IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 6, June 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106297|

Black Friday Sales Prediction Analysis using Machine Learning Techniques

P Poornima [1], Jennifer Joyce B [2]

Assistant Professor, Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology Hyderabad, India

Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology Hyderabad, India

ABSTRACT: The Intelligent Decision Analytical System requires integration of decision analysis and predictions. Business organizations often tend to depend on a heavy knowledge base and demand prediction of sales trends. The accuracy of sales forecasts provides a big impact on business. This is the age of the internet where the amount of data being generated is so huge that man alone is not able to process the data. Several effective tools are used in extracting hidden knowledge from an enormous dataset to enhance accuracy and efficiency of forecasting. Several techniques and measures are also used for sales prediction. The aim of this paper is to analyze the sales transactions captured at a retail store during Black Friday, which highlights multiple shopping experiences, and to analyze purchase amounts from a variety of customers.
The customer's purchasing decisions are analyzed in a way that allows us to see the various factors that allow them to make a purchase. We use machine learning techniques to make our analysis to find the best model for prediction.

This paper aims to analyze how demographic factors influence Black Friday Sales at a retail store through data visualization and predict future Black Friday sales using rigorously chosen machine learning models.

KEYWORDS: Machine Learning, Random Forest, Linear Regression, Decision Tree.

I. INTRODUCTION

Sales forecasting can be defined as the prediction of upcoming sales based on the past sales occurred. Sales forecasting is of paramount importance for companies which are entering new markets or are adding new services, products or which are experiencing high growth. The main reason a company does a forecast is to balance marketing resources and sales against supply capacity planning.

Forecasting can help in answering some expository queries like "Do we have the right mix of price, promotion, and marketing in place to drive demand?", "Do we have enough salespeople to get the volume of orders we have budgeted?", "Do we have the essential demand-side resources in place?" and for these reasons, many of the companies allocate significant financial and human resources to perform this task genuinely, which requires large investment. Manufactured organizations and business houses require an accurate and reliable forecast of sales data so that they don't suffer from losses due to wrong or inaccurate predictions by the model. Companies mainly use sales forecasting to determine two things. First, to determine the current demand level of the service or product in the market. Second, to determine the future demand for a company's services or products.

1.1 Problem Statement

This paper "BLACK FRIDAY SALES PREDICTION ANALYSIS "is mainly based on machine learning algorithms by using JUPYTER Notebook. Black Friday Sales prediction analysis deals with analyzing a retail store's sales data during the period of Black Friday, to understand the purchase decisions being made by customers by surveying demographic data.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106297|

1.2 Existing System

The retail industry needs a huge upgrade in order to survive the changing conditions of the economy. Along with the advances in technologies used in improving the retail sector, useful and accurate information about customer purchases also plays a significant role in it. This information is being gathered and collected so as to make an accurate sales forecast. This information along with the knowledge of subject experts and researchers should be known to retail owners and businesses so as to explore its potential.

1.3 Proposed System

The proposed system suggests sales forecasting to predict upcoming sales based on past sales. Here we focus on product-level forecasts. Sales forecasts affect a company's marketing plan directly. . The marketing department is responsible for how clients and their customers interpret its services and products and compare it against its competitors and use the sales forecast to assess how marketing spending can increase sales and channel demand. It is important to develop effective sales forecasting models in order to generate accurate and robust forecasting results. In the business and economic environment, it is very important to accurately predict various kinds of economic variables such as Past Economic Performance, Current Global Conditions, Current Industry Conditions, Rate of Inflation, Internal Organizational Changes, Marketing Efforts, Seasonal Demands, etc. to develop proper strategies. On the contrary, inaccurate forecasts may lead to inventory shortage, unsatisfied customer demands, and product backlogs. Due to these reasons, outmost importance is given to develop productive models in order to generate robust and accurate results. Additionally, suitable classification methods like Linear Regression (LR), Decision Tree Regression (DTR), Random Forest Regressor (RFR) and Extra Tree Regressor (ETR) are employed for better classification outcomes.

II. LITERATURE SURVEY

1. Sales forecasting of retail stores using ML techniques has aimed to predict the sales of a retail store using different machine learning techniques and determined the best algorithm suited to the particular problem statement. Normal regression techniques have been implemented as well as boosting techniques and it has been found that the boosting algorithms have better results than the regular regression algorithms.
2. Sales forecasting using Deep Neural Network and SHAP techniques : The Neural Network sale prediction model has been used for predicting Walmart's sales. Moreover, we evaluate our NN model on the datasets provided by the Kaggle platform. Experiments have shown that compared with other machine learning models, the NN model achieves superior performance. The RMSE metric is 2.92 and 2.58 lower than the Linear Regression algorithm and the SVM algorithm. Furthermore, to mine attributes of different dimensions to make predictions well, SHAP is utilized to interpret the NN model.
3. Machine Learning Model for sales forecasting using XGBoost :An efficient and accurate sales forecasting model was proposed using machine learning. Initially, feature engineering is conducted for extracting features from historical sales data. Furthermore, eXtreme Gradient Boosting (XGBoost) is used to utilize these features for forecasting the future sales amount. The experiment results on a publicly available Walmart retail goods dataset provided by Kaggle competition to demonstrate the proposed model performs extremely well for sales prediction with less computing time and memory resources.
4. Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms : A predictive model was developed using Xgboost, Linear regression, Polynomial regression, and Ridge regression techniques for forecasting the sales of a business such as Big -Mart, and it was discovered that the model outperforms existing models.

III. METHODOLOGY

Data is an especially important part of any Machine Learning System. Python code has been used for implementation of this paper. Python Code is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106297|

Jupyter Notebooks are a powerful way to write and iterate on your Python code for data analysis. Rather than writing and re-writing an entire program, you can write lines of code and run them one at a time. Then, if you need to make a change, you can go back and make your edit and rerun the program again, all in the same window.

Jupyter Notebook is built off of IPython, an interactive way of running Python code in the terminal using the REPL model (Read-Eval-Print-Loop). The IPython Kernel runs the computations and communicates with the Jupyter Notebook front-end interface. It also allows Jupyter Notebook to support multiple languages. Jupyter Notebooks extend IPython through additional features, like storing your code and output and allowing you to keep markdown notes.

IV. DESIGN AND IMPLEMENTATION

Here, the design and implementation of this particular is shown, wherein we visualize the sales data prediction process and realize the various steps that must be undertaken. We begin by generating a hypothesis such that they favor the needed outcome. We then achieve more accuracy through data exploration. Feature engineering is used to understand the nuances in the data For building our prediction model we segregate training and testing data and finally do an analysis of the performance of the model.

4.1 Sales Data Prediction Process

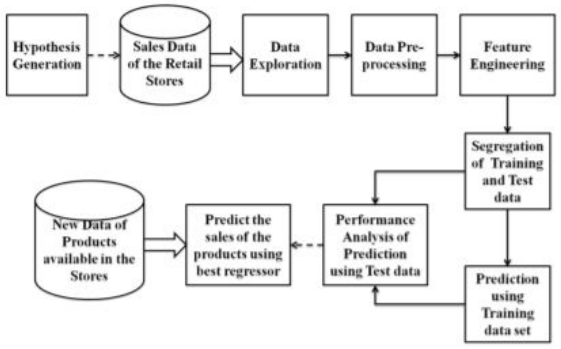


Fig.4.1. Sales Data Prediction

System Modules:

1. **Hypothesis Generation:**

In this step various hypotheses are generated by analyzing the problem statement. The hypotheses are generated in such a way that they favor the needed outcome. So the major idea in this step is to identify the properties of a product and the stores, which can have an impact on sales. Since the forecasting is done on the basis of the products and stores, we can categorize the hypotheses as “Product Level Hypotheses” and “Store Level Hypotheses”.

2. **Data Exploration:**

The first step in Data exploration is to look into the dataset and to discover the information regarding the available and the hypothesized data. The dataset under consideration has the features as the variables.

3. **Data Preprocessing:**

This step typically imputes the missing values and handles the outliers present in the dataset.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106297|

Three common data preprocessing steps are:

- **Formatting**: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.

- **Cleaning**: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely.
- **Sampling**: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

4. Feature Engineering:

During the data exploration step, we identified a few of the nuances in the data. This step deals with those nuances and also is used for creating new variables out of existing variables so that our data will be ready to perform the analysis. As we know that machine learning algorithms take numerical values as input, we encode the categorical variables into numerical variables by using LabelEncoder and OneHotEncoder of sklearn's preprocessing module.

5. Segregation of Training and Testing Data:

A major step in machine learning is to feed some amount of data into the algorithm and to train it to understand the pattern of data. Once the algorithm learns the pattern, one has to feed another dataset to check the level of understanding done by the algorithm. It is a practice to divide the available data into two subsets in the ratio of 4:1 for the purpose of training and testing.

6. Model Building for Prediction:

After the dataset is split into training and testing sets, the training set is fed into the algorithm so that it can learn how to predict the values.

7. Performance analysis and Prediction:

The performance of any regression algorithm is computed by feeding in the test data in the training model. This procedure gives us the idea how well a model has learned the pattern present in the dataset and is able to predict the values of the new data. The performance of any given regression model can be computed using the Root Mean Square Error (RMSE) or the R2 error.

4.2 Implementation and Performance Analysis:
The proposed algorithm is implemented using Python 3.6 and sklearn 0.19.1. The sales forecast dataset from Kaggle has been pre-processed with the steps mentioned above to predict the sales. The dataset consists of 550029 entries and has 12 features. Different models are created and simulated based on this dataset containing 550029 entries.

Sales Prediction Data Model Training. Data Partitioning: The Entire dataset is partitioned into 2 parts: for example, say, 75% of the dataset is used for training the model and 25% of the data is set aside to test the model. To predict future events Machine Learning Algorithms:

Gradient Boost Algorithm : Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. Gradient boosting is one of the boosting algorithms that is used to minimize bias error of the model. The base estimator for the Gradient Boost algorithm is fixed and i.e. *Decision Stump*. Gradient boosting algorithm can be used for predicting not only continuous target variables (as a Regressor) but also categorical target variables (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106297|

Machine Learning Algorithms Used:

1. **Linear Regression :**
   Linear regression is useful for finding relationships between two continuous variables. One is a predictor or independent variable and other is response or dependent variable. It looks for statistical relationships but not deterministic relationships. Relationship between two.

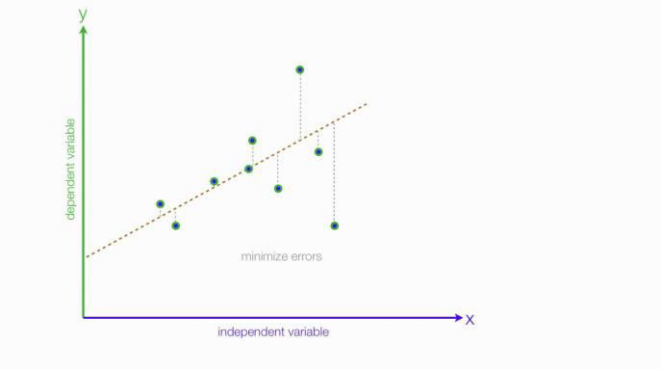


Fig 4.2.1 Linear Regression

variables are said to be deterministic if one variable can be accurately expressed by the other. The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction errors (all data points) are as small as possible. Error is the distance between the point to the regression line.

2. **Decision Tree Regressor:**

   Decision trees build regression or classification models in the form of a tree structure. It breaks down a 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.

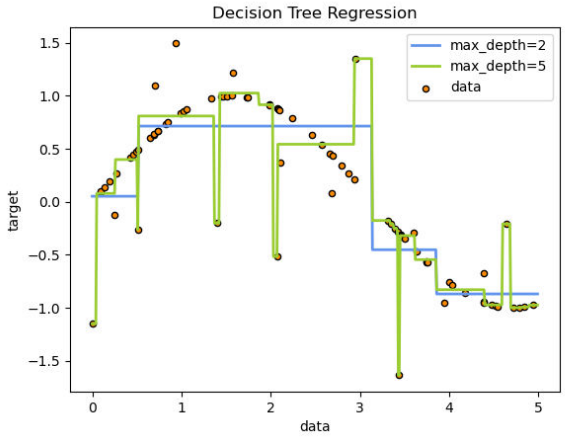


Fig 4.2.2 Decision Tree Regressor

3. **Random Forest Regressor:**

   A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106297|

size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

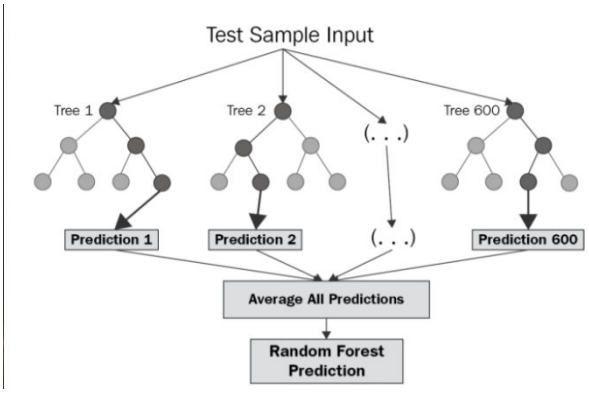


Fig 4.2.3 Random Forest Regressor

4. Extra Tree Regressor:

This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

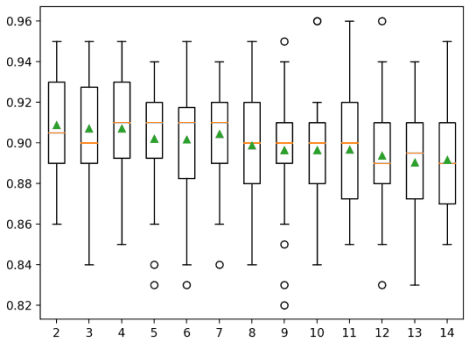


Fig 4.2.4 Extra Tree Regressor

V. EXPERIMENTAL RESULTS

Histogram:

In Matplotlib we can create a Histogram using the hist method. If we pass categorical data like the points column from the coral dataset it will automatically calculate how often each class occurs.

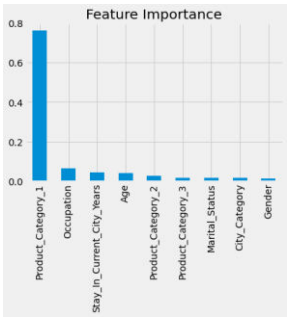
Bar Chart:

A bar-chart can be created using the *bar* method. The bar-chart isn't automatically calculating the frequency of a category, so we are going to use pandas *value_counts* function to do this. The bar-chart is useful for categorical data that doesn't have a lot of different categories.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

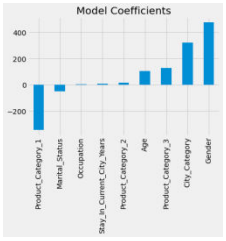
| DOI:10.15680/IJIRSET.2022.1106297|



Graph 5.1 Random Forest Regressor

1. Random Forest Regressor:

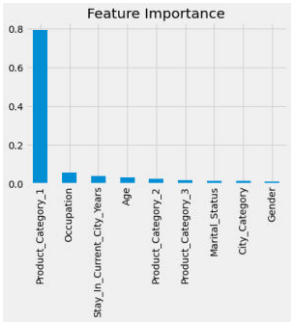
A random forest is **a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset** and uses averaging to improve the predictive accuracy and control over-fitting. The algorithm operates by constructing a multitude of decision trees at training time and outputting the mean/mode of prediction of the individual trees.



Graph 5.2 Linear Regression

2. Linear Regression:

Linear regression is a linear model, e.g. a **model that assumes a linear relationship between the input variables (x) and the single output variable (y).** linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.



Graph 5.3 Decision Tree Regression

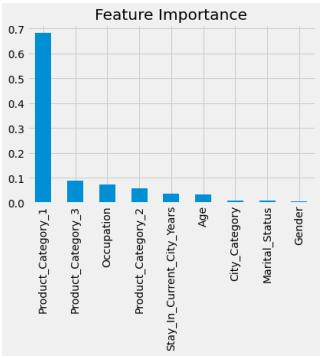
| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106297|

3. Decision Tree Regressor:

Decision trees build regression or classification models in the form of a tree structure. **It breaks down a 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.



Graph 5.4 Extra Tree Regressor

5. Extra Tree Regressor:

An extra-trees regressor. This class **implements a meta estimator that fits a number of randomized decision trees** (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

VI. CONCLUSION AND FUTURE SCOPE

6.1 Conclusion:

In this paper, a model using various algorithms such as Linear regression, Decision tree regression, Random forest and XGB regressor to get the best possible prediction is built.The hyperparameter tuned XGB regressor gives us the best rmse value and r2 score for this problem. Hence, by the obtained results the GradientBoost algorithm is observed as the best predictor for the dataset that is taken from Kaggle having the least RMSE value of 1088.64 and the highest R2 value of 0.59.

6.2 Future Scope:

The sales prediction data of the given dataset is observed in this paper. In the future, for a large dataset, neural networks such as an artificial neural network can be used to build a model which can result in better performance.

REFERENCES

(1) A. Krishna, A. V, A. Aich and C. Hegde, "Sales-forecasting of Retail Stores using Machine Learning Techniques," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), 2018, pp. 160-166, doi: 10.1109/CSITSS.2018.8768765.

(2) J. Chen, W. Koju, S. Xu and Z. Liu, "Sales Forecasting Using Deep Neural Network And SHAP techniques," 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 2021, pp. 135-138, doi: 10.1109/ICBAIE52039.2021.9389930.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106297|

(3) X. dairu and Z. Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2021, pp. 480-483, doi: 10.1109/ICCECE51280.2021.9342304.

(4) R. P and S. M, "Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1416-1421, doi: 10.1109/ICICCS51141.2021.9432109.

(5) S. K. Punjabi, V. Shetty, S. Pranav and A. Yadav, "Sales Prediction using Online Sentiment with Regression Model," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 209-212, doi: 10.1109/ICICCS48265.2020.9120936.

(6) Q. Sun and X. Luan, "The Study of Cluster Predication Method on Sales Forecast Based on Residual Error Modified GM (1, 1)," 2008 International Seminar on Business and Information Management, 2008, pp. 46-49, doi: 10.1109/ISBIM.2008.64.

(7) H. V. Ramachandra, G. Balaraju, A. Rajashekar and H. Patil, "Machine Learning Application for Black Friday Sales Prediction Framework," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 57-61, doi: 10.1109/ESCI50559.2021.9396994.

(8) T. Tandel, S. Wagal, N. Singh, R. Chaudhari and V. Badgujar, "Case Study on an Android App for Inventory Management System with Sales Prediction for Local Shopkeepers in India," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 931-934, doi: 10.1109/ICACCS48705.2020.9074234.

(9) W. Shan, "Research on Refined Sales Management, Data Analysis and Forecasting under Big Data," 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2020, pp. 305-308, doi: 10.1109/MLBDBI51377.2020.00065.

(10) H. Koptagel, D. C. Cıvelek and B. Dal, "Sales Prediction using Matrix and Tensor Factorization Models," 2019 27th Signal Processing and Communications Applications Conference (SIU), 2019, pp. 1-4, doi: 10.1109/SIU.2019.8806437.

(11) Linear Regression - Detailed View : Saishruthi Swaminathan

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details