Bike Renting

Pesala Ravi Kumar

14/08/2019

Contents

1. Introduction 2

1.1 Problem Statement 2

1.2 Data 2

2. Methodology 3

2.1 Pre Processing 3

2.1.1 Outlier Analysis 3

2.1.2 Missing Value Analysis 4

2.1.3 Feature Selection 4

2.2 Modeling 5

2.2.1 Model Selection 5

2.2.2 Decision Tree Classification 7

3. Conclusion & Future Work 10

3.1 Model Evaluation 10

3.1.1 Confusion Matrix 10

3.1.2 Accuracy 10

3.1.2 Precision 10

3.1.2 Recall or Sensitivity 10

Appendix 11

R code 11

Python code 17

**Chapter 2**

**Introduction**

* 1. **Problem Statement**

We need to identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted. Objective of this project is to predict the customer’s future transaction without prior knowledge of customer transaction amount. So this project analyze the past transactions of customer in order to predict the future transaction.

**1.2 Data**

As you can see in the fig.1.2.1 above we have the following 200 columns of customer transactions out of which we need to predict the column “target”.

**Chapter 2**

**Methodology**

**2.1 Pre Processing**

Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis.

**2.1.1 Outlier Analysis**

Outlier, it is an observation which inconsistence to rest of data.

Causes of the outlier are,

* 1. Poor data quality or contamination.
  2. Low quality measurements.
  3. Manual error.
  4. Malfunctioning equipment.
  5. Correct but exceptional data

Effect of outliers is that it will gives data which is not present in data because of the outlier.

Assume data as 1, 3, 5, 7, and 14 now try to impute using mean value. Mean will be 6 which is not present under data so it might reflect the modelling.

Steps to detect an outlier are,

1. Detect variable with outlier using graphical tools
2. Replace all with NA
3. Apply the missing value analysis on NA records to impute new Values.

**2.1.2 Missing Value Analysis**

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

Why missing the value?

Human Error, refuse to answer, optional

In order to handle Missing Values, there are 2 cases Ignore or Impute Missing value.

Before Imputing Understand why value is missing and by plotting graphs.

Delete the observations where you not to impute.

Techniques to impute missing values:

1. Fill with central statistics like
   * 1. Mean
     2. Mode
     3. Median
2. Distance base/ Data Mining Method – KNN Imputation

While imputing using KNN method there may be a chance of getting an error: Not sufficient complete cases for computing neighbors.

In that case we can go for Central Statistics to impute missing values.

1. Prediction method – Machine Learning models.

Selecting the technique for missing values

1. Create a small subset of total data.
2. Delete some values manually.
3. Use multiple methods to fill.
4. See which technique fills correctly.
5. Select that technique for finding missing value analysis.

**2.1.3 Feature Selection**

It is also called as variable selection or attribute selection.

Selecting a subset of relevant features like variables, predictors for model construction.

Advantages of Feature selection is Dimensionality Reduction (Variable reduction).

Techniques to dimensionality reduction,

1. Correlation analysis.
2. Chi-square test of independence.

**2.2 Modeling**

**2.2.1 Model Selection**

You always start your model building from the simplest to more complex so after applying all the models select the best model according to Confusion Matrix, Accuracy and MAPE and etc.

As our data representing target variable is a binary data type variable. So we’ll go for classification algorithms in order to detect the target variable data with respect to customer transaction amount.

**2.2.2 Decision Tree Classification**

Decision Tree builds classification and regression models in the forms of a tree structure. It breaks down dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numeric data.

A decision tree can easily transformed to a set of rules by mapping from the root node to the leaf nodes one by one.

So when does Decision Tree classifier terminate?

1. Either it divided into classes that are pure (only containing members of single class).
2. Some criteria of classifier attributes are met.

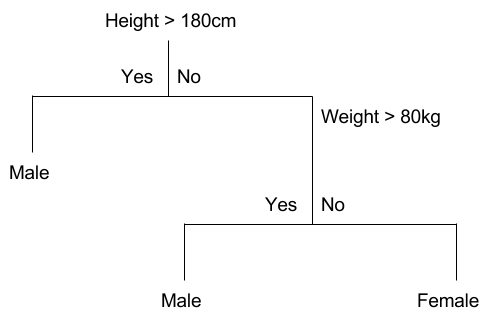
**There are two main types of Decision Trees:**

1. Classification Trees.
2. Regression Trees.

**Classification trees (Yes/No, 1/0 types):**

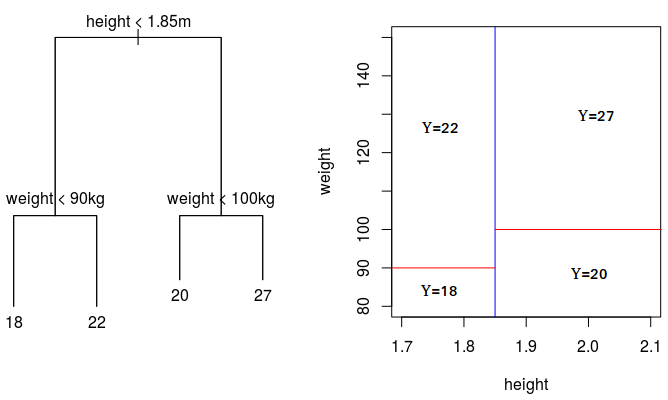
What we’ve seen above is an example of classification tree, where the outcome was a variable like ‘fit’ or ‘unfit’. Here the decision variable is **Categorical/ discrete**.

Such a tree is built through a process known as **binary recursive partitioning**. This is an iterative process of **splitting the data into partitions**, and then splitting it up further on each of the branches.



**Regression trees (Continuous data types):**

Decision trees where the target variable can take **continuous values** (typically real numbers) are called **regression trees**. (E.g. the price of a house, or a patient’s length of stay in a hospital)



## ****Creation of Decision Tree:****

In this method a set of training examples is broken down into smaller and smaller subsets while at the same time an associated decision tree get incrementally developed. At the end of the learning process, a decision tree covering the training set is returned.

The key idea is to use a decision tree to partition the data space into cluster (or dense) regions and empty (or sparse) regions.

In Decision Tree Classification a new example is classified by submitting it to a series of tests that determine the class label of the example. These tests are organized in a hierarchical structure called a decision tree. Decision Trees follow Divide-and-Conquer Algorithm.

## ****Divide and Conquer****



Decision trees are built using a heuristic called **recursive partitioning**. This approach is also commonly known as **divide and conquer** because it splits the data into subsets, which are then split repeatedly into even **smaller subsets**, and so on and so forth until the process stops when the algorithm determines the data within the subsets are **sufficiently homogenous**, or another stopping criterion has been met.

## ****Basic Divide-and-Conquer Algorithm:****

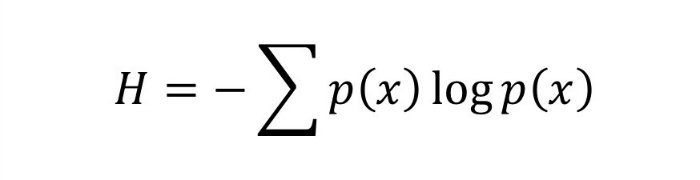
1. Select a test for root node. Create branch for each possible outcome of the test.
2. Split instances into subsets. One for each branch extending from the node.
3. Repeat recursively for each branch, using only instances that reach the branch.
4. Stop recursion for a branch if all its instances have the same class.

## ****Decision Tree Classifier****

* Using the decision algorithm, we start at the tree root and split the data on the feature that results in the **largest information gain (IG)** (reduction in uncertainty towards the final decision).
* In an iterative process, we can then repeat this splitting procedure at each child node **until the leaves are pure**. This means that the samples at each leaf node all belong to the same class.
* In practice, we may set a **limit on the depth of the tree to prevent overfitting**. We compromise on purity here somewhat as the final leaves may still have some impurity.

**Entropy**

Entropy is degree of randomness of elements or in other words it is ***measure of impurity.***Mathematically, it can be calculated with the help of probability of the items as:



P(x) is probability of item x.

It is negative summation of probability times the log of probability of item x.

**Information Gain**

Suppose we have multiple features to divide the current working set. What feature should we select for division? Perhaps one that gives us less impurity.

Suppose we divide the classes into multiple branches as follows, the information gain at any node is defined as

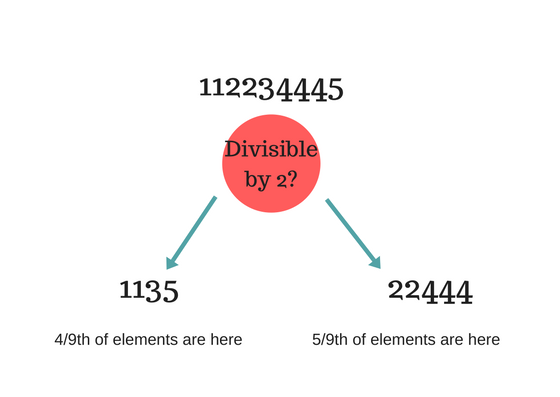
Information Gain (n) =  
 Entropy(x) — ([weighted average] \* entropy (children for feature))

This need a bit explanation!

Suppose we have following class to work with initially

112234445

Suppose we divide them based on property: divisible by 2



Entropy at root level: 0.66  
Entropy of left child: 0.45, weighted value = (4/9) \* 0.45 = 0.2  
Entropy of right child: 0.29, weighted value = (5/9) \* 0.29 = 0.16

**Information Gain** = 0.66 - [0.2 + 0.16] = ***0.3***

Check what information gain we get if we take decision as**prime number instead of divide by 2.**Which one is better for this case?

Decision tree at every stage selects the one that gives best information gain. ***When information gain is 0 means the feature does not divide the working set at all.***

Dividing efficiently based on maximum information gain is key to decision tree classifier. However, in real world with millions of data dividing into pure class in practically not feasible (it may take longer training time) and so we stop at points in nodes of tree when fulfilled with certain parameters.

## ****Advantages of Classification with Decision Trees:****

1. Inexpensive to construct.
2. Extremely fast at classifying unknown records.
3. Easy to interpret for small-sized trees
4. Accuracy comparable to other classification techniques for many simple data sets.
5. Excludes unimportant features.

## ****Disadvantages of Classification with Decision Trees:****

1. Easy to over fit.
2. Decision Boundary restricted to being parallel to attribute axes.
3. Decision tree models are often biased toward splits on features having a large number of levels.
4. Small changes in the training data can result in large changes to decision logic.
5. Large trees can be difficult to interpret and the decisions they make may seem counter intuitive.

## ****Applications of Decision trees in real life:****

1. Biomedical Engineering (decision trees for identifying features to be used in implantable devices).
2. Financial analysis (Customer Satisfaction with a product or service).
3. Astronomy (classify galaxies).
4. System Control.
5. Manufacturing and Production (Quality control, Semiconductor manufacturing, etc.).
6. Medicines (diagnosis, cardiology, psychiatry).
7. Physics (Particle detection).

**Chapter 3**

**Conclusion**

**3.1 Model Evaluation**

The idea of building machine learning models works on a constructive feedback principle. You build a model, get feedback from metrics, make improvements and continue until you achieve a desirable accuracy. Evaluation metrics explain the performance of a model. An important aspect of evaluation metrics is their capability to discriminate among model results.

In our industry, we consider different kinds of metrics to evaluate our models. The choice of metric completely depends on the type of model and the implementation plan of the model.

When we talk about predictive models, we are talking either about a regression model (continuous output) or a classification model (nominal or binary output). The evaluation metrics used in each of these models are different.

In classification problems, we use two types of algorithms (dependent on the kind of output it creates):

1. **Class output**: Algorithms like Decision Tree, SVM and KNN create a class output. For instance, in a binary classification problem, the outputs will be either 0 or 1. However, today we have algorithms which can convert these class outputs to probability. But these algorithms are not well accepted by the statistics community.
2. **Probability output**: Algorithms like Logistic Regression, Random Forest, Gradient Boosting, Adaboost etc. give probability outputs. Converting probability outputs to class output is just a matter of creating a threshold probability.

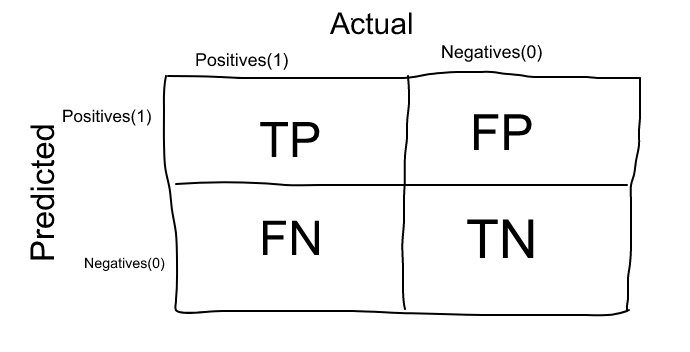
In regression problems, we do not have such inconsistencies in output. The output is always continuous in nature and requires no further treatment.

**3.1.1 Confusion Matrix**

The Confusion matrix is one of the most intuitive and easiest (unless of course, you are not confused) metrics used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes.

Before diving into what the confusion matrix is all about and what it conveys, Let’s say we are solving a classification problem where we are predicting whether a person is having cancer or not.

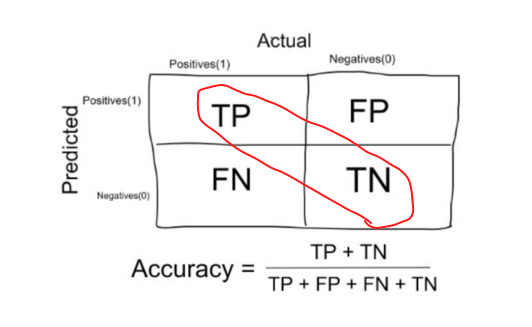
Alright! Now that we have identified the problem, the confusion matrix, is a table with two dimensions (“Actual” and “Predicted”), and sets of “classes” in both dimensions. Our Actual classifications are columns and Predicted ones are Rows.



The Confusion matrix in itself is not a performance measure as such, but almost all of the performance metrics are based on Confusion Matrix and the numbers inside it.

**3.1.2 Accuracy**

Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.



In the Numerator, are our correct predictions (True positives and True Negatives)(Marked as red in the fig above) and in the denominator, are the kind of all predictions made by the algorithm(Right as well as wrong ones).

**When to use Accuracy:**

Accuracy is a good measure when the target variable classes in the data are nearly balanced.

Ex: 60% classes in our fruits images data are apple and 40% are oranges.

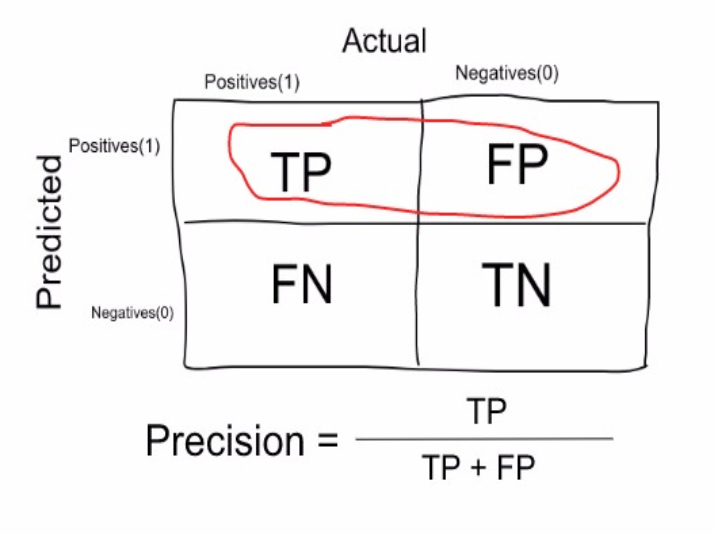
A model which predicts whether a new image is Apple or an Orange, 97% of times correctly is a very good measure in this example.

**When NOT to use Accuracy:**

Accuracy should NEVER be used as a measure when the target variable classes in the data are a majority of one class.

Ex: In our cancer detection example with 100 people, only 5 people has cancer. Let’s say our model is very bad and predicts every case as No Cancer. In doing so, it has classified those 95 non-cancer patients correctly and 5 cancerous patients as Non-cancerous. Now even though the model is terrible at predicting cancer, the accuracy of such a bad model is also 95%.

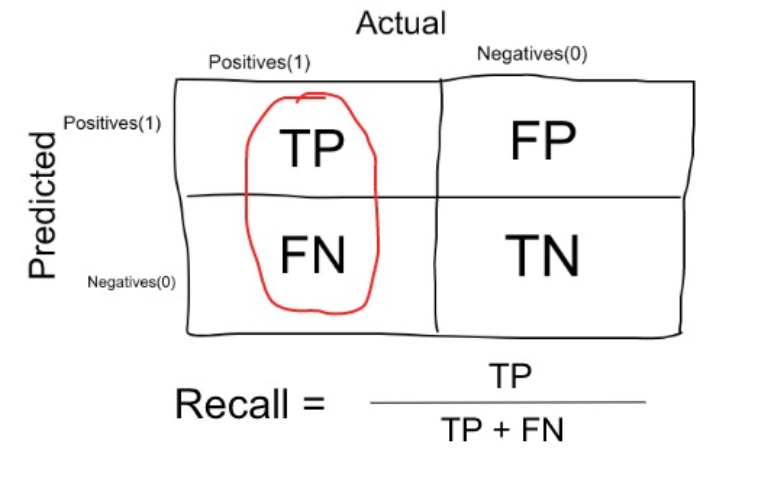
**3.1.2 Precision**

Precision is a measure that tells us what proportion of patients that we diagnosed as having cancer, actually had cancer. The predicted positives (People predicted as cancerous are TP and FP) and the people actually having a cancer are TP.

Ex: In our cancer example with 100 people, only 5 people have cancer. Let’s say our model is very bad and predicts every case as **Cancer**. Since we are predicting everyone as having cancer, our denominator (True positives and False Positives) is 100 and the numerator, person having cancer and the model predicting his case as cancer is 5. So in this example, we can say that **Precision** of such model is 5%.

**3.1.2 Recall or Sensitivity**

Recall is a measure that tells us what proportion of patients that actually had cancer was diagnosed by the algorithm as having cancer. The actual positives (People having cancer are TP and FN) and the people diagnosed by the model having a cancer are TP. (Note: FN is included because the Person actually had a cancer even though the model predicted otherwise).



Ex: In our cancer example with 100 people, 5 people actually have cancer. Let’s say that the model predicts every case as cancer.

So our denominator (True positives and False Negatives) is 5 and the numerator, person having cancer and the model predicting his case as cancer is also 5(Since we predicted 5 cancer cases correctly). So in this example, we can say that the **Recall** of such model is 100%. And Precision of such a model (As we saw above) is 5%

**When to use Precision and When to use Recall?**

It is clear that recall gives us information about a classifier’s performance with respect to false negatives (how many did we miss), while precision gives us information about its performance with respect to false positives (how many did we caught).

**Precision**is about being precise. So even if we managed to capture only one cancer case, and we captured it correctly, then we are 100% precise.

**Recall**is not so much about capturing cases correctly but more about capturing all cases that have “cancer” with the answer as “cancer”. So if we simply always say every case as “cancer”, we have 100% recall.

So basically if we want to focus more on minimizing False Negatives, we would want our Recall to be as close to 100% as possible without precision being too bad and if we want to focus on minimizing False positives, then our focus should be to make Precision as close to 100% as possible.

**Future Work**

**Appendix**

**R code:**

rm(list=ls())

install.packages(c("dmm","dplyr","plyr","reshape","ggplot2","data.table","psych","usdm","caret","DMwR","C50"))

#getwd()

setwd("C:/Users/gopin/Documents/R/CustomerTransaction")

data = read.csv("train.csv", header = T)

savedData = data

# Outlier Analysis

data1 = data[!names(data)%in% c("ID\_code")]

for(i in colnames(data1))

{

#print(i)

values = data1[,i][data1[,i]%in%boxplot.stats(data1[,i])$out]

#print(length(values))

data1[,i][data1[,i]%in%values] = NA

}

#Removing outliers using KNN imputation

require(DMwR)

data1 = knnImputation(data1, k=5)

# Not sufficient complete cases for computing neighbors.

# Removing outliers using MEAN

for(i in colnames(data1))

{

data1$i[is.na(data1$i)] = mean(data1$i,na.rm = T)

}

# Check if any NA present or not

for(i in colnames(data1))

{

print(i)

sum(is.na(data1$i))

}

data1$target[is.na(data1$target)] = 1

newData = data1

# Select Dependent & Independent columns

DependentCols = names(data1)%in% c("target")

dependentData = data1[DependentCols]

independentData = data1[!DependentCols]

independentCols = colnames(independentData)

data = data1[!names(data1)%in% c("ID\_code")]

#Check Multicollinearity

install.packages("usdm")

library(usdm)

vif(data[,-201])

vifcor(data[,-201], th=0.9)

# Logistic Regression because of Categorical Target variable

logistic\_model = glm(target~. , data = data)

logistic\_prediction = predict(logistic\_model)

logistic\_prediction = ifelse(logistic\_prediction > 0.5, 1, 0)

# KNN

# data = data1[!names(data1)%in% c("ID\_code")]

trainData\_index = sample(1:nrow(data), 0.8 \* nrow(data), prob = NULL)

trainData = data[trainData\_index,]

testData = data[-trainData\_index,]

# library(class)

# knn\_pred = knn(trainData[,2:200], testData[,2:200], trainData[,1])

#install.packages("C50")

library(C50)

cols = names(trainData)

trainData[cols] = lapply(trainData[cols] , factor)

str(trainData)

C50\_model = C5.0(target~., trainData, trails = 100, rules = TRUE)

write(capture.output(summary(C50\_model)), "C50\_rules.txt")

C50\_prediction = predict(C50\_model, testData[,2:201], type="class")

confusionMatrix = table(target, C50\_prediction)

ConfusionMatrix(confusionMatrix)

TN = confusionMatrix[0,0]

FN = confusionMatrix[1,0]

TP = confusionMatrix[1,1]

FP = confusionMatrix[0,1]

totalObservations = (TN + FN + TP + FP)

# Accuracy

accuracy = (TP + TN)/totalObservations

# Precision

precision = TP/(TP + FP)

# Recall

recall = TP/(TP + FN)

# False Negative Rate

fnRate = FN/(FN + TP)

library(ggplot2)

library(scales)

library(psych)

library(gplots)

# ggplot(aes(x='var\_0', y='target'), data = data) +

# geom\_point(stat = "identity", fill="Blue") +

# theme\_bw()+

# ylab("target") + xlab("var\_0") +

# ggtitle("Scatter plot")

ggplot(data, aes(as.factor("var\_1"), "target")) +

geom\_point() +

labs(y = "target", x = "var\_1");

ggplot(data, aes(as.factor("var\_2"), "target")) +

geom\_point() +

labs(y = "target", x = "var\_2");

ggplot(data, aes(as.factor("var\_3"), "target")) +

geom\_point() +

labs(y = "target", x = "var\_3");

ggplot(data, aes(as.factor("var\_4"), "target")) +

geom\_point() +

labs(y = "target", x = "var\_4");

**Python code:**