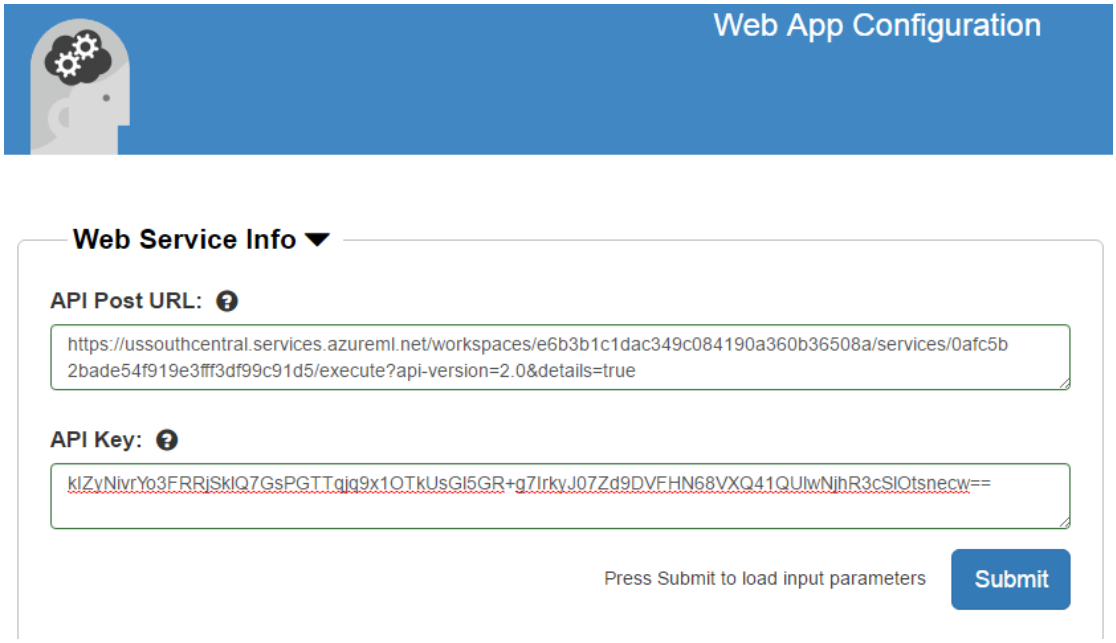
In Azure studio, we create an Azure ML Request-Response Service Web App, and input the API key and URL of models.



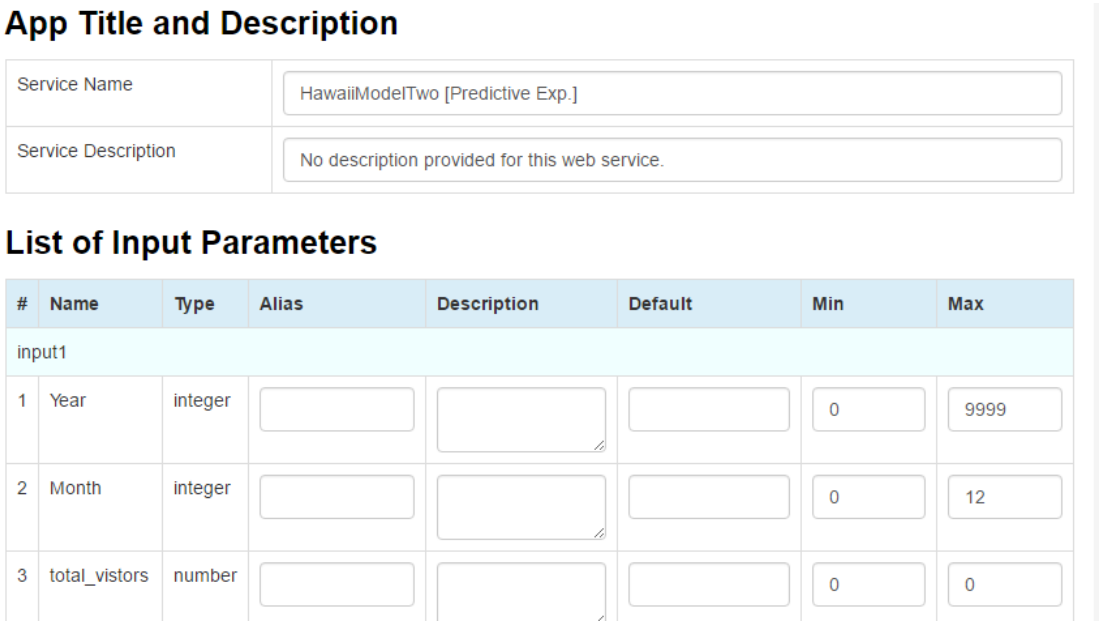
This is the process to create a Request/Response Service Web App in Azure Studio. After inputting name of App name and subscription, Azure will create a Web app for us. And it will give us a URL of this Web app.



Now, this page is just an empty page, and it does not have any function. We need click the URL and configure this page.



Now, we can copy the API key and URL of one of model to here to implements that model to this page.



App Title and Description

Service Name	HawaiiModelTwo [Predictive Exp.]
Service Description	No description provided for this web service.

List of Input Parameters

#	Name	Type	Alias	Description	Default	Min	Max
input1							
1	Year	integer				0	9999
2	Month	integer				0	12
3	total_vistors	number				0	0

After implementing models, we can set some default value to this page. In this setting page, we can change the alias, max and min value of each feature. We also can set default value to any of feature.

List of Output Parameters

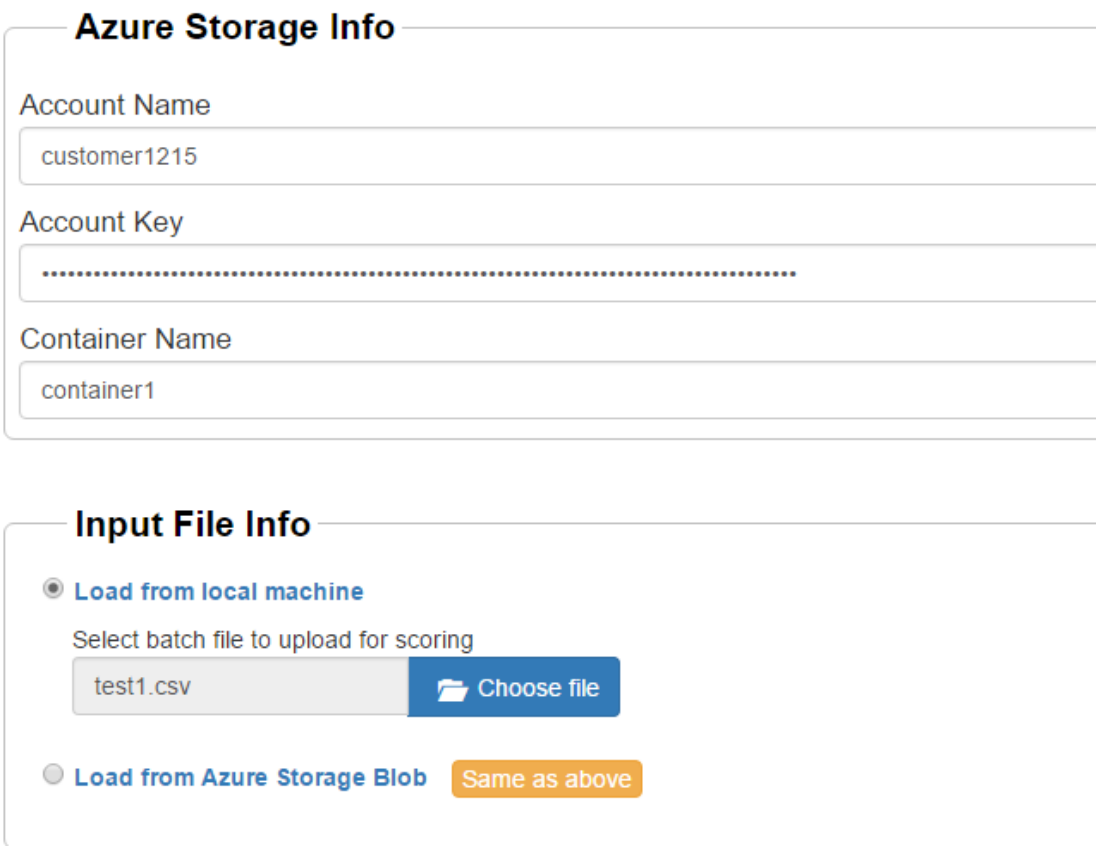
#	Name	Type	Alias	Enabled
1	Year	integer		ON
2	Month	integer		ON
3	total_vistors	number		OFF
4	Average high temperature	number		ON
5	Average low temperature	number		ON
6	Average temperature	number		ON
7	Island	string		ON
8	Scored Labels	number	Total_Visitors	ON

In this setting page, we also can choose what features we want to show in the predict result table. Because this model is for predicting Total_visitors, so we decided to not show the original Total_Visitors value in our result. Result will show the value of Island, Country, Year, Month, High temperature, Low temperature and Average temperature user input, and the predictive Total_Visitors value.

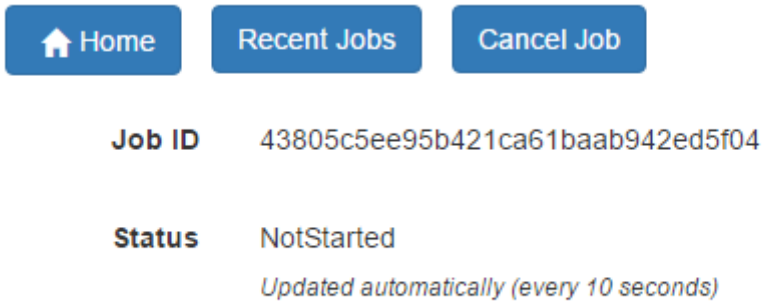
12.4 RESULT TESTING

We created three predictive web pages for three models separately. The screen shots of result testing process are shown below.

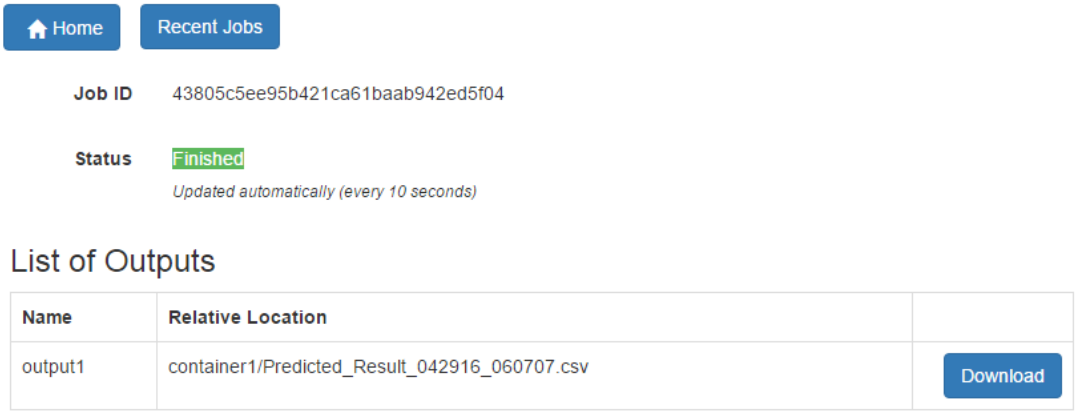
12.4.1 PREDICTION TEST (VISITOR AMOUNT OF EACH ISLAND FROM EACH COUNTRY)



Users must input own Storage Account, Container and Access Key. These are used for receiving the result of this prediction model. Also, user must give a input to this model, uploading from local or from Storage Blob. Users can choose the same Storage Blob to give an input CSV and receive the result CSV.



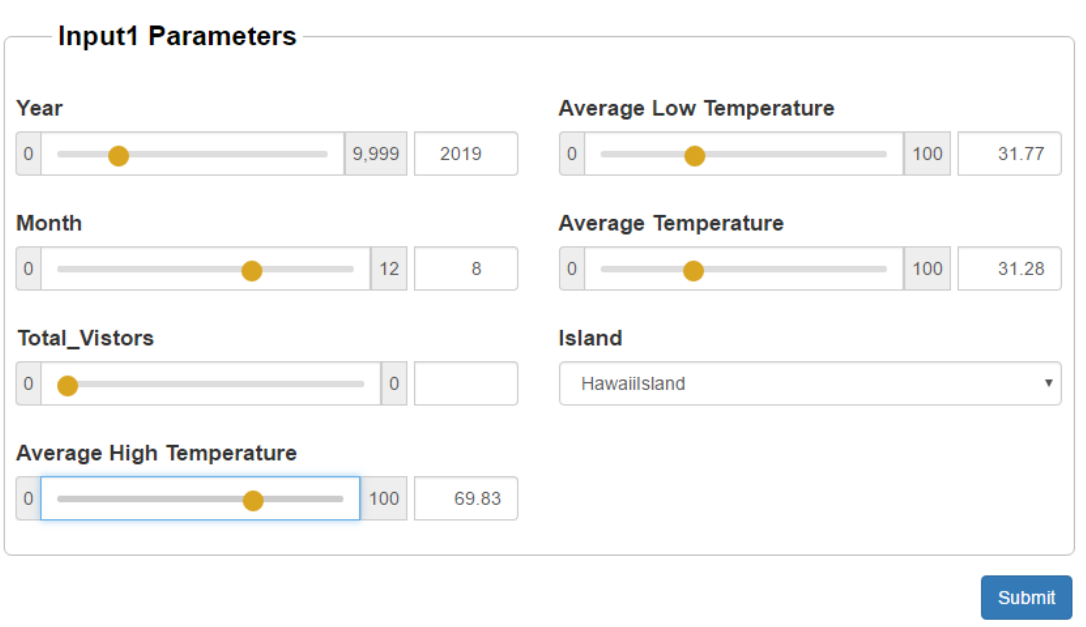
Each prediction has its own unique Job ID. When Status show “Finished”, you can check the predicted result.



Users also can download the result CSV which has been stored in user's Storage Account automatically. The predicted values will be display in the last column of this CSV with the header "Score Labels".

	Year	Month	Average h	Average l	Average t	Island	Country	Scored Labels
2	2016	1	63.26	48.31	55.87	HawaiiIsl	US_West	488145.3
3	2016	1	63.26	48.31	55.87	HawaiiIsl	US_East	303612.1
4	2016	1	63.26	48.31	55.87	HawaiiIsl	Canada	119174.4
5	2016	1	63.26	48.31	55.87	HawaiiIsl	Japan	580266.8
6	2016	1	63.26	48.31	55.87	Kaua'i	Canada	406784.6
7	2016	1	63.26	48.31	55.87	Kaua'i	Japan	651760.2
8	2016	1	63.26	48.31	55.87	Lanai	US_West	293998.1
9	2016	1	63.26	48.31	55.87	Lanai	US_East	295429.4
10	2016	1	63.26	48.31	55.87	Lanai	Canada	184560.1

12.4.2 PREDICTION TEST (TOTAL VISITOR AMOUNT OF EACH ISLAND)



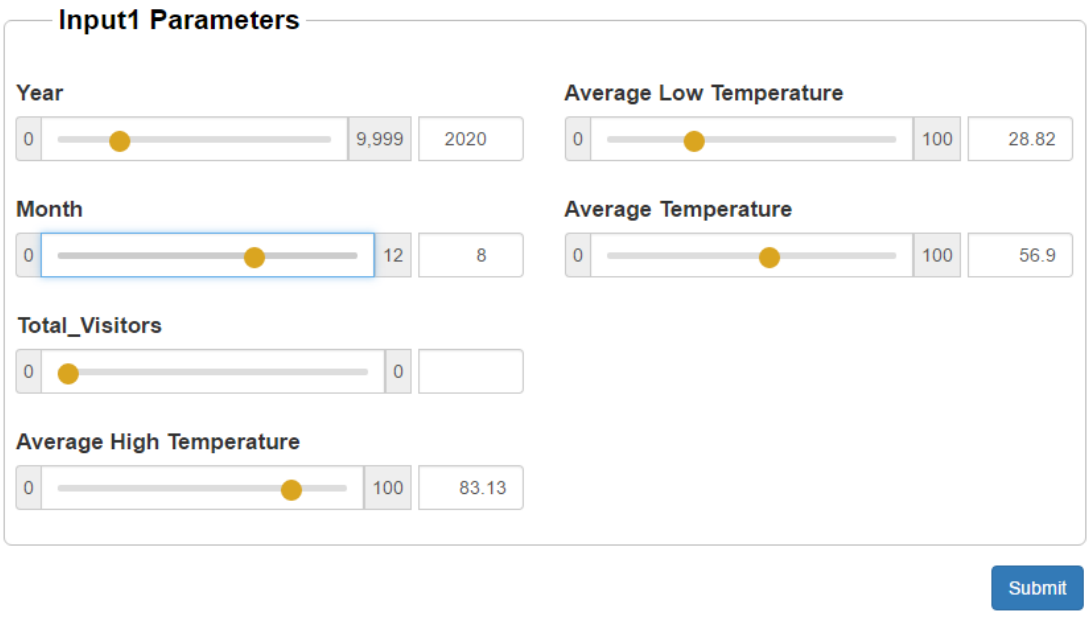
Users should input each parameter's value, and select the Island. Because this model is for predicting the total visitor amount of one island in a future month, user need not input visitor value.

Result

Label	Value
output1	
Year	2019
Month	8
Average High Temperature	69.83
Average Low Temperature	31.77
Average Temperature	31.28
Island	HawaiiIsland
Total_Visitors	122981.1953125

After clicking Submit button, we can get a result table. The predicted total visitor amount in 08/2019 is shown in the last line. This result come from the model2 created in Azure Machine Learning Workspace. About the result accuracy, we talked about it in Creating Models part before.

12.4.3 PREDICTION TEST (TOTAL VISITOR AMOUNT OF ENTIRE HAWAII)



Also, users should input each parameter's value as the input of this model. Because this model is for predicting the total visitor amount of Hawaii in a future month, user need not input visitor value.

Result

Label	Value
output1	
Year	2020
Month	8
Average High Temperature	83.13
Average Low Temperature	28.82
Average Temperature	56.9
Total_Visitors	786226.4375

After clicking Submit button, we can get a result table. The predicted total visitor amount in 08/2020 is shown in the last line. This result come from the model3 created in Azure Machine Learning Workspace. About the result accuracy, we talked about it in Creating Models part before.

13. UI & INTEGRATION

13.1 UI TOOLS

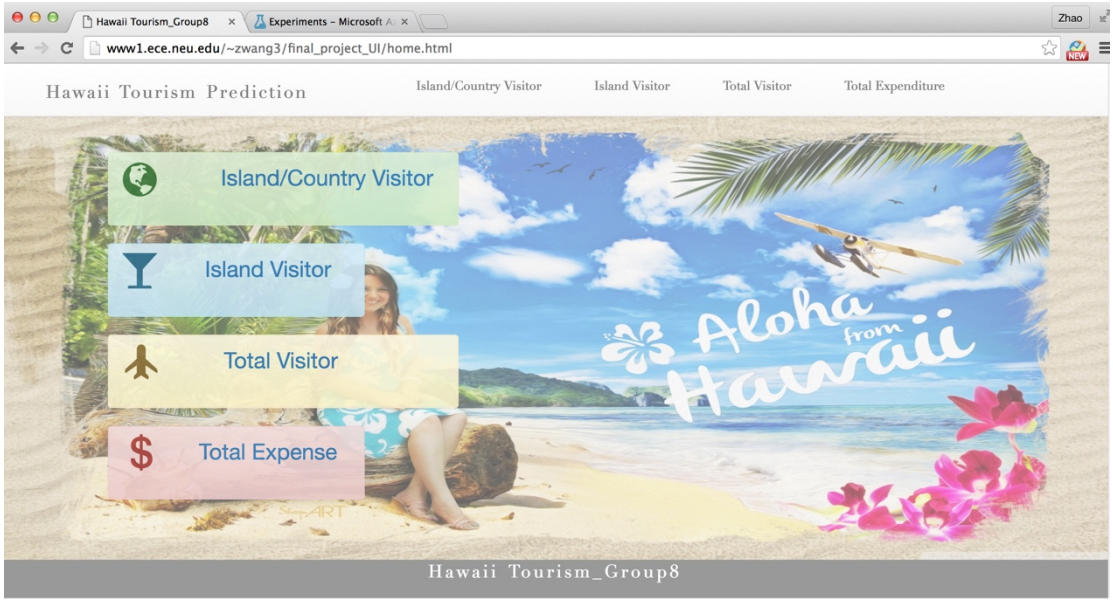
We develop our User Interface using HTML, CSS, Bootstrap and JavaScript. And we also combine our UI with web service UI from Azure Studio.

13.2 WEB APP EXPLANATION

1) Home Page

URL: http://www1.ece.neu.edu/~zwang3/final_project_UI/home.html

The home page contains four parts: Island/Country Visitor Model; Island Visitor Model; Total Visitor Model; Total Expense Model. You can click each of them to go to the corresponding model page.



2) Island/Country Visitor Prediction Page

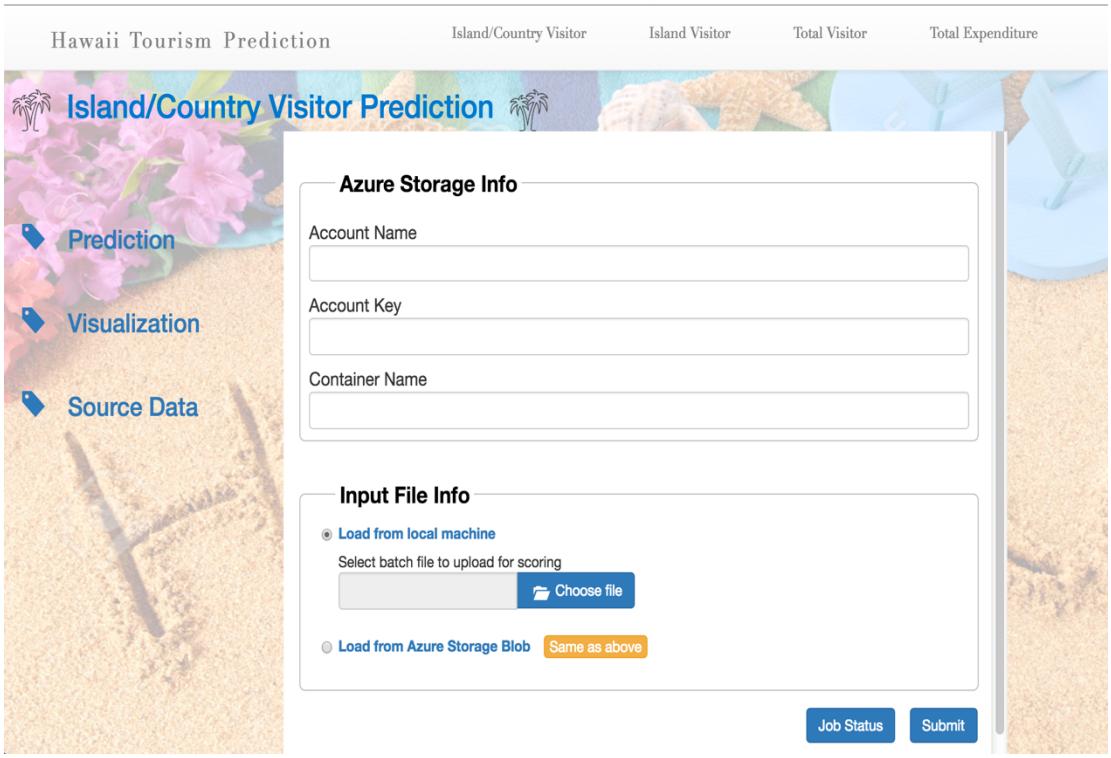
After click "Island/country Visitor" menu, you will enter "Island/Country Visitor Prediction" page. It contains three parts, prediction part, visualization part, and source data part.

In prediction page, you can input following Azure Storage Information:

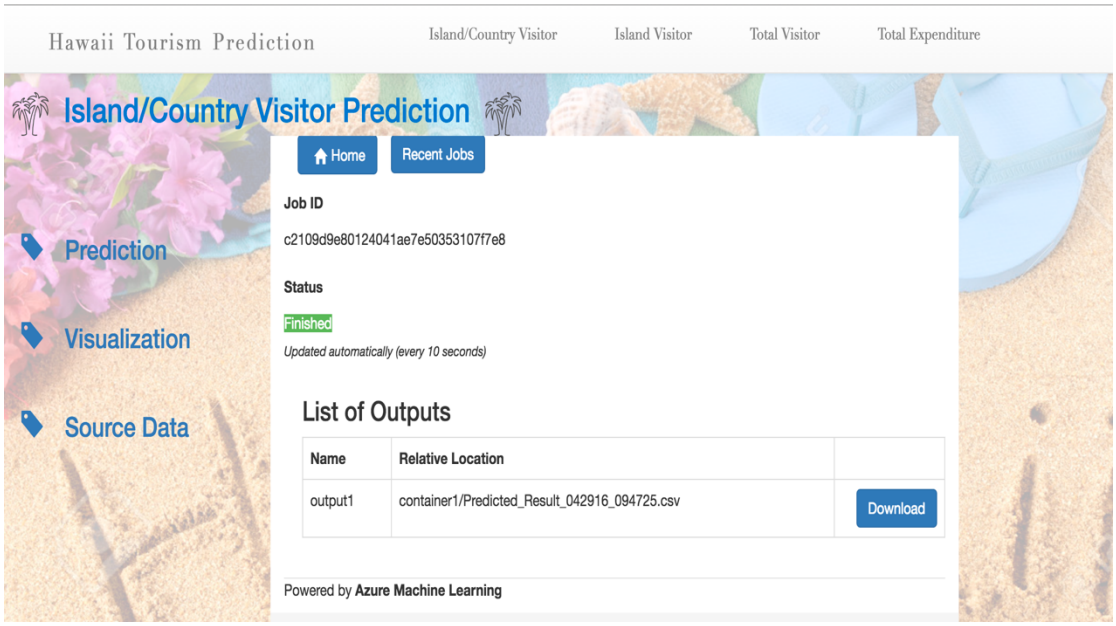
Account Name: customer1215

AccountKey: 6lcGoWsDu9wTEZki0RsNHAV3fttcQwcRYKxBE7pBTvZ26T5z8N5Y9fnNGHKFXqGW8qu4smyPK+0OcAAYJ9w4Zw==

Container Name: container1

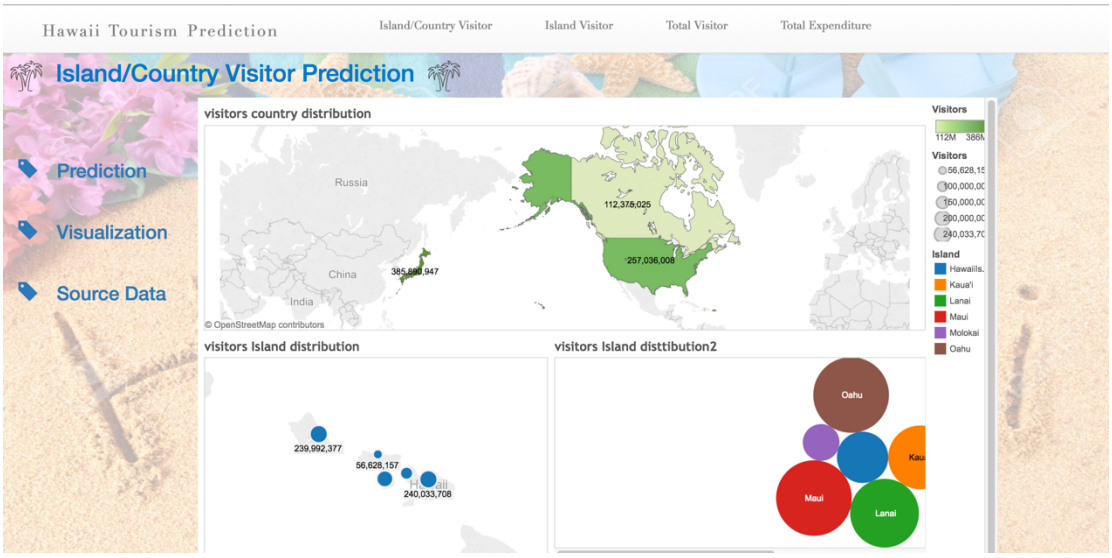


And then you can choose file from blob storage or upload file from your computer, and click submit button. After prediction, user can download the predict result.



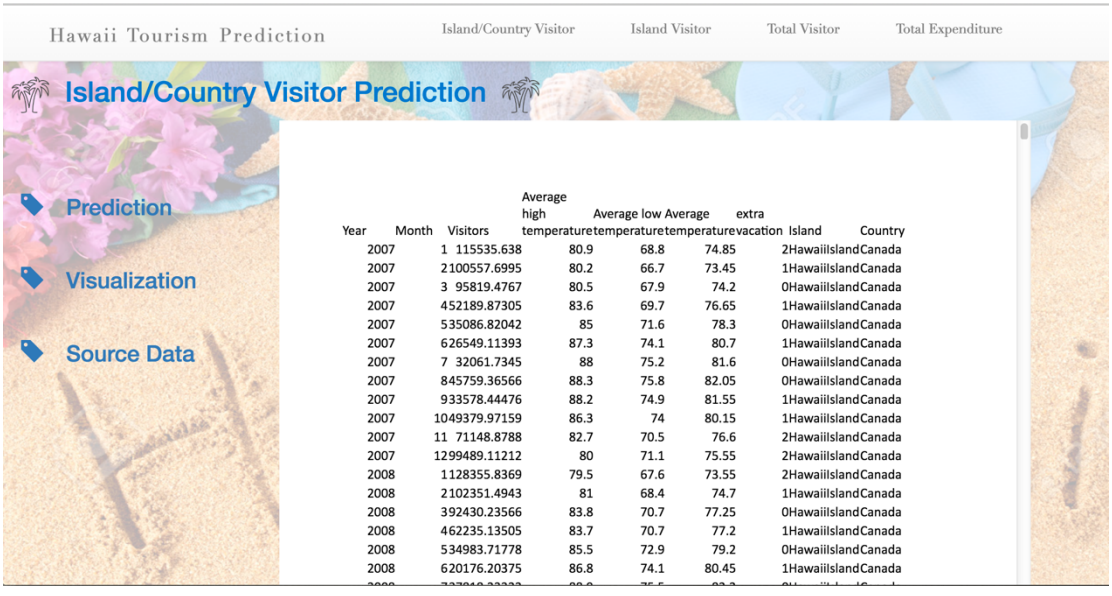
3) Island/Country Visitor visualization Page

After click the “Visualization” menu, user can view the visualization analysis for “island/country visitor dataset”. We use Tableau public for that.



4) Island/Country Visitor SourceData Page

After click the “Source Data” menu, user can view the “island/country visitor dataset”.

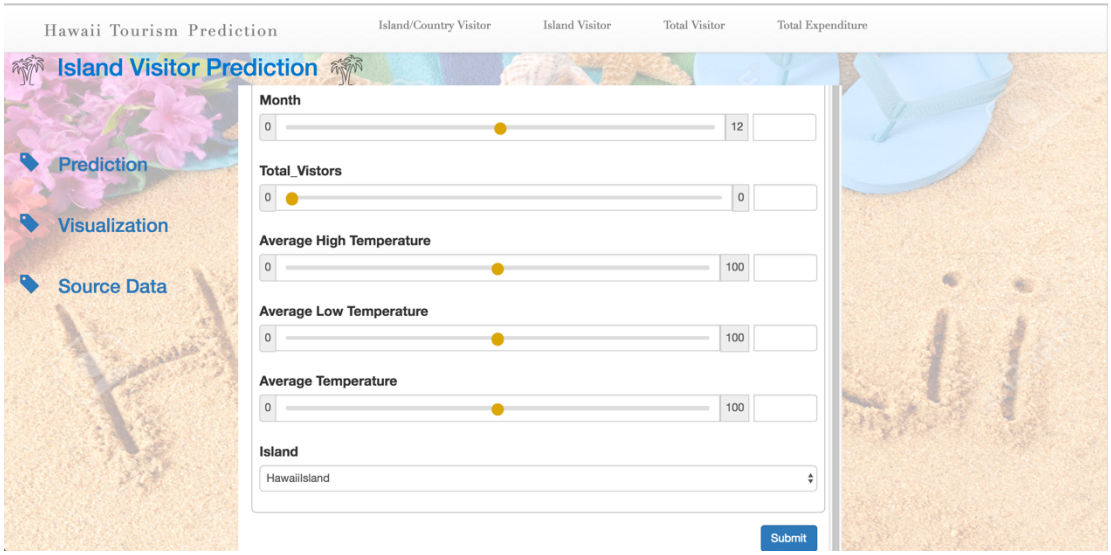


Year	Month	Visitors	Average high temperature	Average low temperature	Average temperature	extra vacation	Island	Country
2007	1	115535.638	80.9	68.8	74.85	2	HawaiiIsland	Canada
2007		2100557.6995	80.2	66.7	73.45	1	HawaiiIsland	Canada
2007	3	95819.4767	80.5	67.9	74.2	0	HawaiiIsland	Canada
2007		452189.87305	83.6	69.7	76.65	1	HawaiiIsland	Canada
2007		535086.82042	85	71.6	78.3	0	HawaiiIsland	Canada
2007		626549.11393	87.3	74.1	80.7	1	HawaiiIsland	Canada
2007	7	32061.7345	88	75.2	81.6	0	HawaiiIsland	Canada
2007		845759.36566	88.3	75.8	82.05	0	HawaiiIsland	Canada
2007		933578.44476	88.2	74.9	81.55	1	HawaiiIsland	Canada
2007		1049379.97159	86.3	74	80.15	1	HawaiiIsland	Canada
2007	11	71148.8788	82.7	70.5	76.6	2	HawaiiIsland	Canada
2007		1299489.11212	80	71.1	75.55	2	HawaiiIsland	Canada
2008		1128355.8369	79.5	67.6	73.55	2	HawaiiIsland	Canada
2008		2102351.4943	81	68.4	74.7	1	HawaiiIsland	Canada
2008		392430.23566	83.8	70.7	77.25	0	HawaiiIsland	Canada
2008		462235.13505	83.7	70.7	77.2	1	HawaiiIsland	Canada
2008		534983.71778	85.5	72.9	79.2	0	HawaiiIsland	Canada
2008		620176.20375	86.8	74.1	80.45	1	HawaiiIsland	Canada

5) Island Visitor Prediction Page

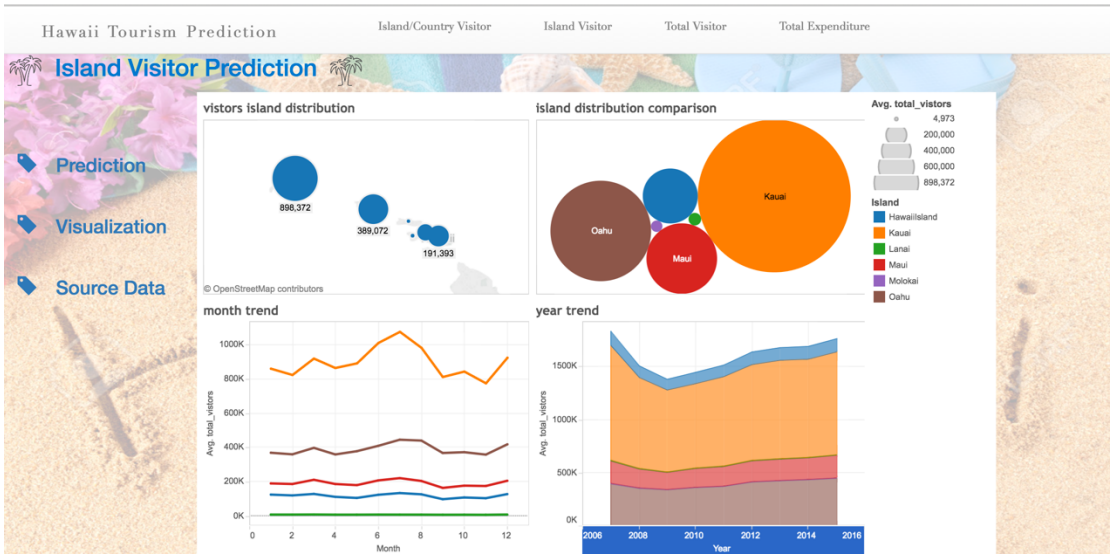
After click “Island Visitor” menu, you will enter “Island Visitor Prediction” page. It contains three parts, prediction part, visualization part, and source data part.

You can input month, high/low temperature, island, and then click “submit” button to predict visitor numbers for one island.



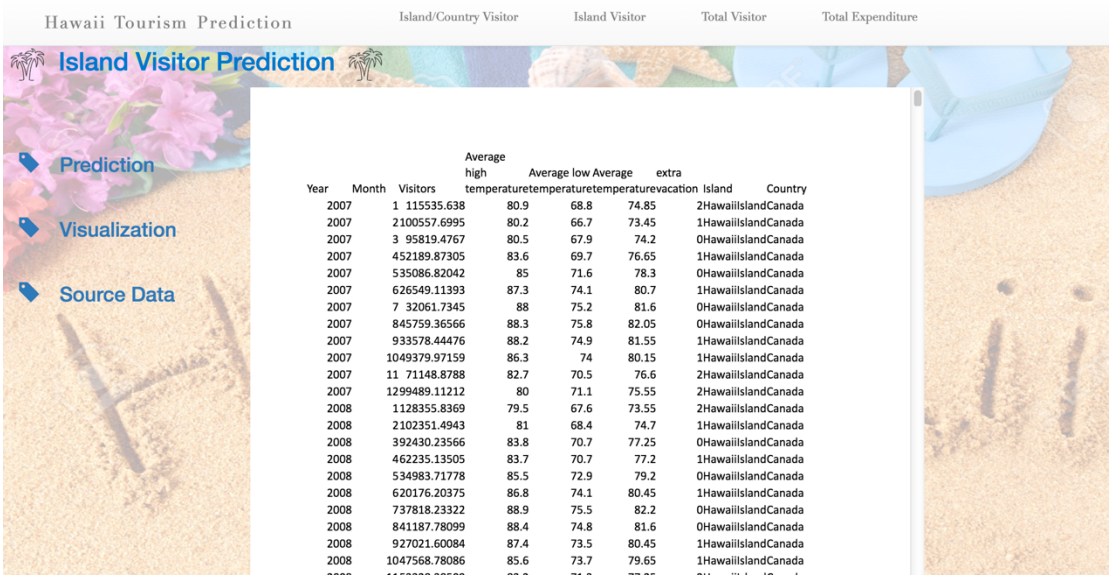
6) Island Visitor Visualization Page

After click the “Visualization” menu, user can view the visualization analysis for “island visitor dataset”. We use Tableau public for that.



7) Island Visitor SourceData Page

After click the “Source Data” menu, user can view the “island visitor dataset”.

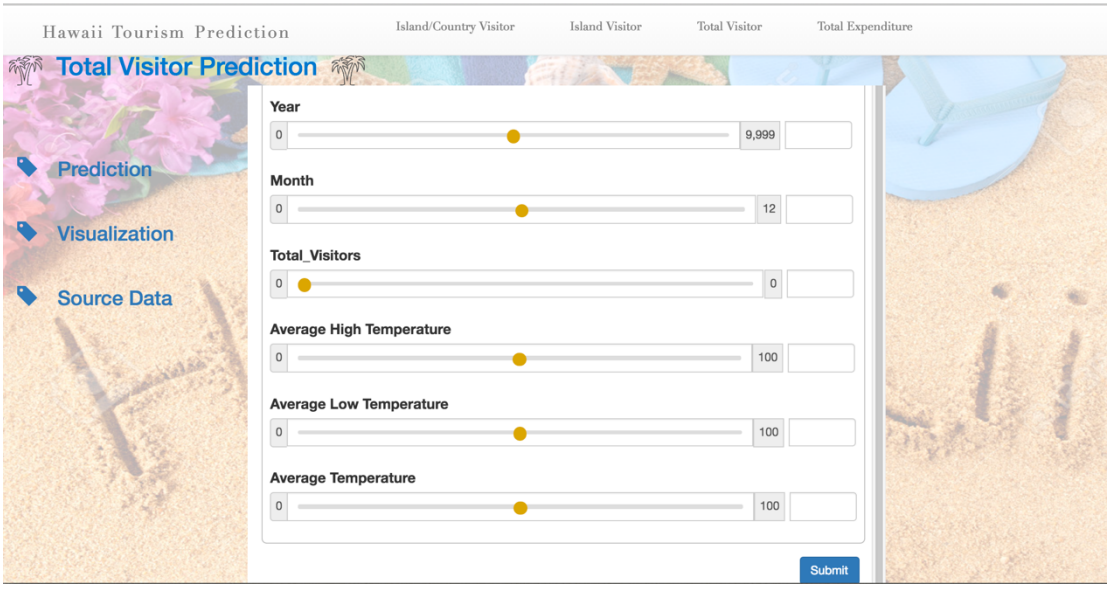


Year	Month	Visitors	Average high temperature	Average low temperature	Average temperature	extra vacation	Island	Country
2007	1	115535.638	80.9	68.8	74.85	2	HawaiiIsland	Canada
2007		2100557.6995	80.2	66.7	73.45	1	HawaiiIsland	Canada
2007	3	95819.4767	80.5	67.9	74.2	0	HawaiiIsland	Canada
2007		452189.87305	83.6	69.7	76.65	1	HawaiiIsland	Canada
2007		535086.82042	85	71.6	78.3	0	HawaiiIsland	Canada
2007		626549.11393	87.3	74.1	80.7	1	HawaiiIsland	Canada
2007	7	32061.7345	88	75.2	81.6	0	HawaiiIsland	Canada
2007		845759.36566	88.3	75.8	82.05	0	HawaiiIsland	Canada
2007		933578.44476	88.2	74.9	81.55	1	HawaiiIsland	Canada
2007		1049379.97159	86.3	74	80.15	1	HawaiiIsland	Canada
2007	11	71148.8788	82.7	70.5	76.6	2	HawaiiIsland	Canada
2007		1299489.11212	80	71.1	75.55	2	HawaiiIsland	Canada
2008		1128355.8369	79.5	67.6	73.55	2	HawaiiIsland	Canada
2008		2102351.4943	81	68.4	74.7	1	HawaiiIsland	Canada
2008		392430.23566	83.8	70.7	77.25	0	HawaiiIsland	Canada
2008		462235.13505	83.7	70.7	77.2	1	HawaiiIsland	Canada
2008		534983.71778	85.5	72.9	79.2	0	HawaiiIsland	Canada
2008		620176.20375	86.8	74.1	80.45	1	HawaiiIsland	Canada
2008		737818.23322	88.9	75.5	82.2	0	HawaiiIsland	Canada
2008		841187.78099	88.4	74.8	81.6	0	HawaiiIsland	Canada
2008		927021.60084	87.4	73.5	80.45	1	HawaiiIsland	Canada
2008		1047568.78086	85.6	73.7	79.65	1	HawaiiIsland	Canada

8) Total Visitor Prediction Page

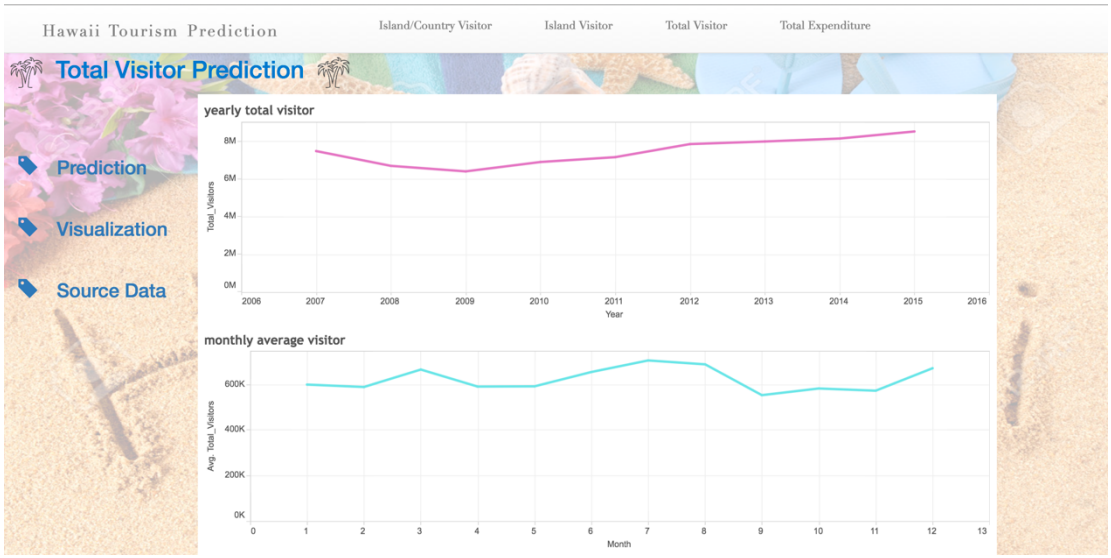
After click “Total Visitor” menu, you will enter “Total Visitor Prediction” page. It contains three parts, prediction part, visualization part, and source data part.

You can input month, high/low temperature, and then click “submit” button to predict visitor numbers for entire Hawaii area.



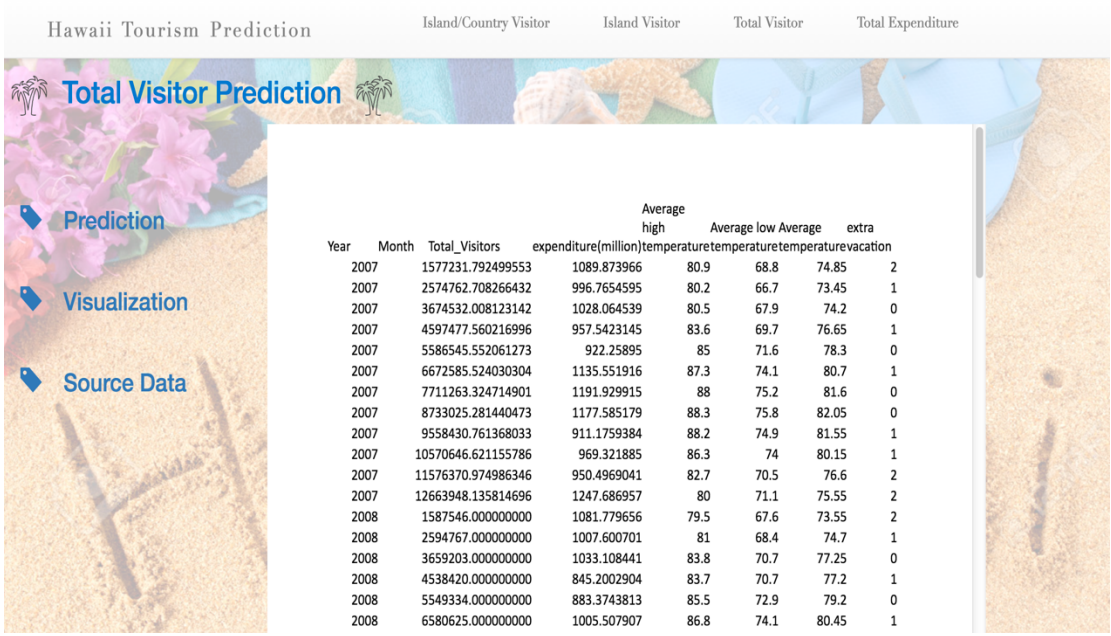
9) Total Visitor Visualization Page

After click the “Visualization” menu, user can view the visualization analysis for “total visitor dataset”. We use Tableau public for that.



10) Total Visitor SourceData Page

After click the “Source Data” menu, user can view the “total visitor dataset”.

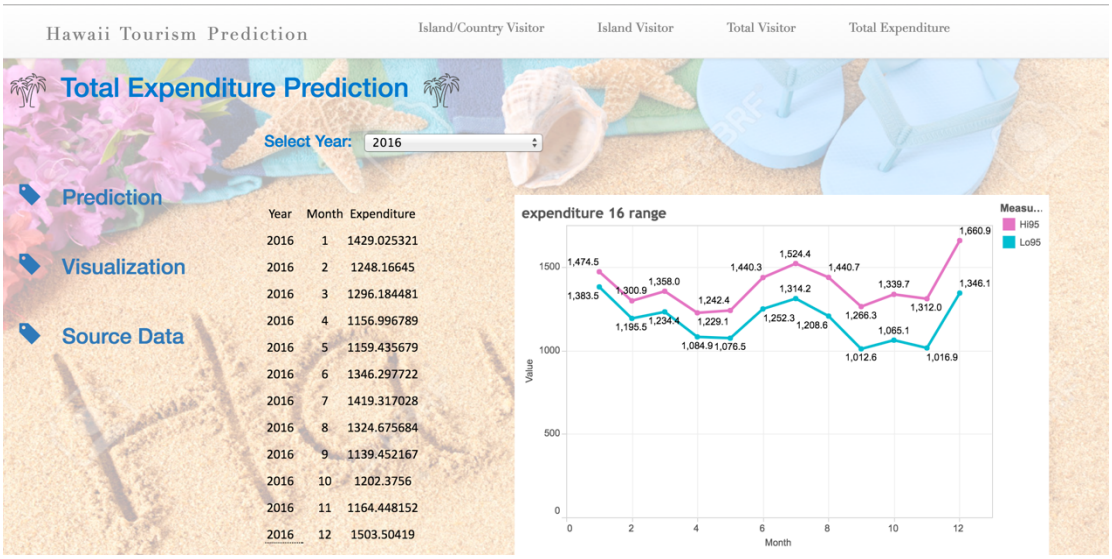


Year	Month	Total_Visitors	expenditure(million)	Average high temperature	Average low temperature	Average temperature	extra vacation
2007		1577231.792499553	1089.873966	80.9	68.8	74.85	2
2007		2574762.708266432	996.7654595	80.2	66.7	73.45	1
2007		3674532.008123142	1028.064539	80.5	67.9	74.2	0
2007		4597477.560216996	957.5423145	83.6	69.7	76.65	1
2007		5586545.552061273	922.25895	85	71.6	78.3	0
2007		6672585.524030304	1135.551916	87.3	74.1	80.7	1
2007		7711263.324714901	1191.929915	88	75.2	81.6	0
2007		8733025.281440473	1177.585179	88.3	75.8	82.05	0
2007		9558430.761368033	911.1759384	88.2	74.9	81.55	1
2007		10570646.621155786	969.321885	86.3	74	80.15	1
2007		11576370.974986346	950.4969041	82.7	70.5	76.6	2
2007		12663948.135814696	1247.686957	80	71.1	75.55	2
2008		1587546.000000000	1081.779656	79.5	67.6	73.55	2
2008		2594767.000000000	1007.600701	81	68.4	74.7	1
2008		3659203.000000000	1033.108441	83.8	70.7	77.25	0
2008		4538420.000000000	845.2002904	83.7	70.7	77.2	1
2008		5549334.000000000	883.3743813	85.5	72.9	79.2	0
2008		6580625.000000000	1005.507907	86.8	74.1	80.45	1

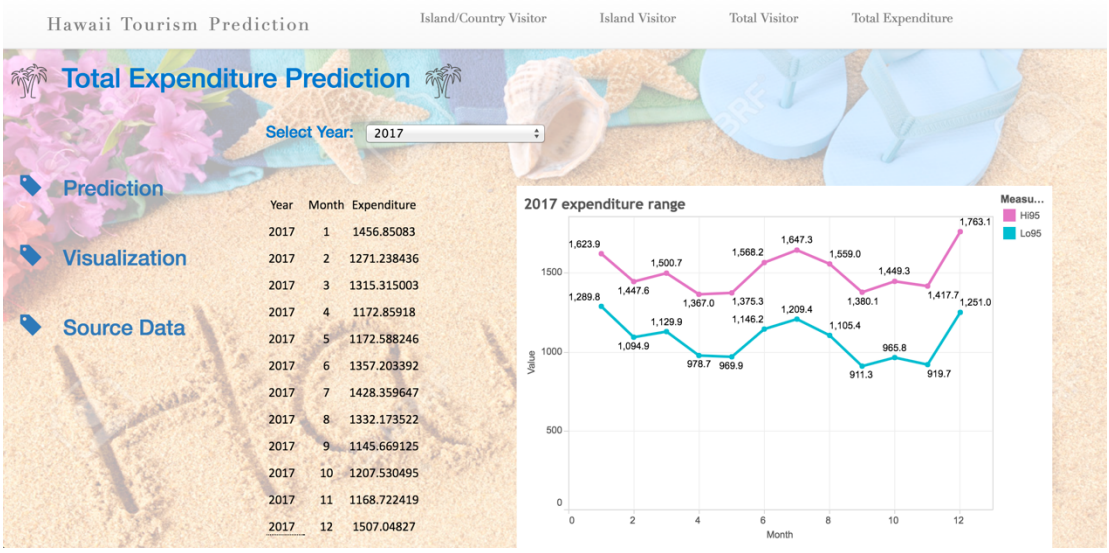
11) Total Expenditure Prediction Page

After click “Total Expenditure” menu, you will enter “Total Expenditure Prediction” page. It contains three parts, prediction part, visualization part, and source data part.

In prediction page, if user select year 2016, it will show the value and chart for predicted monthly expenditure in 2016.



In prediction page, if user select year 2017, it will show the value and chart for predicted monthly expenditure in 2017.

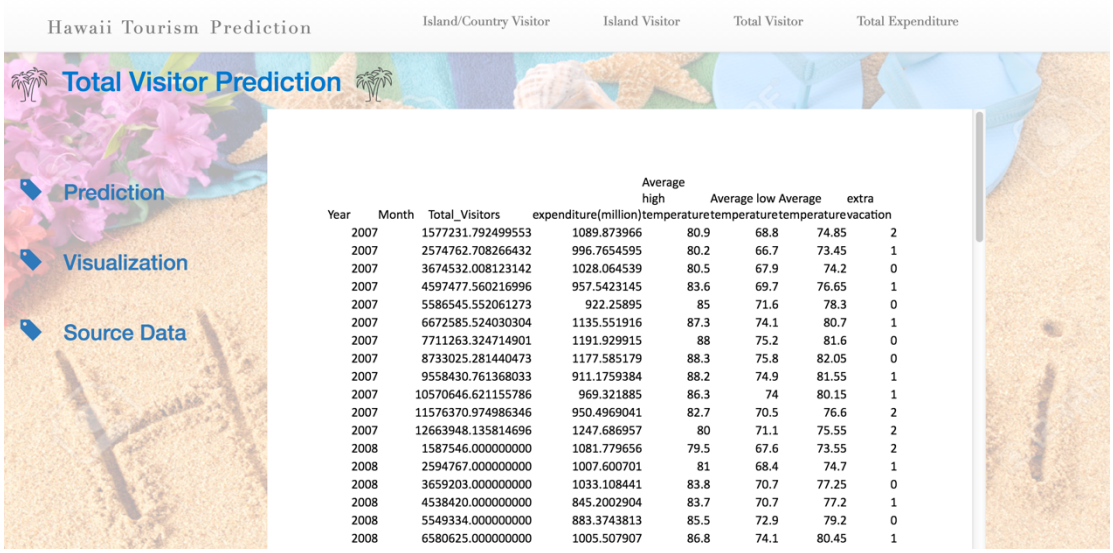


12) Total Expenditure Visualization Page

After click the “Visualization” menu, user can view the visualization analysis for “total expenditure dataset”. We use Tableau public for that.

13) Total Expenditure SourceData Page

After click the “Source Data” menu, user can view the “total visitor/expenditure dataset”.



Year	Month	Total_Visitors	expenditure(million)	Average high temperature	Average low temperature	Average temperature	extra vacation
2007		1577231.792499553	1089.873966	80.9	68.8	74.85	2
2007		2574762.708266432	996.7654595	80.2	66.7	73.45	1
2007		3674532.008123142	1028.064539	80.5	67.9	74.2	0
2007		4597477.560216996	957.5423145	83.6	69.7	76.65	1
2007		5586545.552061273	922.25895	85	71.6	78.3	0
2007		6672585.524030304	1135.551916	87.3	74.1	80.7	1
2007		7711263.324714901	1191.929915	88	75.2	81.6	0
2007		8733025.281440473	1177.585179	88.3	75.8	82.05	0
2007		9558430.761368033	911.1759384	88.2	74.9	81.55	1
2007		10570646.621155786	969.321885	86.3	74	80.15	1
2007		11576370.974986346	950.4969041	82.7	70.5	76.6	2
2007		12663948.135814696	1247.686957	80	71.1	75.55	2
2008		1587546.000000000	1081.779656	79.5	67.6	73.55	2
2008		2594767.000000000	1007.600701	81	68.4	74.7	1
2008		3659203.000000000	1033.108441	83.8	70.7	77.25	0
2008		4538420.000000000	845.2002904	83.7	70.7	77.2	1
2008		5549334.000000000	883.3743813	85.5	72.9	79.2	0
2008		6580625.000000000	1005.507907	86.8	74.1	80.45	1

14. TIME SCHEDULE

Time	Content
April 18-19	Decide the topic and find datasets, finish the proposal.
April 20-22	Pro-process the datasets.
April 23	Visualization and analyze datasets.
April 24-26	Compare and choose the algorithms and building & evaluate the model.
April 26-28	Develop the web service and UI.
April 29	Maintenance and improvement, finish the report.
April 30	Presentation, we will deliver a fully functioning website, full source code, report and ppt.

15. CHALLENGES

1. Let user upload his dataset and build prediction model based on his dataset, in order to improve model flexibility accuracy.

2. Better integration: optimize UI and used JavaScript to improve user experience.