iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** VIII **Month of publication:** August 2019

DOI: http://doi.org/10.22214/ijraset.2019.8073

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VIII, Aug 2019- Available at www.ijraset.com

Machine Learning using Exploratory Analysis to Predict Taxi Fare

Gunjan Panda[1], Supriya P. Panda[2]
[1]*B.Tech (CSE), Final year student, The North Cap University, Gurgaon,*
[2]*Professor (CSE) Manav Rachna International Institute of Research and Studies, Faridabad*

Abstract: ***Predictive analytics uses archival data to predict the future events. Typically, past data is used to build a mathematical model that captures important trends. That predictive model is then used on current data to predict the future or to suggest actions to take for optimal outcomes. Predictive analytics has received a lot of attention in recent years due to advances in supporting technology, particularly in the areas of big data and machine learning. Companies also use predictive analytics to create more accurate forecasts, such as forecasting the fare amount for a cab ride in the city. These forecasts enable resource planning for instance, scheduling of various cab rentals to be done more effectively. For a cab rental start-up company, the fare amount is dependent on a lot of factors. This research aims to understand all patterns and to apply analytics for fare prediction. The proposed work is to design a system that predicts the fare amount for a cab ride in the city. The aim is to build regression models, which will predict the continuous fare amount for each cab ride and help prediction depending on multiple time-based, positional and general factors.***
Keywords: ***Predictive Analytics, Forecasting, Regression Models,*** *Random Forest,* ***Decision Tree, K-NN***

I. INTRODUCTION

Machine learning (ML) is closely related to computational statistics, which focuses on making predictions using computers. Data mining (DM) is a field of study within ML and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics. Machine learning tasks are classified into several broad categories.

In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. Classification algorithms and regression algorithms are examples of supervised learning Regression algorithms are named for their continuous outputs, meaning they may have any value within a range [1].

In unsupervised learning, the algorithm builds a mathematical model from a set of data that contains only inputs and no desired output labels.

Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points. Unsupervised learning can discover patterns in the data and can group the inputs into categories, as in feature learning. Dimensionality reduction is the process of reducing the number of "features", or inputs, in a set of data.

Machine learning and data mining often employ the same methods and overlap significantly, but while ML focuses on prediction, based on known properties learned from the training data, data mining focuses on the discovery of (previously) unknown properties in the data. This is the analysis step of knowledge discovery in databases (KDD) [1]. DM uses many ML methods, but with different goals; on the other hand, ML also employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy.

A. Supervised Learning Approach

Supervised learning algorithms include classification and regression. Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range.

To build any model, the first step is to recognize the problem statement and choose the appropriate category that fits in. Since the discourse statement “fare amount for a cab ride in the city” fits into forecasting, Regression is used as it helps in predicting continuous fare amount for the future. Regression is a Supervised Learning approach as the target variable, which is fare-amount and is known beforehand. Regression is used as it helps in predicting continuous fare amount for the future.

B. Data

The aim is to build regression models that will predict the continuous fare amount for each of the cab-rides depending on multiple time-based, positional and generic factors. This problem statement falls under the category of forecasting which deals with predicting continuous values for the future (the continuous value is the fare amount of the cab ride).Fig.1 shows a sample of the data set[2] that will be used to predict the fare amount of a cab ride.

fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
4.5	2009-06-15 17:26:21 UTC	-73.844311	40.721319	-73.841610	40.712278	1.0
16.9	2010-01-05 16:52:16 UTC	-74.016048	40.711303	-73.979268	40.782004	1.0
5.7	2011-08-18 00:35:00 UTC	-73.982738	40.761270	-73.991242	40.750562	2.0
7.7	2012-04-21 04:30:42 UTC	-73.987130	40.733143	-73.991567	40.758092	1.0
5.3	2010-03-09 07:51:00 UTC	-73.968095	40.768008	-73.956655	40.783762	1.0

Fig.1: a sample of the data set (Kaggle[2])

There are six predictor variables and one target variable which are listed as follows:

Predictors:

1) Pickup_datetime : timestamp value indicating when the cab ride started.
2) Pickup_longitude: float for longitude coordinate of where the cab ride started.
3) Pickup_latitude: float for latitude coordinate of where the cab ride started.
4) Dropoff_longitude: float for longitude coordinate of where the cab ride ended.
5) Dropoff_latitude: float for latitude coordinate of where the cab ride ended.
6) Passenger_count: an integer indicating the number of passengers in the cab ride.

Target: fare_amount

Data structures upon proper data type conversion are shown in Fig.2:

```
cab_df.dtypes

fare_amount                       float64
pickup_datetime       datetime64[ns, UTC]
pickup_longitude                  float64
pickup_latitude                   float64
dropoff_longitude                 float64
dropoff_latitude                  float64
passenger_count                   float64
dtype: object
```

Fig.2: Data structures following data type conversion

II. METHODOLOGY

One common methodology is the Cross-Industry Standard Process for DM or (CRISP-DM) model [1,2]. This is a five process model that provides a fluid framework for devising, creating, building, testing, and deploying machine learning solutions.

A. Exploratory Data Analysis

After identifying the approach, the next step is preprocessing the data. Looking at data refers to exploring the data, refining the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis (EDA) [3].

To initiate this process, any of the probability distributions of the variables are considered. Most analysis like regression, require the data to be normally distributed. This can be visualized at a glance by looking at the probability distributions or probability density functions of the variable. Fig.3a and Fig.3b plot the probability density functions of few variables available in the data as well as the dependent fare_amount variable. The distribution depicts the skewness of the data points, which indicate the presence of outliers.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VIII, Aug 2019- Available at www.ijraset.com

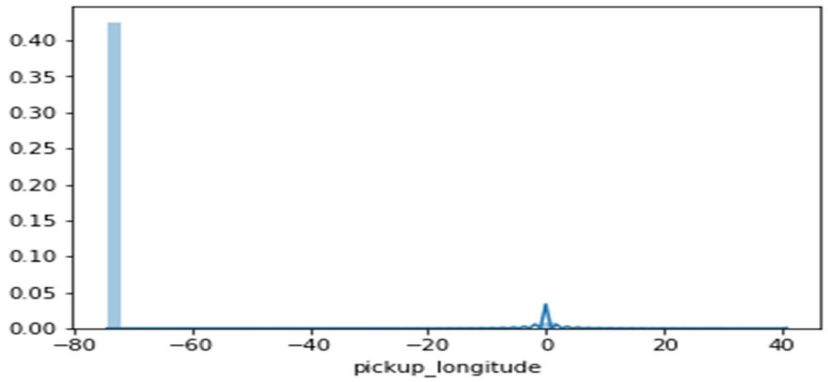


Fig.3a: plotting probability density functions

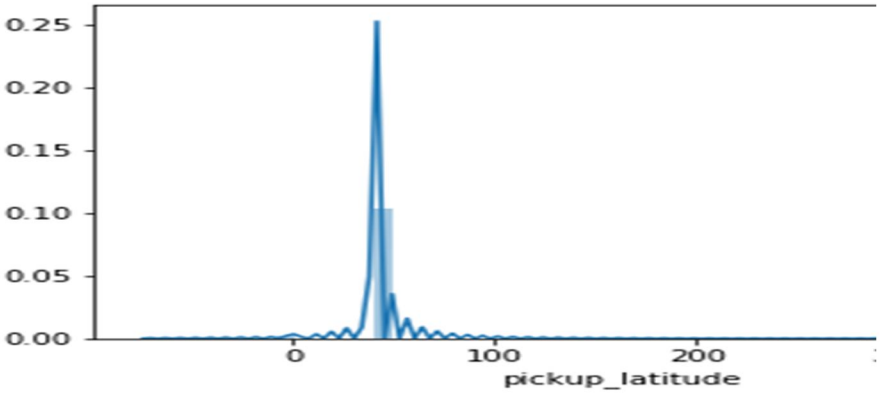


Fig.3b: plotting probability density functions

B. Missing Value Analysis

Once proper data conversion is done next step is to analyze the missing values. As per data, the missing value percentages of the variables are as in Fig.4

variables	missing_percentage
passenger_count	0.342338
fare_amount	0.155608
pickup_datetime	0.000000
pickup_longitude	0.000000
pickup_latitude	0.000000
dropoff_longitude	0.000000
dropoff_latitude	0.000000

Fig.4: missing value percentages of variables

As only passenger_count and fare_amount have missing values and the percentage is less than thirty percent, so the missing values are imputed. On randomly assigning NA to one of these values for passenger_count and then filling the content using the three methods: mean, median and K-NN, it is found that the median gives the closest estimate to this actual value. Similarly, for the fare_amount, mean yields the nearest value to the real value. Hence, the missing values for passenger_count are filled with median and that for the fare_amount are filled with mean. After filling in missing values data looks like as shown in Fig.5.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VIII, Aug 2019- Available at www.ijraset.com

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	16066.000000	16066.000000	16066.000000	16066.000000	16066.000000	16066.000000
mean	15.005185	-72.462693	39.914675	-72.462233	39.897852	2.626478
std	430.139359	10.578707	6.826797	10.575384	6.187276	60.741672
min	-3.000000	-74.438233	-74.006893	-74.429332	-74.006377	0.000000
25%	6.000000	-73.992156	40.734935	-73.991182	40.734647	1.000000
50%	8.500000	-73.981697	40.752605	-73.980170	40.753566	1.000000
75%	12.500000	-73.966837	40.767381	-73.963642	40.768015	2.000000
max	54343.000000	40.766125	401.083332	40.802437	41.366138	5345.000000

Fig.5: filling in missing values

C. Outlier Analysis

As depicted in Fig.4, there are a lot of noisy data so it's important to clean the data for better model performance. In this case, a classic approach, namely Turkey's method is used for removing outliers. We visualize the outliers using boxplots. Fig.6a Fig.6b, Fig.6c, and Fig.6d plot the boxplots of four of the six predictor variables (as a sample) and the target variable. A lot of useful inferences can be made from these plots such as a lot of outliers and extreme values are seen in each of the data sets.

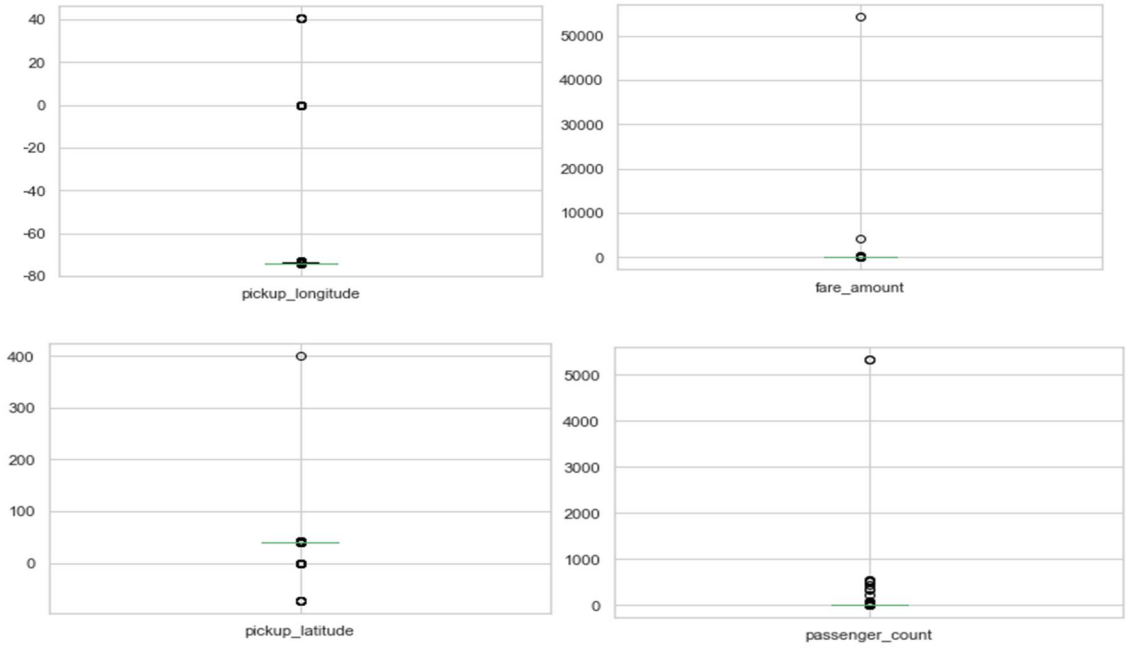


Fig.6a-d: noisy data pre-cleaning

Upon removing the outliers, data is now refined as shown in Fig.7.

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	11824.000000	11824.000000	11824.000000	11824.000000	11824.000000	11824.000000
mean	8.563625	-73.981886	40.752954	-73.981039	40.753368	1.262432
std	3.816751	0.016014	0.021029	0.016500	0.021894	0.544297
min	2.500000	-74.018108	40.693504	-74.019535	40.694260	1.000000
25%	5.700000	-73.992841	40.738610	-73.992088	40.738855	1.000000
50%	7.700000	-73.982749	40.753510	-73.982093	40.754582	1.000000
75%	10.500000	-73.971549	40.766776	-73.970891	40.767358	1.000000
max	22.100000	-73.932999	40.810862	-73.935237	40.811445	3.000000

Fig.7: refined data

D. Feature Selection

Because all the variables are numeric the important features are extracted using the correlation matrix. As is seen from Fig.8, all the variables are important for predicting the fare_amount since none of the variables have a high correlation factor (considering the threshold as 0.9), so all the variables for model building are kept.

Fig.8: the correlation matrix

Another method for feature selection is Random Forest. Fig.9 mentions the process of Random Forest to extract the importance of each variable using R-programming [4].

```
model_rf = randomForest(fare_amount~., train,importance = TRUE, ntree = 300)
importance (model_rf,type = 1)
```

	%IncMSE
pickup_longitude	22.7374655
pickup_latitude	22.5690452
dropoff_longitude	23.4859649
dropoff_latitude	24.9081478
passenger_count	-0.6184569
year	71.9721697
month	9.3666233
day	-0.1673588
weekday	5.7074033
hour	31.4455704
distance	145.7308873

Fig.9: the process of Random Forest

As is seen, distance has the highest prediction power for fare_amount whereas passenger_count and day have the least prediction power.

E. Feature Engineering

It is important to infer some knowledge from the existing data and come up with more valuable information. As the dataset already has datetime variable, further the year, month, day, weekday and hours are calculated that might have an effect on the fare and to further perform some EDA on the data. Also as the longitude, latitude points are there, the distance traveled per ride is easily calculated to derive a relationship between the fare amount and the distance. To calculate the distance Haversine Distance formula is used and the distance in kilometers is found. The Haversine formula [5] calculates the shortest distance between two points on a sphere using their latitudes and longitudes measured along the surface. It is important for use in navigation.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VIII, Aug 2019- Available at www.ijraset.com

III. MODELING

A. *Model Selection*

In the early stages of analysis during pre-processing, it is understood that fare_amount is dependent on multiple behaviors. Therefore, it's important to build a model in such a way that it takes in all the required inputs and fits the model in such a way that it gives the most accurate result amongst all the other models. The dependent variable can fall in any of the four categories: Nominal, Ordinal, Interval, and Ratio. Three approaches are taken and compared:

B. *Decision Tree*

A decision tree is a tree-like graph with nodes representing the place where an attribute is picked and queried; edges represent the answers to the query, and the leaves represent the actual output or class label. Decision trees are nonlinear [6]. Decision Tree algorithms are referred to as Classification and Regression Trees (CART) [7].

Max Depth: larger the dataset harder to visualize so the maximum branching is taken as five, and fit=DecisionTreeRegressor(max_depth=5).fit(train.iloc[:,1:],train.iloc[:,0]).

Herein, the maxDepth is chosen as 5.

C. *Random Forest*

Random forest is a tree-based algorithm, which involves building several trees (decision trees), then combining their output to improve the generalization ability of the model. The method of combining trees is known as an ensemble method. The ensemble is a combination of weak learners (individual trees) to produce a strong learner. Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical.

D. *K-NN*

The K-Nearest Neighbors (K-NN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The K-NN algorithm assumes that similar things exist close. K-NN makes predictions using the training dataset directly. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. For regression this might be the mean output variable, in classification, this might be the mode (or most common) class value. To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance.

IV. MODEL EVALUATION

The quality of a regression model is how well its predictions match up against actual values, and Error metrics are used to judge the quality of a model, which enables us to compare regressions against other regressions with varied parameters [7].

A. *Root Mean Squared Error (RMSE)*

The Root Mean Squared Error (RMSE) and R-Squared are used for dealing with time series forecasting and continuous variables.

The RMSE indicates the absolute fit of the model to the data, whereas R-Squared is a relative measure of fit [1, 8].

RMSE must be compared with the dependent variable as RMSE is in the same units as the dependent variable.

1) *Smaller The Result, Better The Performance Of The Model:* To understand how well the independent variables “explain” the variance in the model, the R-Squared formula is used.
2) *For The R-Squared, The Closer The Value To 1, The Better The Performance Of The Model:* According to the underlying model, Table-1 describes its error metrics.

Table-1

Model	RMSE Score	R Square
Decision Tree	2.1091782572686593	0.6790662606031613
Random Forest	1.9953043471990368	0.7127850099199918
K- NN	2.6025145530297236	0.5297013732475271

V. CONCLUSION

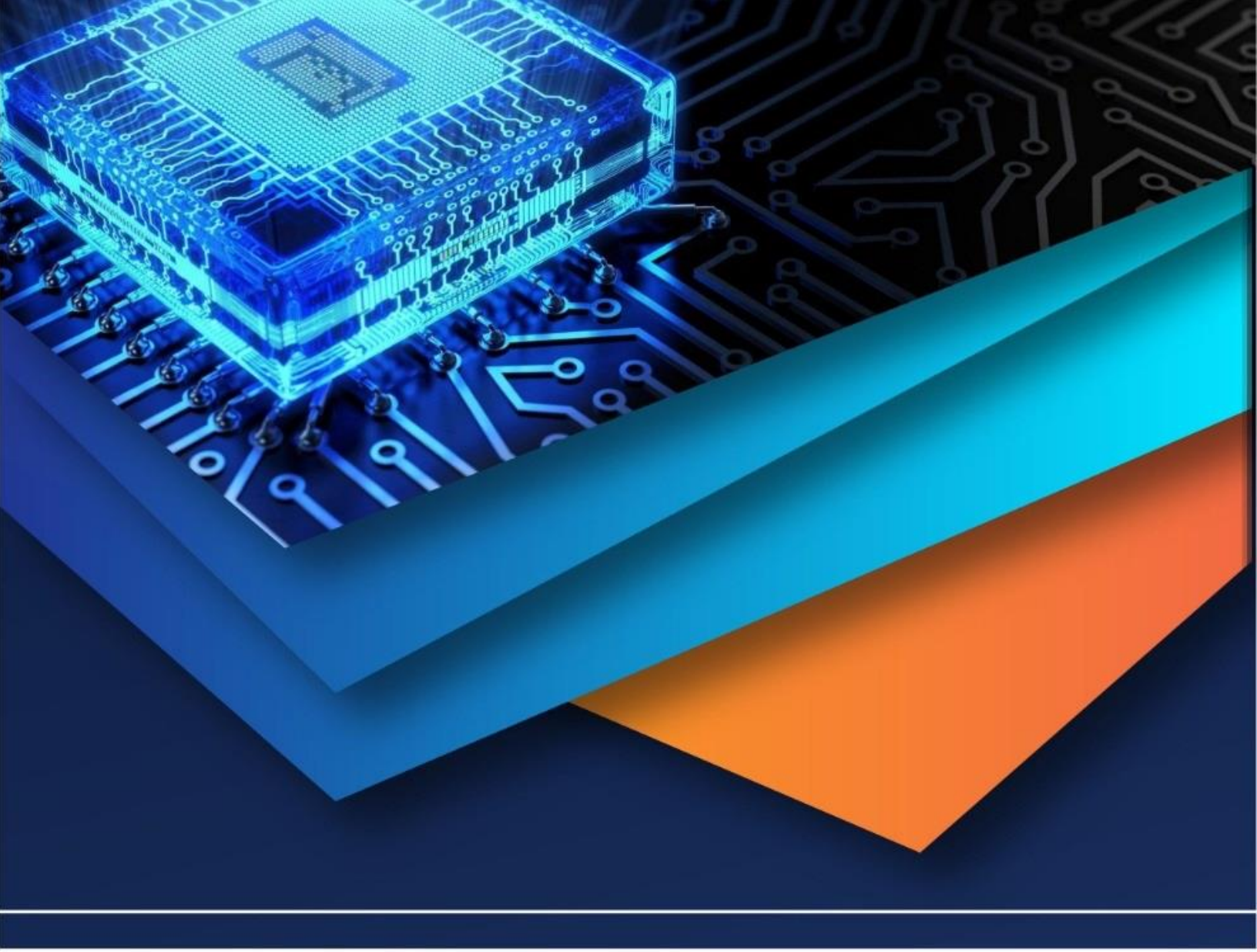
The quality of a regression model depends on the matchup of predictions against actual values. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical. Random Forest can be used to solve both regression and classification problems. The K-NN algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. Decision trees are nonlinear; unlike linear regression, there is no equation to express the relationship between independent and dependent variables. Out of the three models left, Random Forest is the best model as it has the lowest RMSE score and highest R-Squared score, which explains the highest variability and tells us how well the model fits in this data.

VI. FUTURE SCOPE

As is known, with an increase in the number of features; underlying equations become a higher-order polynomial equation, and it leads to overfitting of the data. Generally, it is seen that an overfitted model performs worse on the testing data set, and it is also observed that the overfitted model performs worse on additional new test data set as well. A kind of normalized regression type - Ridge Regression may be further considered.

REFERENCES

[1] John D. Kelleher**,** Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (The MIT Press), 1st Edn.
[2] Data science work with public datasets, [Online] Available: https://www.kaggle.com,
[3] Chong Ho Yu, Exploratory data analysis in the context of data mining and re-sampling, International Journal of psychological research, 2010, vol.3, No.1
[4] Ruben Oliva Ramos, Jen Stirrup, Advanced Analytics with R and Tableau, Packt Publishing, 2017, ISBN: 9781786460110
[5] Machine Learning in Python,[Online] Available https://scikit-learn.org/stable/
[6] T. Divya and A. Sonali, "A survey on Data Mining approaches for Healthcare", International Journal of Biosciences and Bio-Technology, vol. 5, no. 5, (2013), pp. 241-266.
[7] Sebastian Raschka, Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, [Online] Available https://arxiv.org/abs/1811.12808,2016
[8] Weijie Wang and Yanmin Lu, Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model, ICMEMSCE, IOP Publishing, 324 (2018), doi:10.1088/1757-899X/324/1/012049



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY