Cab Fare Prediction

Date : 27-11-2019

Author : Pritam Sonawane

Introduction

Problem Statement

Data

Methodology

Pre Processing

Model Development

Decision tree for regression

Random Forest for regression

Regression Analysis

Model Evaluation

Introduction

Problem Statement:

You are a cab rental start-up company. You have successfully run the pilot project and now want to launch your cab service across the country. You have collected the historical data from your pilot project and now have a requirement to apply analytics for fare prediction. You need to design a system that predicts the fare amount for a cab ride in the city.

Data-set :

The details of data attributes in the dataset are as follows -

```
In [57]: ###########################################Explore the data##########################################
         ## Read the data
         train_df <- read.csv("./train_cab.csv", stringsAsFactors=FALSE)
         dim(train_df)
         test_df <- read.csv("./test.csv", stringsAsFactors=FALSE)
         # column names
         names(train_df)
         # datatypes
         str(train_df)
```

```
16067   7

'fare_amount'   'pickup_datetime'   'pickup_longitude'   'pickup_latitude'   'dropoff_longitude'   'dropoff_latitude'   'passenger_count'

'data.frame':   16067 obs. of  7 variables:
 $ fare_amount      : chr  "4.5" "16.9" "5.7" "7.7" ...
 $ pickup_datetime  : chr  "2009-06-15 17:26:21 UTC" "2010-01-05 16:52:16 UTC" "2011-08-18 00:35:00 UTC" "2012-0
21 04:30:42 UTC" ...
 $ pickup_longitude : num  -73.8 -74 -74 -74 -74 ...
 $ pickup_latitude  : num  40.7 40.7 40.8 40.7 40.8 ...
 $ dropoff_longitude: num  -73.8 -74 -74 -74 -74 ...
 $ dropoff_latitude : num  40.7 40.8 40.8 40.8 40.8 ...
 $ passenger_count  : num  1 1 2 1 1 1 1 1 1 2 ...
```

From above data we can see fare amount having char datatype also from datetime we can extract year,date and time for model development

As we can see from dataset the target variable contains continuous values so our task here is to build a regression model which will predict fare amount for car.

Here we have independent variables like pickup_datetime, pickup_longitude, pickup_latitude, dropof_longitude, dropof_latitude, passenger_count and fare_amount is a target variable

From datetime we can get year , month , time features which can be useful for model building. Also from extract feature like distance from pickup and dropoff latitude ,longitude respectively.

Methodology

Pre Processing

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.
Steps Involved in Data Preprocessing:

1. Data Cleaning:

This step is important because in most situations data provided by the customer has a bad quality or just cannot be directly fed to some kind of ML model. It includes data type conversion, data validation, handling dates, handling nominal and categorical variables. But in this case dataset variable data type is as follows:

- We need to convert fare amount variable to numeric value and pickup_datetime to datetime
- From pickup_datetime i have extracted date, year, and time variable

The great circle distance or orthodromic distance is the shortest distance between two points on a sphere (or the surface of Earth). In order to use this method, we need to have the coordinates of point A and point B.

Find the value of longitude in radians:

Value of Longitude in Radians, long = Longitude / (180/pi) OR

Value of Longitude in Radians, long = Longitude / 57.29577951

to get the distance between point A and point B use the following formula:

Distance, d = 3963.0 * arccos[(sin(lat1) * sin(lat2)) + cos(lat1) * cos(lat2) * cos(long2 – long1)]

2. Basic-data-exploration:

```
In [188]: train_df.describe()
```

Out[188]:

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	16027.000000	16027.000000	16027.000000	16027.000000	16027.000000	15972.000000
mean	11.269636	-72.472867	39.920276	-72.472420	39.903395	2.624444
std	9.375223	10.544818	6.813129	10.541479	6.170564	60.918805
min	2.500000	-74.438233	-74.006893	-74.227047	-74.006377	0.000000
25%	6.000000	-73.992153	40.734950	-73.991182	40.734732	1.000000
50%	8.500000	-73.981704	40.752615	-73.980182	40.753577	1.000000
75%	12.500000	-73.966848	40.767366	-73.963666	40.768008	2.000000
max	96.000000	40.766125	401.083332	40.802437	41.366138	5345.000000

The results show 8 numbers for each column in your original dataset. The first number, the **count**, shows how many rows have non-missing values.

The second value is the **mean**, which is the average. Under that, **std** is the standard deviation, which measures how numerically spread out the values are.

3. Missing value analysis:

The concept of missing values is important to understand in order to successfully manage data. If the missing values are not handled properly then we may end up drawing an inaccurate inference about the data.

We can impute missing values by mean ,median of variable or for categorical variable we use mode.

Here there are some missing values found in fare amount, passenger count ,date, time and year :

```
In [60]: ######################################Missing Values Analysis###############################################

missing_val = data.frame(apply(train_df,2,function(x){sum(is.na(x))}))
missing_val
```

A data.frame: 12 × 1

	apply.train_df..2..function.x...
	<int>
fare_amount	24
pickup_datetime	0
pickup_longitude	0
pickup_latitude	0
dropoff_longitude	0
dropoff_latitude	0
passenger_count	55
pickup_date	1
pickup_mnth	1
pickup_yr	1
pickup_hour	1

To impute missing values in fare_amount I have used mean value imputation technique

For passenger count i used median value and i dropped missing values rows for date,time and year variable

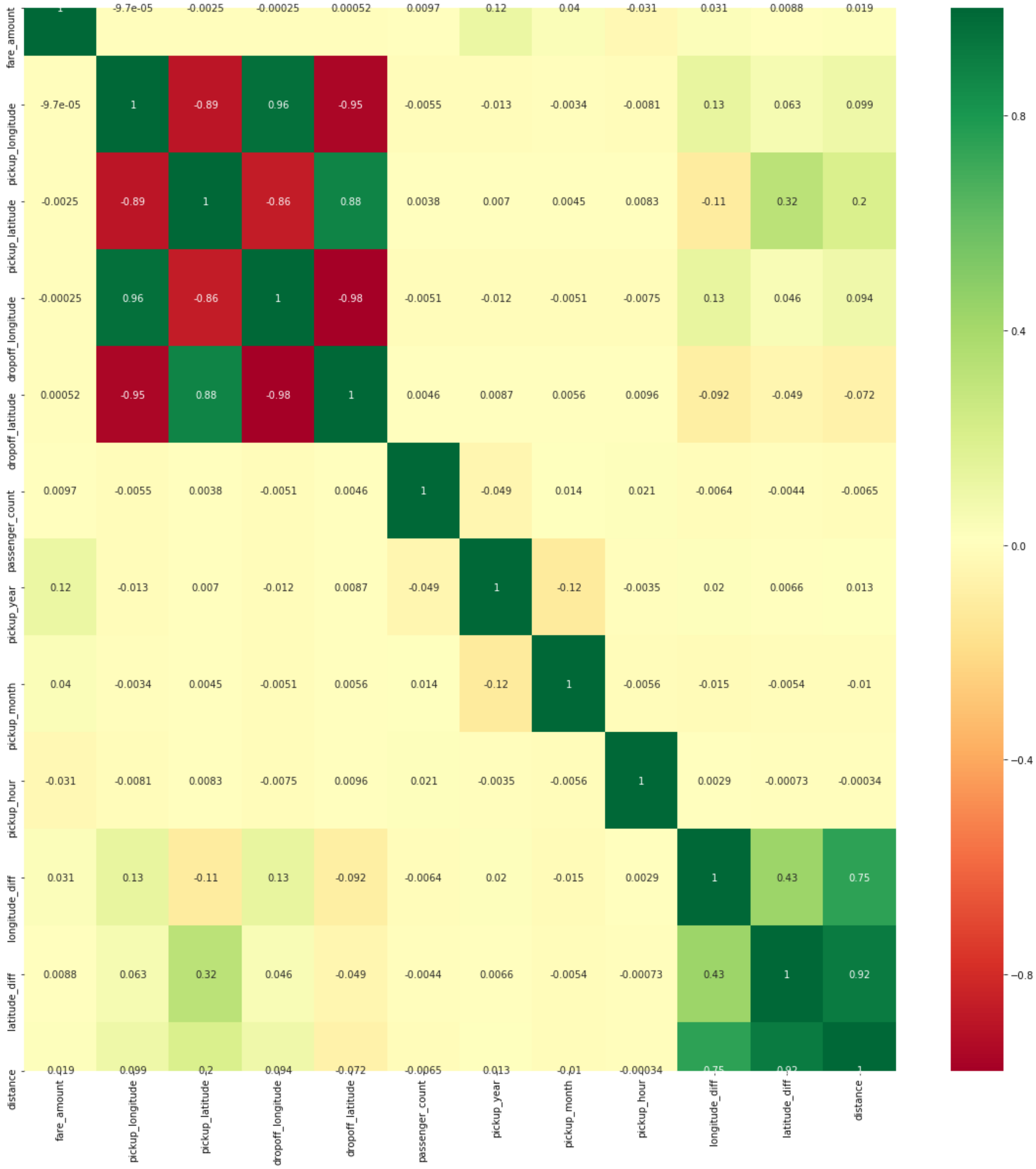
3. Outlier analysis:

An outlier is an element of a data set that distinctly stands out from the rest of the data. It can affect the overall observation made from the data series.

From date description shown above we observed that there is 1 outlier in **pickup_latitude,**

Also fare_amount have -ve values ,zero values which are removed

4. Feature engineering :



From above correlation graph I understand that there are few multicollinearity in the dataset.

Multicollinearity means independent variables are highly correlated to each other

If two variables are correlated it's hard to tell which affect the dependent variable

In feature selection we have dropped pickup_datetime as we have extracted more relevant features from it like year , date, and time

Decision tree for regression

- Here our target variable having continuous values hence we have to use regression model in which by variance we decide best splits
- lower values of variance clearly leading to more pure node and high value of variance lead to impure node
- We will use variance reduction method for node splitting:
  In the anova method
  the splitting criteria is
  SST − (SSL + SSR), where SST = P(yi − y¯)2 is the sum of squares for the node, and SSR, SSL are the sums of squares for the right and left son, respectively.
  This is equivalent to choosing the split to maximize the between-groups sum-of-squares in a simple analysis of variance. This rule is identical to the regression option for tree
- rpart for regression
  rpart(formula, data=, method=,control=) where

  formula : is in the format

  outcome ~ predictor1+predictor2+predictor3+ect.

  data : training dataset

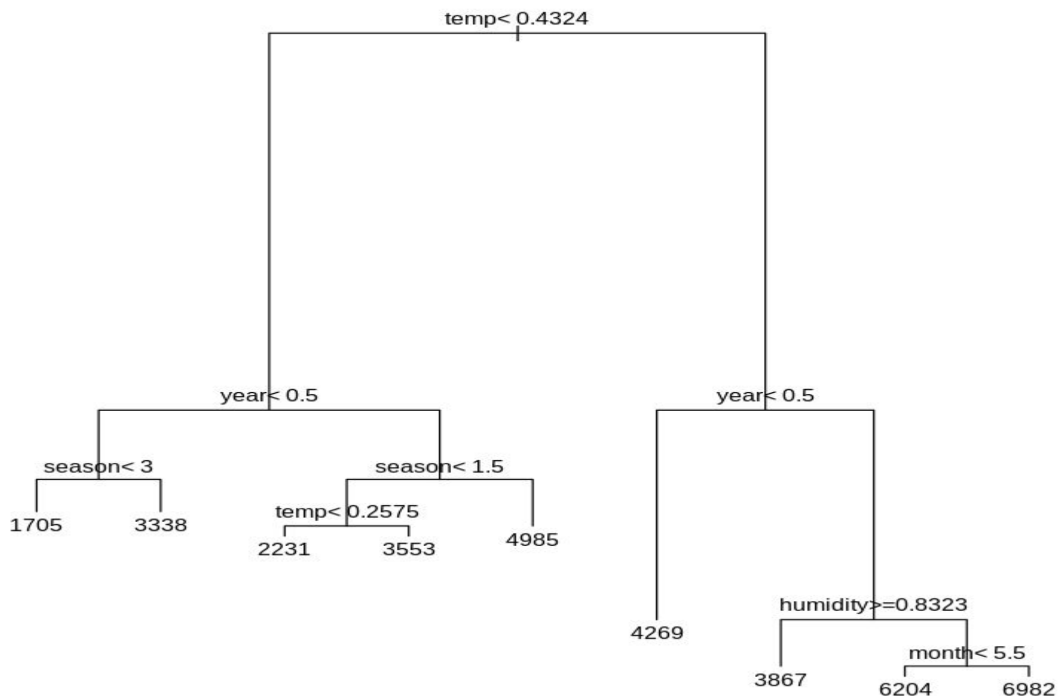
  method :

  "class" for a classification tree

  "anova" for a regression tree

  control : optional parameters for controlling tree growth.

Decision tree model visualisation



Model Evaluation :

Mean Absolute Error (MAE) and Root mean squared error (RMSE) are two of the most common metrics used to measure accuracy for continuous variables.

Mean Absolute Error (MAE): It is the amount of error in your measurements.

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

MAE = 1/n $\sum_{i=0}^{n} yi - yj$

Here, yi=predicted value,yj actual value and n is total set of predictions

Root mean squared error (RMSE): RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It's the square root of the average of squared differences between prediction and actual observation.

RMSE = $\sqrt{1/n \sum_{i=0}^{n}(yi - yj)}$

Both MAE and RMSE express average model prediction error in units of the variable of interest. Both metrics can range from 0 to ∞ and are indifferent to the direction of errors.

Decision tree model

They are negatively-oriented scores: Lower values are better.

- mean absolute percentage error = 0.26600049363014
- accuracy = 73.399950636986 %
- Root mean square error = 5.38910114584255

Multiple Linear Regression

Regression is a parametric technique used to predict continuous (dependent) variable given a set of independent variables. Mathematically, regression uses a linear function to approximate (predict) the dependent variable given as: Y = βo + β1X + ∈ where, Y - Dependent variable X - Independent variable βo - Intercept β1 - Slope ∈ - Error

- βo and β1 are known as coefficients. This is the equation of simple linear regression.
- Error is an inevitable part of the prediction-making process. No matter how powerful the algorithm we choose, there will always remain an (∈) irreducible error The formula to calculate coefficients goes like this: β1 = Σ(xi - xmean)(yi-ymean)/ Σ (xi - xmean)^2 where i= 1 to n (no. of obs.)

βo = ymean - β1(xmean)

In [103]:
```r
##################################################Regression Analysis###########################################
#the base function lm is used for regression.
regmodel <- lm(fare_amount ~ ., data = train)
summary(regmodel)
```

```
Call:
lm(formula = fare_amount ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.943  -5.138  -2.739   1.308  83.462 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       8.595e+01  1.366e+02   0.629 0.529091    
pickup_longitude  -2.348e-02  1.037e-01  -0.226 0.820880    
dropoff_longitude -6.864e-03  9.630e-02  -0.071 0.943180    
pickup_latitude   -4.199e-02  1.865e-01  -0.225 0.821885    
dropoff_latitude  -1.769e-02  1.740e-01  -0.102 0.919033    
passenger_count    8.068e-02  6.517e-02   1.238 0.215733    
pickup_date       -5.336e-03  9.576e-03  -0.557 0.577360    
pickup_mnth02      5.288e-01  4.875e-01   1.085 0.278130    
pickup_mnth03      9.122e-01  6.834e-01   1.335 0.181977    
pickup_mnth04      1.695e+00  9.405e-01   1.803 0.071477 .  
pickup_mnth05      1.698e+00  1.210e+00   1.403 0.160698    
pickup_mnth06      1.386e+00  1.495e+00   0.928 0.353592    
pickup_mnth07      1.460e+00  1.785e+00   0.818 0.413624    
```

```
pickup_mnth08      2.926e+00  2.069e+00   1.414 0.157385    
pickup_mnth09      2.993e+00  2.361e+00   1.267 0.205016    
pickup_mnth10      3.585e+00  2.643e+00   1.356 0.174978    
pickup_mnth11      3.206e+00  2.930e+00   1.094 0.273898    
pickup_mnth12      3.297e+00  3.216e+00   1.025 0.305295    
pickup_yr2010      1.572e+00  3.508e+00   0.448 0.654047    
pickup_yr2011      3.766e+00  6.992e+00   0.539 0.590145    
pickup_yr2012      6.990e+00  1.049e+01   0.666 0.505342    
pickup_yr2013      9.884e+00  1.399e+01   0.707 0.479872    
pickup_yr2014      1.242e+01  1.748e+01   0.711 0.477380    
pickup_yr2015      1.467e+01  2.098e+01   0.699 0.484449    
pickup_hour       -4.172e-02  1.268e-02  -3.289 0.001008 ** 
distance           8.769e-04  2.485e-04   3.529 0.000419 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.359 on 12801 degrees of freedom
Multiple R-squared:  0.0222,	Adjusted R-squared:  0.02029 
F-statistic: 11.62 on 25 and 12801 DF,  p-value: < 2.2e-16
```

- Intercept - This is the βo value. It's the prediction made by model when all the independent variables are set to zero.
- Estimate - This represents regression coefficients for respective variables.
- Std. Error - This determines the level of variability associated with the estimates.
- t value - t statistic is generally used to determine variable significance, i.e.
  if a variable is significantly adding information to the model.
- t value > 2 suggests the variable is significant.

- p value - It's the probability value of respective variables determining their significance in the model.

  p value < 0.05 is always desirable.

  From the above values we can say that there are few values are significant like pickup hour , distance and pickup month

Multiple Linear Regression model

- mean absolute percentage error = 0.509387049648153
- accuracy = 49.0612950351847 %
- Root mean square error = 8.96655073661903.

Random Forest for regression

A random forest allows us to determine the most important predictors across the explanatory variables by generating many decision trees and then ranking the variables by importance.

Random subsets are created from original dataset

At each node in the decision tree only random set of features are considered to decide the best split

The decision tree model is fitted on each set of the subset

The final decision is calculated by averaging the prediction from all decision trees.

```
In [117]: # Number of variables randomly sampled as candidates at each split. Note that the default values are different for
          # where p is number of variables and for regression we take mtry=(p/3)
          ##############################Random Forest model##################################

          rf_2=randomForest(fare_amount ~ . , data = train,mtry =4,ntree=100 ,nodesize =10 ,importance =TRUE)
```

Number of variables randomly sampled as candidates at each split.

From above

mtry :

Number of variables randomly sampled as candidates at each split. Note that the default values are different for classification (sqrt(p) where p is the number of variables in x) and regression (p/3)

ntree : Number of trees to grow.

nodesize : Minimum size of terminal nodes. Setting this number larger causes smaller trees to be grown (and thus take less time). Note that the default values are different for classification (1) and regression (5).

Random Forest model

- mean absolute percentage error = 0.200908952833695
- accuracy = 79.9091047166305 %
- Root mean square error = 4.60001705106375