DATA MINING

Analysis of Bike sharing dataset

April 13, 2014

UT DALLAS
NAVEEN JINDAL
SCHOOL OF MANAGEMENT

Project Report for Analysis of bike sharing dataset

MIS-6324 Intro. to business intelligence software and techniques

Prepared by

Abhay Satish Joshi

Under the guidance of Professor

Kelly Slaughter, PhD
Clinical Professor
Information Systems
University of Texas at Dallas

Table of Contents

1.Introduction to Data Mining

Data mining (the analysis step of the "Knowledge Discovery in Databases" process), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

Data mining involves six common classes of tasks:

- Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule (Dependency modelling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function which models the data with the least error.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

A wide range of industries - including retail, finance, heath care, manufacturing transportation, and aerospace - are already using data mining tools and techniques to take advantage of historical data.
Specific uses of data mining are Market segmentation, Customer churn, Fraud detection, Direct marketing, Interactive marketing, Market basket analysis, Trend analysis.

2. Background of the dataset

Bike sharing systems are new generation of traditional bike rentals where the whole process from membership, rental and return back is automatic. Through these systems, user is able to easily rent a bike from a particular location and return it back at another location. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

2.1 Description of dataset

This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bike share system with the corresponding weather and seasonal information. The size of the file containing the data set is 1.10 MB (1,156,736 bytes).The Dataset contains 17389 records and 16 attributes. The hourly dataset has been consolidated into the day dataset.

Source:
https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset
HadiFanaee-T
Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto
INESC Porto, Campus da FEUP
Rua Dr. Roberto Frias, 378
4200 - 465 Porto, Portugal

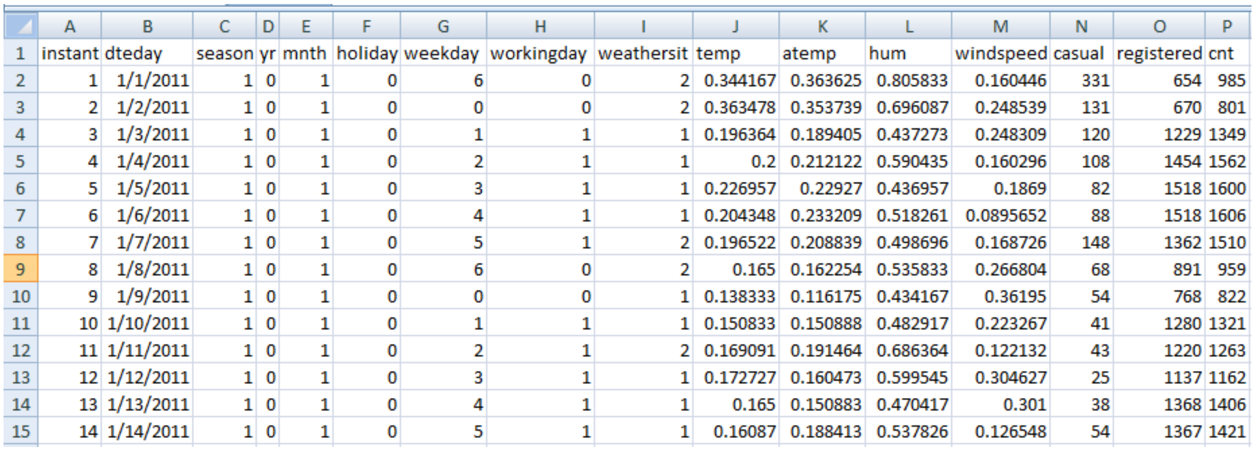
Original Source: http://capitalbikeshare.com/system-data
Weather Information: http://www.freemeteo.com
Holiday Schedule: http://dchr.dc.gov/page/holiday-schedule

List of variables present in the dataset

- instant: record index
- dteday : date
- season : season
- yr : year
- mnth : month
- hr : hour
- holiday : weather day is holiday or not
- weekday : day of the week
- working day
- weathersit
- temp : Normalized temperature in Celsius.
- atemp: Normalized feeling temperature in Celsius.
- hum: Normalized humidity.
- windspeed: Normalized wind speed

- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

Snapshot of the dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
2	1	1/1/2011	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
3	2	1/2/2011	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
4	3	1/3/2011	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
5	4	1/4/2011	1	0	1	0	2	1	1	0.2	0.212122	0.590435	0.160296	108	1454	1562
6	5	1/5/2011	1	0	1	0	3	1	1	0.226957	0.22927	0.436957	0.1869	82	1518	1600
7	6	1/6/2011	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.0895652	88	1518	1606
8	7	1/7/2011	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
9	8	1/8/2011	1	0	1	0	6	0	2	0.165	0.162254	0.535833	0.266804	68	891	959
10	9	1/9/2011	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.36195	54	768	822
11	10	1/10/2011	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321
12	11	1/11/2011	1	0	1	0	2	1	2	0.169091	0.191464	0.686364	0.122132	43	1220	1263
13	12	1/12/2011	1	0	1	0	3	1	1	0.172727	0.160473	0.599545	0.304627	25	1137	1162
14	13	1/13/2011	1	0	1	0	4	1	1	0.165	0.150883	0.470417	0.301	38	1368	1406
15	14	1/14/2011	1	0	1	0	5	1	1	0.16087	0.188413	0.537826	0.126548	54	1367	1421

Figure 1Snapshot of the dataset

3.Outline of Analysis

The primary outline of this data mining analysis is to predict the success or failure of a bike rental company based on the profitability on a particular day depending upon weather attributes.

Three important means of analysis include:

- Plot the relationship between key weather elements with respect to profitability and provide insight into the trends.
- Design a decision tree to classify whether a day is profitable or not based upon combination of weather elements.
- Build models and select the optimum model which effectively predicts probability of a day being profitable.

4. The Methodology

The methodology used to achieve the analysis for each of the above means is described below:

- Plot the relationship between key weather elements with respect to profitability and provide insight into the trends.

  Key weather elements such as season, temperature, humidity and windspeed are plotted with respect to count using weka and insight to profitability is gained with respect to each of these variables independently.

- Design a decision tree to classify whether a day is profitable or not based upon weather elements.

  The weather elements are combined and taken as independent variables to construct a decision tree to gain insight to ascertain as to whether the day is profitable or not. The decision tree is built in R using the tree command.
  This helps in predicting whether the day is profitable or not taking into account the overall weather conditions of the day establishing a strong relationship between the weather elements and the class label.

- Build models and select the optimum model which effectively predicts probability of a day being profitable.

  To better the result and get an accurate measurement as to whether the day turned out to be profitable, Models are constructed using the weather elements as independent variables and the variable profitable as the class label using logistic binomial regression in R. The best model is finally selected having the lowest AIC value and gives the most accurate prediction of probability of answering whether the day turned out to be profitable.

5. Pre-processing the dataset

Data pre-processing plays a very important in many deep learning algorithms. For the purpose of classification and logistic regression , the categorical values are converted into numeric measurements. A class label profitable is assigned. The details are mentioned below.

All the categorical values are converted into numeric measurements. Below are the conversion methods used

- **season** : season (1-spring, 2-summer, 3-fall, 4-winter)
- **yr** : year (0- 2011, 1-2012)
- **mnth** : month (1 to 12)
- **workingday** : if day is neither weekend nor holiday is 1, otherwise is 0.
- **weathersit** :
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

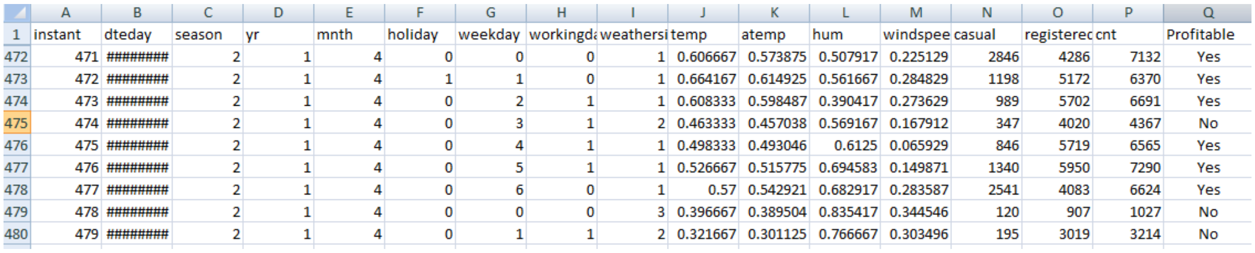
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

5.1 Class label

To build a decision tree/perform classification, The input fields to be classified and the class label field have to be specified. The class label field is also called target field. The class label field contains the class labels of the classes to which the records in the source data are attributed during the classification.

"Profitable" is used as the class label in the dataset. Making an assumption of 1$ per bike for a rental, The Median of count i.e. 4548 is considered as the profitable count. So for a particular day to be profitable it has to generate more than $4548. Hence any record below the median value is considered as Non-Profitable and above as Profitable.

Snapshot with Class label Profitable



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	instant	dteday	season	yr	mnth	holiday	weekday	workingda	weathersi	temp	atemp	hum	windspee	casual	registerec	cnt	Profitable
472	471	########	2	1	4	0	0	0	1	0.606667	0.573875	0.507917	0.225129	2846	4286	7132	Yes
473	472	########	2	1	4	1	1	0	1	0.664167	0.614925	0.561667	0.284829	1198	5172	6370	Yes
474	473	########	2	1	4	0	2	1	1	0.608333	0.598487	0.390417	0.273629	989	5702	6691	Yes
475	474	########	2	1	4	0	3	1	2	0.463333	0.457038	0.569167	0.167912	347	4020	4367	No
476	475	########	2	1	4	0	4	1	1	0.498333	0.493046	0.6125	0.065929	846	5719	6565	Yes
477	476	########	2	1	4	0	5	1	1	0.526667	0.515775	0.694583	0.149871	1340	5950	7290	Yes
478	477	########	2	1	4	0	6	0	1	0.57	0.542921	0.682917	0.283587	2541	4083	6624	Yes
479	478	########	2	1	4	0	0	0	3	0.396667	0.389504	0.835417	0.344546	120	907	1027	No
480	479	########	2	1	4	0	1	1	2	0.321667	0.301125	0.766667	0.303496	195	3019	3214	No

Figure 2Snapshot of the dataset with Class label Profitable

6. Implementation

The implementation is done using Weka and R.

6.1 Plot the relationship between key weather elements with respect to profitability and provide insight into the trends

Key weather elements are plotted against count to gain insight into individual elements contributions' towards the day being profitable.

The below plots are from using Weka and the legend being:

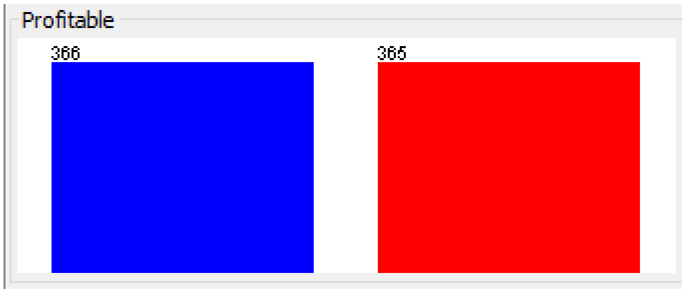


Figure 3 Profitability graph

Season :

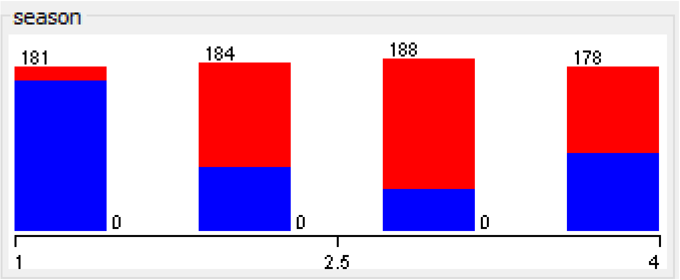


Figure 4 Season vs Profitability Graph

This plot reveals that the profitability is very low during spring. Improves during summer. There is a slight reduction during fall and again rises during winter.

Temperature:

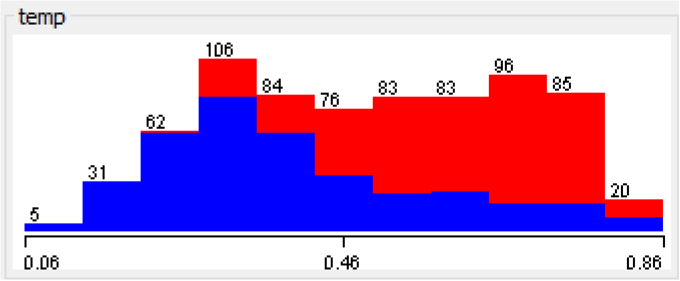


Figure 5 Temperature vs Profitability

The plot provides information that when temperature is between 18.86 degree centigrade and 31.98 degree Celsius the profitability is high. Profitability being highest when the temperature of 26.24 degree Celsius.

Humidity:

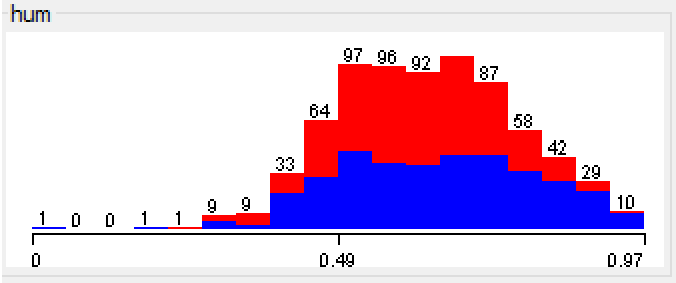


Figure 6 Humidity vs Profitability

The plot suggests that the profitability is maximum when humidity is between 48% and 59%.

Windspeed:

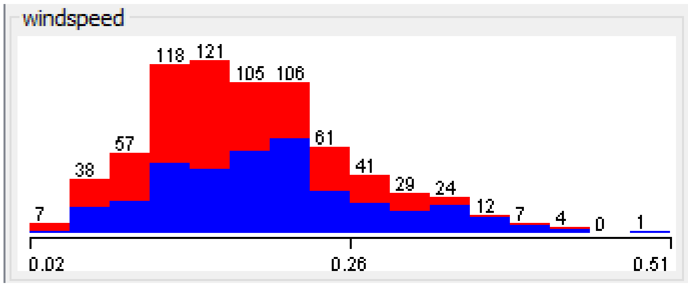


Figure 7 Windspeed vs Profitability

The plot reveals that when windspeed is between 7.5 and 11.65 mph, the profitability is highest.

All these weather elements independently provide insight into predicting whether the day is profitable or not. But, Since the weather of a particular day is determined by all these elements in combination, The profitability is better measured as a combination of these elements.

6.2 Design a decision tree to classify whether a day is profitable or not based upon weather elements

In order to construct a decision tree/perform classification analysis on the data, A class label of profitable is incorporated into the dataset as mentioned above in section 5.1.

The class label was assigned using commands in R. Classification tree was built using the class label as target field and combination of weather elements.

```
>install.packages("tree")   # Install the tree Package

>library(tree)  # open tree library
```

```
>day.tree<- read.csv("day.csv")  # read the csv file into day.tree frame

>profitable<- ifelse(day.tree$cnt<=4548,"No","Yes")  # assign class label condition

> data.tree2 <- data.frame(day.tree,profitable)  # introduce class label into data set

> data.tree3<- tree(formula = profitable~
season+holiday+weekday+workingday+weathersit+temp+atemp+hum+windspeed,data =
data.tree2)   # construct the tree using target field and independent variables

>summary(data.tree3)  # summary of the tree

Classification tree:

tree(formula = profitable ~ season + holiday + weekday + workingday +

weathersit + temp + atemp + hum + windspeed, data = data.tree2)

Variables actually used in tree construction:

[1] "temp"     "season"    "hum"      "atemp"     "windspeed"

Number of terminal nodes:  12

Residual mean deviance:  0.7483 = 538 / 719

Misclassification error rate: 0.1696 = 124 / 731

>plot(data.tree3)  # plot the tree

>text(data.tree3)  # label the branches of the tree
```

Snapshot of the plot using R

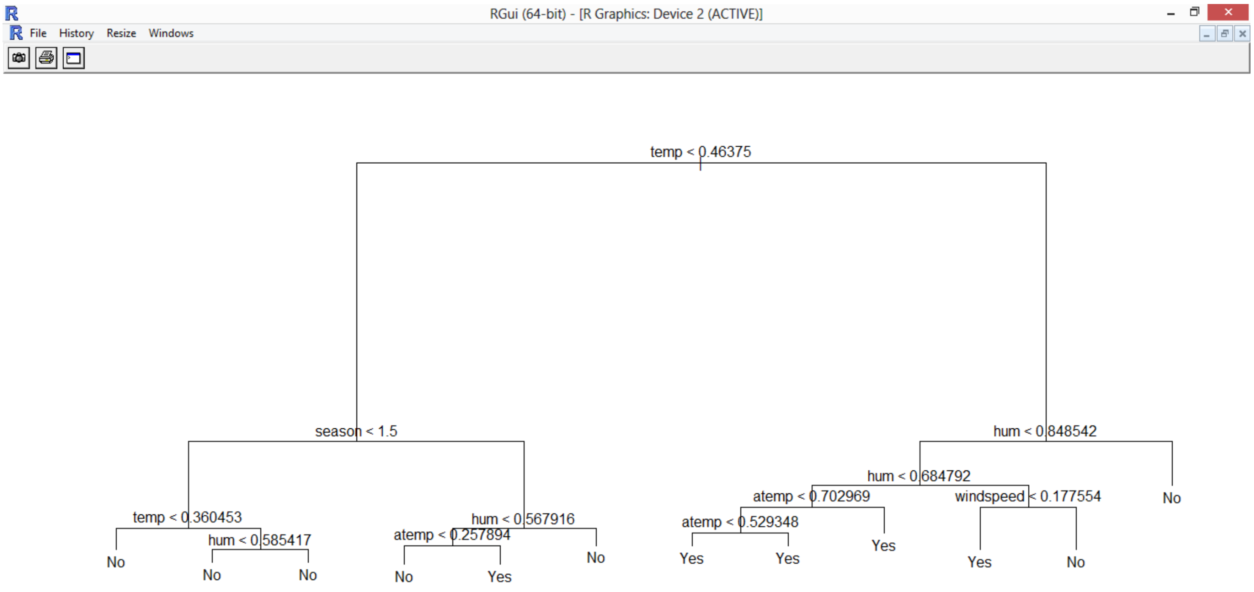


Figure 8 Classification-Tree Representation in R

```
> day.treesnip <- snip.tree(data.tree3, nodes=c(4,12) ) # prune the tree

> plot(day.treesnip)  # plot pruned tree

> text(day.treesnip) # label pruned tree
```

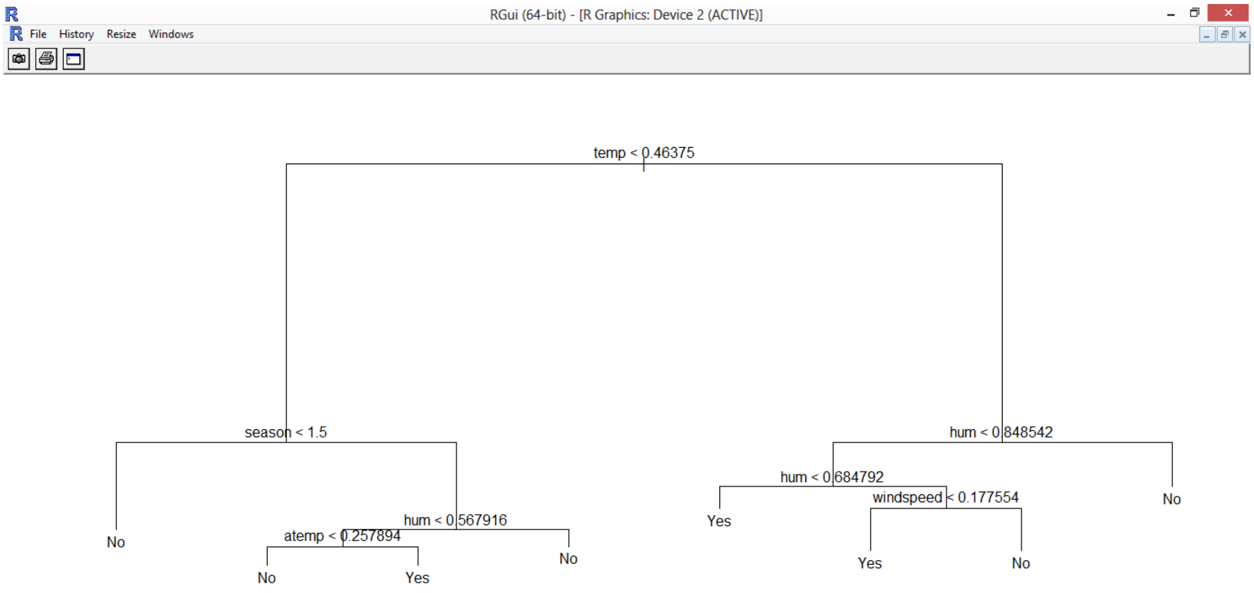
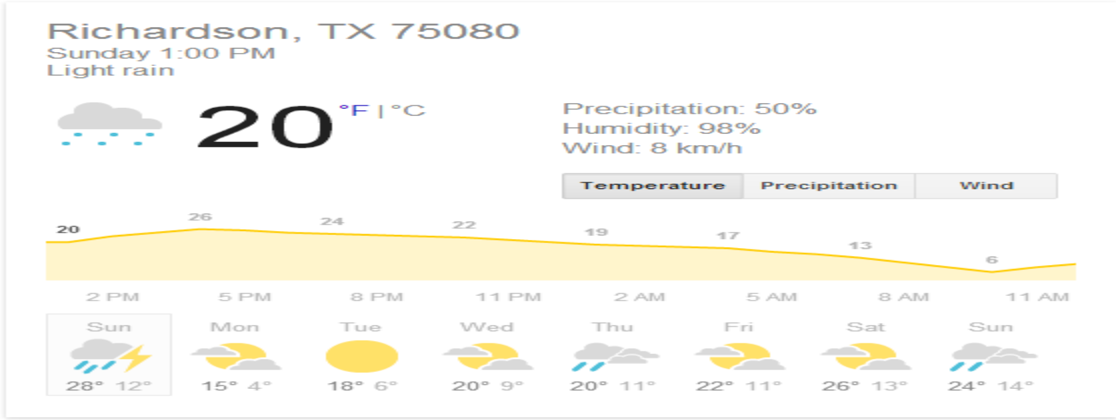


Figure 9 Classification-Pruned Tree(based on temp)



Richardson, TX 75080
Sunday 1:00 PM
Light rain

20 °F | °C

Precipitation: 50%
Humidity: 98%
Wind: 8 km/h

The snapshot provides weather in Richardson for 13th April 2014.

cross validating the pruned tree, temp >0.46375 and humidity >0.848542, I arrived at a class label of NO or Not profitable. Since Richardson experienced showers, It is highly unlikely that the bike rental company made profit on the day. This cross validation further strengthens the validity of the decision tree constructed.

6.3 Build models and select the optimum model which effectively predicts probability of a day being profitable

```
> day.file<- read.csv("C:/Users/Abhay S Joshi/Desktop/day.csv")  # read in the csv file

> day.model <- glm(Profitable ~season+ temp+ hum+windspeed,family = binomial,data = day.file) # command to create the best model for logistic regression based on recurring trials

> summary(day.model)  # summary of the model

Call:

glm(formula = Profitable ~ season + temp + hum + windspeed, family = binomial,

    data = day.file)

Deviance Residuals:

    Min       1Q   Median       3Q      Max

-2.8198  -0.7028  -0.1294   0.7679   2.1869
```

```
Coefficients:

               Estimate Std.     Error        z value      Pr(>|z|)

(Intercept)     -0.66410         0.62077       -1.070      0.285

season           0.50995         0.09028       5.649       1.62e-08 ***

temp            7.70032          0.65358       11.782       < 2e-16 ***

hum             -5.52919         0.76600       -7.218      5.27e-13 ***

windspeed       -5.20813         1.30800        -3.982     6.84e-05 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1013.38  on 730  degrees of freedom

Residual deviance:  693.66  on 726  degrees of freedom

AIC: 703.66

Number of Fisher Scoring iterations: 5
```

Substituting values from instant 1 where count of bike rentals is 985 and hence not profitable, Into the model :

```
>sig.model = predict.glm(day.model,data.frame(season = 1,hum = 0.805833,windspeed = 0.160446, temp = 0.344167),type = "response",se.fit = T)

>sig.model$fit

         1

0.05758544
```

Therefore, this day has 0.05 probability of being profitable based upon the model.

Substituting values from instant 560 where count of bike rentals is 7499 and hence profitable, Into the model :

```
>sig.model = predict.glm(day.model,data.frame(season = 3,hum = 0.485833,windspeed = 0.08085, temp = 0.731667),type = "response",se.fit = T)

>sig.model$fit

     1

0.967469
```

Therefore, This day has 0.96 probability of being profitable based upon the model.

This substitution of values validates the model constructed.

7. Conclusion

From the analysis I can conclude that the probability of a day being profitable depends significantly on the windspeed, temperature, humidity. They are also the factors which determine the season.

8. Business Solution Proposed

In order to improve business during non profitable weather conditions as suggested by the model, I suggest organizing events and catering first hand to events such as

- Indoor Kids Bike training
- Indoor Track biking
- Indoor Dirt biking

During these events bike marts can also concentrate on selling bike accessories like gloves, helmets, riding goggles, jerseys etc, beverages, food and providing bike health checkups which will help improve their business and also act as marketing event to promote their bike mart brand.

To Further enhance profitability during profitable weather conditions, I suggest :

- Kids mountain biking
- women's bike events
- Identifying camping groups and organizing biker trail events during such camps

9.Table of Figures :