IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005139|

Car Price Prediction Using Regression Techniques of Machine Learning

Sk.Shammi[1] , A.Shanmukha [2] , K. Swathi[3] , G.Sravya[4] , A,Phani Kumar[5]

Assistant Professor, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India[1]

UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India[2]

UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India[3]

UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India[4]

UG Student, Department of CSE, Tirumala Engineering College, NRT, Andhra Pradesh, India[5]

ABSTRACT: For most of the People buying a car for second hand or third hand may include several researches for an accurate price for that car is more challenging and tricky task . The prices of cars depends on a number of collaborative factors. The factors that effect the price includes year of purchase,fuel type,number of years used ,no.of owners, and many more. The goal of this research paper is to predict the efficient pricing for cars with respect to the people budgets and priorities . By analyzing previous market trends and price ranges, and also upcoming developments future prices will be predicted.The data set is taken from the various machine learning plat forms like Kaggle and UCI. In this study an attempt has been made to construct a machine learning model to evaluate the price based on various factors that affect price This paper presents a comparison of Machine Learning algorithms such as Lasso, Linear Regression, Decision Tree Regressor, Ridge, Ada Boost Regressor, Gradient Boosting Regressor, Random Forest Regressor on the Kaggle Car Price Dataset. Based on the results of the Regression model which gives the highest accuracy will be used for the Car Price Prediction.

KEYWORDS: Carprice, Regression, Linear and Ridge

I.INTRODUCTION

Machine learning is one of the applications of artificial intelligence (AI) that provides computers, the ability to learn automatically and improve from experience instead of explicitly programmed. It focuses on developing computer programs that can access data and use it to learn from themselves. The main aim is to allow computers to learn automatically without human intervention and also adjust actions accordingly. Accurate car price prediction involves expert knowledge, because price usually depends on many distinctive features and factors. Typically, most significant ones are brand and model, age, current price and fuel type. Car prices had a great deal of attention in automobile research. The prediction of car price is a challenging task, which can offer automated prediction about the car.

II. LITERATURE SURVEY

Pow, Nissan, Emil Janulewicz, and L. Liu [11] used four regression techniques namely Linear Regression, Support Vector Machine, K-Nearest Neighbors (KNN)and Random Forest Regression and an ensemble approach by combining KNN and Random Forest Technique for predicting the car price value.The ensemble approach predicted the prices with least error and applying PCA didn't improve the prediction error.

Data mining is widely used in the research field such as prediction of car price it is a multidisciplinary field. Using data mining researchers are developing various techniques in-order to predict the car price prediction with high accuracy.

S.DilliArasu and Dr. R. Thirumalaiselvi has worked on missing values in a dataset of car price prediction . Missing values in dataset will reduce the accuracy of our model as well as prediction results. They find solution over this

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005139|

problem that they performed a recalculation process on CPP stages and by doing so they got up with unknown values. They replaced missing values with recalculated values.

III. PROPOSED SYSTEM

Experimental Setup: This experiment was conducted on Intel® Core™ i7 Processors with 64bit Windows 10 Pro. Anaconda 5.1.0 Python distribution is used in this experiment. The dataset used in thisproject is Car Price Prediction is Kaggle car Price data set which is obtained from kaggle Machine Learning repository [10]. The dataset contains 10 attributes which are used to predict the car price.

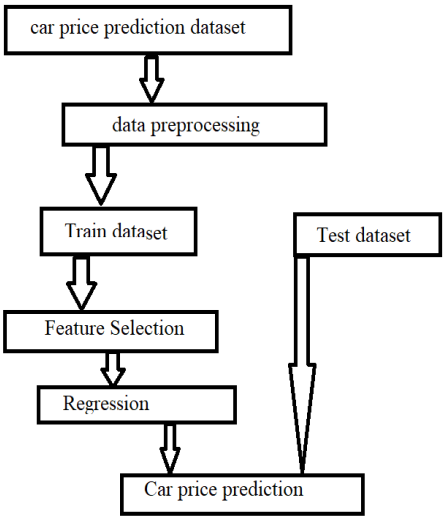

Fig-1: Architecture of Proposed System

IV. DATA DESCRIPTION AND UNDERSTANDING

Data Description: The dataset used for the House Price Prediction which is obtained from Kaggle Machine Learning repository [10]. The dataset contains 9 attributes which are used to predict the price

A. Year-year of the car purchased
B. .Selling_Price-price of the car sold
C. Present_Price-present price of that car
D. Kms_Driven-total number kms that car has driven before selling
E. Fuel_Type-type of the fuel
F. Seller-Type-type of the seller
G. Transmission-Transmission type of gears
H. Owner-no.of owners
I. Current year-present year
J. No_year- No.of years car has been used

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005139|

Car_Name	Year	Selling_Pri	Present_Pr	Kms_Drive	Fuel_Type	Seller_Typ	Transmissi	Owner
ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
swift	2014	4.6	6.87	42450	Diesel	Dealer	Manual	0
vitara brez	2018	9.25	9.83	2071	Diesel	Dealer	Manual	0
ciaz	2015	6.75	8.12	18796	Petrol	Dealer	Manual	0
s cross	2015	6.5	8.61	33429	Diesel	Dealer	Manual	0
ciaz	2016	8.75	8.89	20273	Diesel	Dealer	Manual	0
ciaz	2015	7.45	8.92	42367	Diesel	Dealer	Manual	0
alto 800	2017	2.85	3.6	2135	Petrol	Dealer	Manual	0
ciaz	2015	6.85	10.38	51000	Diesel	Dealer	Manual	0
ciaz	2015	7.5	9.94	15000	Petrol	Dealer	Automatic	0
ertiga	2015	6.1	7.71	26000	Petrol	Dealer	Manual	0
dzire	2009	2.25	7.21	77427	Petrol	Dealer	Manual	0
ertiga	2016	7.75	10.79	43000	Diesel	Dealer	Manual	0
ertiga	2015	7.25	10.79	41678	Diesel	Dealer	Manual	0
ertiga	2016	7.75	10.79	43000	Diesel	Dealer	Manual	0
wagon r	2015	3.25	5.09	35500	CNG	Dealer	Manual	0

Fig-2: Dataset

A.Data Understanding:

The goal is to build a model which predict the price.We offten use correlation and chi2 for feature selection.The data is splitinto 2 parts such as train data. Thedata set has numerical and categorical data .The Regressionmodels should not be trained with categorical data. Using Label Encoder and One Hot Encoding the categorical data is convertedto numerical and then feeded to the model.

B. Data Pre-Processing: Pre-processing is the first step while creating the machine learning model [9]. It is the process of converting raw dataset into cleaned dataset. Raw data contains noise, missing values, duplicate values which is not suitable for machine learning model.

The general steps in data pre-processing are:

- Converting categorical features into numerical variables inorder to fit linear regression model.
- Imputing null records with appropriate values.
- Scaling of data
- Removing Outliers

C. Data Distribution: Before feeding data to an algorithm, we have to apply transformations to our data which is referred as pre-processing. By performing pre- processing, the raw data which is not feasible for analysis is converted into clean data. In-order to achievebetter results using a model in Machine Learning, data format has to be in a proper manner. The data should bein a particular format for different algorithms. For example, if we consider Random Forest algorithm it does not support null values. So that those null values have to be managed using raw data [13].

Data cleaning should be done on missing data and erroneous data. Data cleaning can be done by filling missing values manually or by attribute mean or median or the most probable value. It is nothing but data normalization in data processing used to standardize the range of independent variables. This feature is very useful in data pre-processing step.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005139|

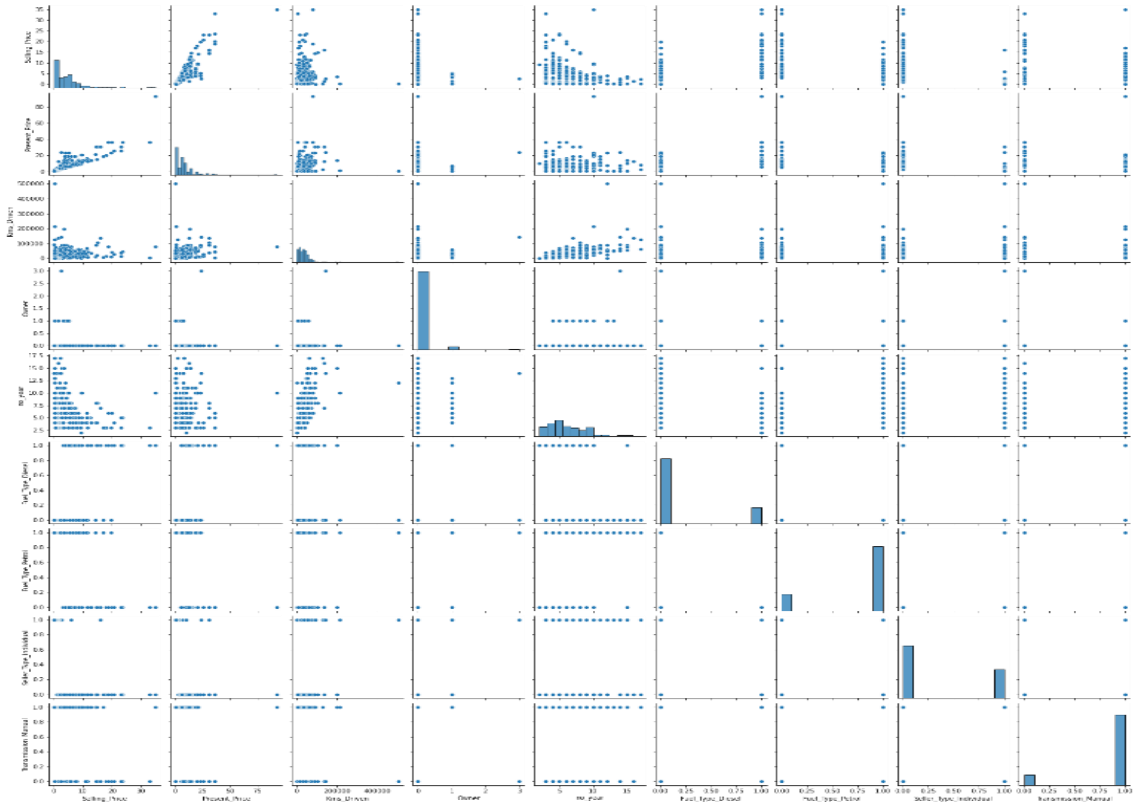

Fig3:Pair plot of all attributes of Dataset

We check the relation between the different different features and their relationship

Here the train data and the test data are splitted. Then the required algorithms are performed to obtain the better accuracy result of the car price All these data preprocessing steps have been carried out in Jupyter notebook python with necessary imports like pandas,numpy,sklearn and matplot . Before feeding data to an algorithm we have to apply transformations to our data which is referred as pre-processing. By performing pre-processing the raw data which is not feasible for analysis is converted into clean data. In-order to achieve better results using a model in Machine Learning, data format has to be in a proper manner. The data should be in a particular format for different algorithms. For example, if we consider Random Forest algorithm it does not support null values. So that those null values have to be managed using raw data.

Correlation: Correlation describes the linear relationship between the two continuous variables . Correlation is used when there is no identified response variable. It is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates extent to which those variables increase or decrease in parallel.A negative correlation indicates extent to which one variable increase as the other decreases.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005139|

Fig-4:Correlation

The following graph represents the variables and their precision towards the accuracy. here most part of the graph shows the values that are linear hence the alogithims works almost perfectly with utmost precision

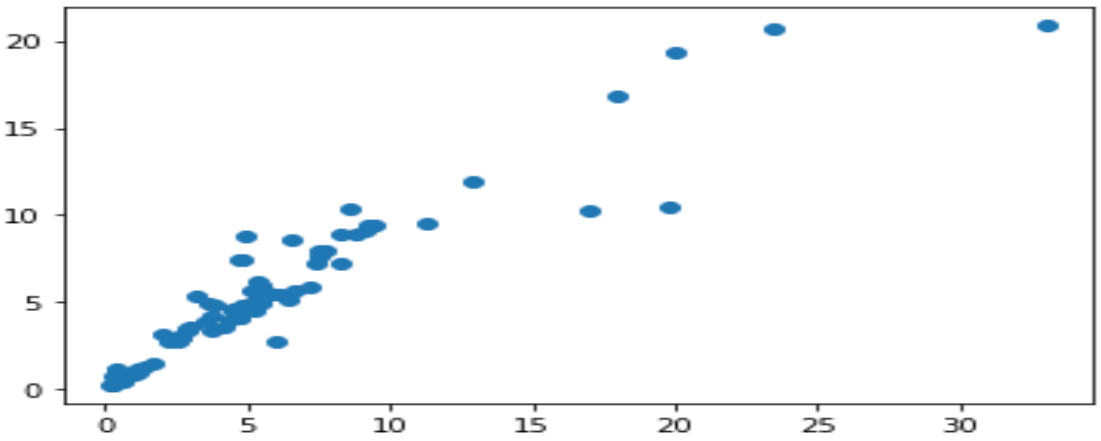

Fig-5:scatterplot

II. REGRESSION MODELS ANDESTIMTIONS

Linear regression is one of the important and well known algorithm in machine learning.It draws a relation between two or multiple features.It is a statistical linear machine learning algorithm that is used for predictive analysis. Here, the predicted analysis is continuous and has a constant slope which is used to predict values within a continuous or real range such as salary, age, product, sales, price . If no.of features are 2 it is known as Linear Regression and no.of features are more than 2 it is known as Multiple Regression.

The following algorithms are used formodel building:

- Linear Regression
- Lasso
- Ridge
- DecisionTreeRegressor

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005139|

- RandomForestRegressor
- AdaBoostRegressor
- GradientBoostingRegressor

1. Linear Regression:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Hypothesis function for Linear Regression :y= θ1+ θ2*x

x:input training data (univariate – one inputvariable(parameter))

y:labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and

θ2 values. **θ1:**intercept **θ2:**coefficient of x

Hypothesis function for Multiple Regression :

y = b1x1 + b2x2 + … + bnxn + c.

Here, bi's (i=1,2…n) are the regression coefficients, which represent the value at which the criterion variable changes when the predictor variable changes

RMSE:

Root Mean Squared Error (RMSD) of an estimator measures the average of error squares i.e. the average squared difference between the estimated values and true value. It is a risk function, corresponding to the expected value of the squared error loss. It is always non – negativeand values close to zero are better. The MSE is the secondmoment of the error (about the origin) and thus incorporates both the variance of the estimator and its bias.

Cross Validation Score with splits 5 and size0.2

array([0.81060255, 0.8060041, 0.81666227,
0.78280382, 0.78306781])

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

2.Lasso:

Linear Regression [2] model considers all the features

3.Ridge:

Ridge [2] is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased,

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005139|

and variances are large, this results in predicted values to be far away from the actual values.

For any type of regression machine learning models, the usual regression equation forms the base which is written as:

Y = XB + e

Where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated, and e represents the errors are residuals.

Once we add the lambda function to this equation, the variance that is not evaluated by the general model is considered. After the data is ready and identified to be part of L2 regularization, there are steps that one can undertake.

equally relevant for prediction. When there are manyfeatures in the dataset and even some of them are not relevant for the predictive model. This makes the model more complex with a too inaccurate prediction on the test set (or overfitting). Such a model with high variance does not generalize on the new data. So, Lasso Regression comes for the rescue. It introduced an L1 penalty (or equal to the absolute value of the magnitude of weights) in the cost function of Linear Regression. t. In mathematical form, the model can be expressed as y=xb+e Here, y is the dependent variable x refers tofeatures in matrix form and b refers to regression coefficients and e represents residuals.

4.Huber:

An alternative apprioach is based on slightly modified model loss function called huber loss(for a single sample).the parameter th defines the threshold that makes a function switch from a squared error to an absolute one.in this way the magnitude of the loss changes accordingly passing from a quadratic behaviour to a linear one when the points are supposed to be outliners.in this way their contribution to the global cost function is reduced and the hyperplane will remain closer to the majority of points even in presence of outliners .
Another common situation in which robust estimation is used occurs when the data contain outliers. In the presence of outliers that do not come from the same data-generating process as the rest of the data, least squares estimation is inefficient and can be biased. Because the least squares predictions are dragged towards the outliers, and because the variance of the estimates is artificially inflated, the result is that outliers can be masked. (In many situations, including some areas of geostatistics and medical statistics, it is precisely the outliers that are of interest.)
Although it is sometimes claimed that least squares (or classical statistical methods in general) are robust, they are only robust in the sense that the type I error rate does not increase under violations of the model. In fact, the type I error rate tends to be lower than the nominal level when outliers are present, and there is often a dramatic increase in the type II error rate. The reduction of the type I error rate has been labelled as the *conservatism* of classical methods.

4.Random Forest Regressor:

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging.The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

Decision trees are a popular method for various machine learning tasks. Tree learning "come[s] closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", say Hastie *et al.*, "because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and produces

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005139|

inspectable models. However, they are seldom accurate".[3]:352

5. Elastic net CV

The elastic net method overcomes the limitations of the LASSO (least absolute shrinkage and selection operator) method which uses a penalty function based on

Use of this penalty function has several limitations.[1] For example, in the "large *p*, small *n*" case (high-dimensional data with few examples), the LASSO selects at most n variables before it saturates. Also if there is a group of highly correlated variables, then the LASSO tends to select one variable from a group and ignore the others. To overcome these limitations, the elastic net adds a quadratic part to the penalty (which when used alone is ridge regression (known also as Tikhonov regularization). The estimates from the elastic net method are defined by

The quadratic penalty term makes the loss function strongly convex, and it therefore has a unique minimum. The elastic net method includes the LASSO and ridge regression: in otherwords, each of them is a special casewhere or . Meanwhile, the naive version of elastic net method finds an estimator in a two-stage procedure : first for each fixed it finds the ridge regression coefficients, and then does a LASSO type shrinkage. This kind of estimation incurs a double

amount of shrinkage, which leads to increased bias and poor predictions. .

III. PERFORMANCE EVALUATION

The model is build using several regression algorithms.The below accuracy plot represents the various scores of algorithms when a model is trained with that particular algorithm.The Random forest Regression given good scores.

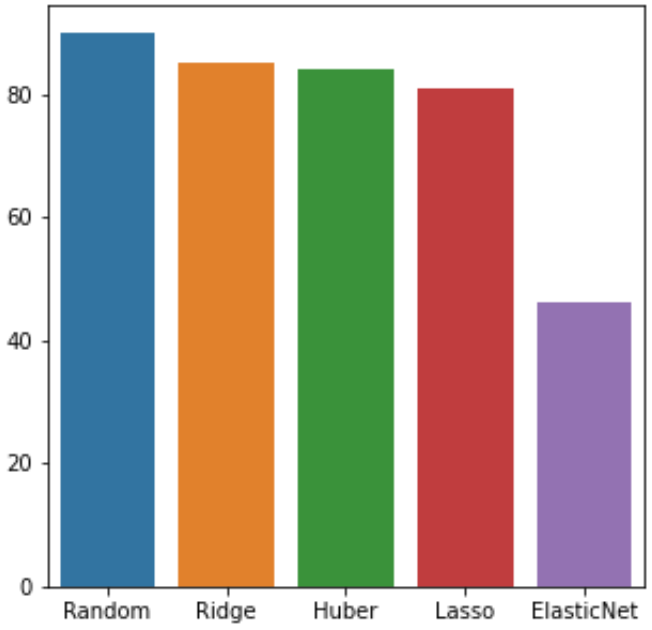

Fig-6:Accuracy Plot

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005139|

IV. CONCLUSION AND FUTURE SCOPE

We have used 5 different algorithms like random forest regressor,Ridge regressor, huber regressor ,Lasso regressor,Elastic net cv regressor fort he prediction of car prices. out of these algorithms the accuracy value of the random forest regressor is 91% the accuracy value of ridge regressor is 88% the accuracy value of huber regressor is 80% the accuracy value of lasso regressor is 80% the accuracy value of elastic net cv is 46% hence random forest regressor exhibits the higher accuracy value out of any other algorithms .hence random forest regression algorithm is used for the prediction of the car prices .

This project further can be developed as android application to overcome the limitation of accessing the system by only desktop and also tell the clients about the prices of the similar cars.we also try to send notification to the client about the prices of the cars for the clients who already have account in our application ans send the notification to them depending upon the recent searches of that client .`

REFERENCES

1. Theobald, O. (2017). Machine learning for absolutebeginners
2. S. Abhishek.:Ridge regression vs Lasso,How these twopopular ML Regression techniques work. Analytics India magazine,2018
3. Minaie, N. (2019). A Beginner's Guide to Selecting Machine Learning Predictive Models in Python, https://towardsdatascience.com/the-beginners-guide-to- selecting-machine-learning-predictive-models-in-python-
4. Little, R. J., & Rubin, D. B. (2014). Statisticalanalysis with missing data (Vol. 333). John Wiley & Sons.
5. Barnett, V., & Lewis, T. (1974). Outliers in statistical data. Wiley.
6. Friedman, J. 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29(5):1189–1232.
7. R. T. Azuma et al., "A survey of augmented reality," Presence, vol. 6, no. 4, pp. 355–385, 1997.
8. https://scikit-learn.org/stable/install.html
9. S.Raheel.Choosing the right encoding method-Label vsOne hot encoder. Towards datascience,2018
10. Raj, J. S., & Ananthi, J. V. (2019). Recurrent neural networks and nonlinear prediction in support vector machines.Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details