Mini Project Report

Entitled

Prediction of Sales

Submitted to the Department of Electronics Engineering in Partial Fulfilment for the Requirements for the Degree of

Bachelor of Technology
(Electronics and Communication)

: Presented & Submitted By :

Karni Tarun, Kadimi Sri Surya Sashank

Roll No. (U20EC123, U20EC145)

B. TECH. VI (EC), 6^{th} Semester

: Guided By :

Dr. Kishor Upla

Assistant Professor, SVNIT

(Year: 2022-23)

DEPARTMENT OF ELECTRONICS ENGINEERING

SARDAR VALLABHBHAI NATIONAL INSTITUTE OF TECHNOLOGY

Surat-395007, Gujarat, INDIA.

Sardar Vallabhbhai National Institute Of Technology

Surat - 395 007, Gujarat, India

DEPARTMENT OF ELECTRONICS ENGINEERING

CERTIFICATE

This is to certify that the Mini-Project Report entitled "**Prediction of Sales**" is presented & submitted by Karni Tarun, Kadimi Sri Surya Sashank, bearing Roll No. U20EC123, U20EC145, of B.Tech. VI, 6^{th} Semester in the partial fulfillment of the requirement for the award of B.Tech. Degree in Electronics & Communication Engineering for academic year 2022-23.

They have successfully and satisfactorily completed their **Mini-Project** in all respects. We, certify that the work is comprehensive, complete and fit for evaluation.

Dr. Kishor Upla
Assistant Professor & Project Guide

Abstract

During the last few decades, sales forecasting has become an important factor in marketing industry that has gained a lot of attention for its ability to improve market operations and product sales. Historically, businesses have concentrated majorly on a standard statistical model. But in recent years, machine learning approaches and the usage of its algorithms have drawn greater attention.

This work will help to identify the important features which will influence the sales of a product. In this project demonstration we have predicted the sales of iphone by using machine learning algorithms such as Support Vector Regression, K- Nearest Neighbour algorithm and Logistic Regression which are expected to give the best results for the available data. The data is preprocessed and cleaned to ensure that it is ready for analysis. The performance of the models is evaluated using metrics such as accuracy, precision, recall, and F1-score.

Table of Contents

Page

Abstract . . . iii

Table of Contents . . . iv

List of Figures . . . v

List of Abbreviations . . . vi

Chapters

1 Introduction . . . 1

1.0.1 Objectives . . . 1

1.1 Preprocessing . . . 1

1.2 Machine Learning . . . 3

2 Methodology . . . 5

2.1 Proposed Model . . . 5

3 Observations and Results . . . 7

3.1 Feature Extraction . . . 7

3.2 Performance metrics . . . 7

List of Figures

1.1 Types of Machine Learning 3

3.1 Correlation of features 7
3.2 Performance of various ML models 8

List of Abbreviations

KNN	K-Nearest Neighbour
SVM	Support Vector Machine
SVC	Support Vector Classifier
CSV	Comma Seperated Values
PSNR	Peak Signal to Noise Ratio
PCA	Principal Component Analysis

Chapter 1
Introduction

In past times, businesses produced things without taking demand or sales volume into account. Data about the demand for items on the market is needed for any manufacturer to decide whether to expand or decrease the production of multiple units. Companies that compete in the market without taking these values into account are at risk of falling out. Many companies adopt specific requirements for assessing their sales and demand.

Accurate and on time revenue forecasting, also known as sales forecasting or revenue forecasting, can give businesses involved in the manufacturing, shipping, or retail of goods an important insight in today's extremely competitive environment and rapidly evolving consumer market.

Companies used to make sales predictions randomly in the past. But the availability of sales data nowadays and other related information can now be used through Machine Learning techniques to predict the sale of products. The data sets are separated into 2 parts i.e, training data and test data. Training data is utilized to train the model and test data is used to assess the trained model. We split the data set in 80-20 ratio where 80 percent data is used to train the model and the remaining 20 percent data is used for testing.

1.0.1 Objectives

1. Converting data into an appropriate form using various pre-processing techniques for the implementation of Machine Learning algorithms.
2. Finding critical features that will most influence sales of the product.
3. To determine the appropriate Machine Learning algorithm for sales forecasting.
4. Selecting various metrics to compare the performance of the applied Machine Learning algorithms.

1.1 Preprocessing

Preprocessing is a critical step in Machine Learning (ML) that involves preparing data for use in model training. This step is crucial because the quality of the data can significantly impact the performance of the model. Preprocessing techniques help to clean

and transform the raw data into a format that is suitable for ML. Common preprocessing techniques include handling missing values, duplicates, and errors in the data, transforming categorical data into numerical data, and scaling features to the same level.

Categorical data refers to data that is non-numerical and can take on a limited number of values, such as colors, countries, or job titles. Machine learning algorithms generally require numerical data as input, so categorical feature encoding is necessary to convert categorical data into numerical data. There are several methods of categorical feature encoding, some of which include:

1. One-Hot Encoding: This method creates a new binary feature for each unique category in a categorical feature. The value of the binary feature is 1 if the original feature had that category and 0 otherwise. One-hot encoding is commonly used when the number of categories is small.

2. Label Encoding: This method assigns a unique numerical value to each category in a categorical feature. Label encoding is useful when the number of categories is large, and one-hot encoding would result in a high-dimensional feature space.

3. Ordinal Encoding: This method assigns numerical values to categories based on their order or rank. For example, if the categories are ”low,” ”medium,” and ”high,” ordinal encoding would assign the values 1, 2, and 3, respectively.

4. Binary Encoding: This method encodes categories using binary digits. Each category is assigned a unique binary code, and the resulting binary digits are used as the feature values.

The choice of categorical feature encoding method depends on several factors, such as the number of categories, the nature of the categories, and the machine learning algorithm used. It is essential to choose an appropriate encoding method to ensure that the encoded features accurately represent the original data and improve the performance of the machine learning model.

Data splitting is also an essential technique that involves dividing the data into separate sets for training, validation, and testing.

Dimensionality reduction is another preprocessing technique that involves reducing the number of features in the data, which can improve model performance and prevent overfitting. Additionally, handling imbalanced data is important to ensure that the model is not biased towards one class.

Overall, preprocessing is an essential step in ML that requires careful consideration and understanding of the data. By applying appropriate preprocessing techniques, we can help ensure that the model is accurate and performs well.

1.2 Machine Learning

Machine Learning is the area of study which enables machines to learn without being explicitly programmed. In general, Machine Learning is a program that can manage various tasks by analyzing and exploring data. Common Machine Learning applications such as email spam detection, credit card fraud, stock predictions, smart assistants, product recommendations, self-driving cars, sentiment analysis, etc.

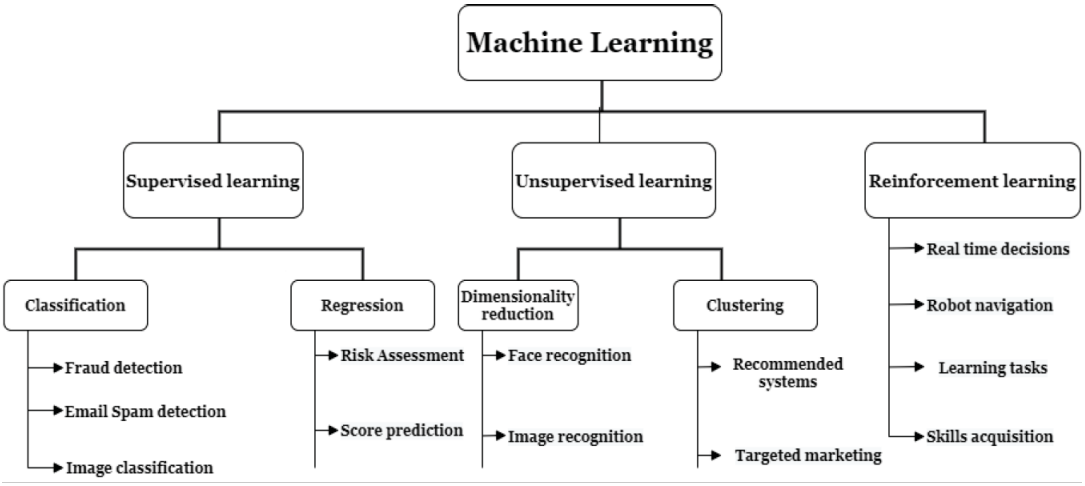

Figure 1.1: Types of Machine Learning

Support Vector Machine

Support Vector Machine or SVM is one of the most common Supervised Learning algorithms used for both Classification and Regression issues. The SVM algorithm aims to build the best line or decision boundary that can divide n-dimensional space into conveniently place the new data point in the right category in the future. The optimal choice boundary is called a hyper plane. SVM chooses extreme points vectors that help to create a hyper plane. Such extreme cases are called help vectors. The equation for Support Vector Regression is:

$$f(x) = x^{'}\beta + b \tag{1.1}$$

Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

K-Nearest Neighbor(KNN) Algorithm

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. This algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using KNN algorithm. It can be used for Regression as well as for Classification but mostly it is used for the Classification problems. It is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Chapter 2
Methodology

Many factors come into play in prediction of sales of a product. We have researched for this topic and chose iPhone purchase prediction and got a brief understanding from ref1.

2.1 Proposed Model

The basic attributes considered in the existing model are gender, age, salary used certain machine learning models to predict whether a person is able to buy the product or not. We felt this wasn't sufficient and worked upon building a dataset with other features such as employment status, locality the person stays and also if the person previously belongs to apple ecosphere. Many more attributes could be added but for understanding and developing a machine learning model we felt these would be useful.

In the existing model there was not much pre-processing required as the data didn't have much attributes. But as we have included more attributes and as they are categorical in nature, it is very evident the data had to be converted into numerical to be helpful in analysing it. Feature scaling is an important step to not miss out as the machine learning considers everything on the same scale while the features ideally wouldn't be so. The given data is split into training data to train the model and testing data to analyse the performance of the model. Algorithms such as SVM, Logistic regression, KNN are used to train this data. Inorder to evaluate the performance of the model confusion matrix method is used. As different ML models are used a comparision study between them is also done based on the metrics. Instead of having a seperate validation dataset we give inputs here itself to make new predictions. But as we did this we got very less accuracy, so inorder to improve the model performance we had to do certain feature engineering and hyperparameter tuning and thus results have drastically improved as well.

So our proposed model implements the following steps to predict the sales of iPhone:

1. **Data loading:** The project uses a CSV file as the data source. The data is loaded using the pandas library, and the independent variables (gender, age, and salary) are stored in the X variable, while the dependent variable (purchase) is stored in the y variable.

2. **Data preprocessing:** The project preprocesses the data to convert the categorical variable (gender) into a numerical variable using LabelEncoder. This step is necessary because most machine learning algorithms cannot handle categorical data.

3. **Split Data into training and testing:** After preprocessing generally the data is split into training and testing purpose. If additional unseen data is given also known as validation data it is used to make predictions on new data. Unfortunately in our case we don't have validation dataset. So we perform all the evaluation on test data itself.

4. **Feature scaling:** Feature scaling is an essential step in machine learning projects because it helps to normalize the data and make it easier for the algorithm to converge. In this project, StandardScaler is used to scale the features (age and salary) to a common scale.

5. **Model selection:** Logistic regression is used as the machine learning algorithm in this project. The model is built using the LogisticRegression function from the sklearn.linear_model library.

6. **Model evaluation:** The project evaluates the performance of the logistic regression classifier using various evaluation metrics such as accuracy score, precision score, recall score, and confusion matrix.

7. **Prediction:** After building the model, the project uses it to predict whether a customer will buy an iPhone based on their gender, age, and salary. The project provides eight sample predictions, with four samples for male and four for female customers, each with different ages and salaries.

In summary, the methodology used in this project includes data loading, data preprocessing, feature scaling, model selection, model evaluation, and prediction. These steps are essential in any machine learning project, and they are critical to the success of the project.

Chapter 3
Observations and Results

The data has several categorical data where we used label encoding and they are converted to numeric. For example

3.1 Feature Extraction

In order to extract useful features we can use different techniques like correlation matrix, PCA i.e principal component analysis. Here, we can see the correlation of each feature w.r.t the iPhone purchase of a customer. So by observing the above values it is quite clear

```
Will Purchase iPhone    1.000000
Salary                  0.285522
Employment Status       0.269726
Has Apple Products      0.198477
Location                0.157183
Gender                  0.022590
Age                    -0.226008
```

Figure 3.1: Correlation of features

that the purchase of iPhone is highly correlated with Salary and employment Status of a person. Also it has a high negative correlation with Age thus these 3 features can be considered very crucial for prediction of sales. But we can also see the gender plays a negligible role in this and thus such features can be neglected. So we take a correlation threshold of 0.15 and neglect all the features having correlation below it. Note that the absolute value has to be taken as negative correlation is also to be considered for better prediction.

3.2 Performance metrics

In machine learning, accuracy, precision, and recall are three commonly used metrics to evaluate the performance of a classification model.

- Accuracy: Accuracy is a measure of how often the model makes correct predictions. It is calculated as the number of correct predictions divided by the total number of predictions.

- Precision: Precision is a measure of the proportion of true positives among the instances predicted as positive. It is calculated as the number of true positives divided by the total number of predicted positives.
- Recall: Recall is a measure of the proportion of true positives among the actual positive instances. It is calculated as the number of true positives divided by the total number of actual positives.

On implementing the algorithms on the data set, we have obtained the following results:

```
KNN
[[89 31]
 [ 9 71]]
Accuracy score: 0.8
Precision score: 0.696078431372549
Recall score: 0.8875
Logistic Regression
[[86 34]
 [24 56]]
Accuracy score: 0.71
Precision score: 0.6222222222222222
Recall score: 0.7
SVM
[[85 35]
 [21 59]]
Accuracy score: 0.72
Precision score: 0.6276595744680851
Recall score: 0.7375
```

Figure 3.2: Performance of various ML models

On obtaining the results we can observe that KNN algorithm gives the best accuracy score among other algorithms used. The accuracy of the predictions can be improved by using advanced machine learning algorithms such as neural networks, decision trees, and regression models. It is important to have access to quality data that is clean, accurate, and relevant to the task at hand also quantity plays a role i.e small dataset can produce accurate results.

Conclusion

Sales forecasting plays a vital role in the business sector in every field. With the help of the sales forecasts, sales revenue analysis will help to get the details needed to estimate both the revenue and the income. Different types of Machine Learning techniques such as Support Vector Regression, KNN algorithm and Logistic Regression have been evaluated on iphone sales data to find the critical factors that influence sales to provide a solution for forecasting sales. After performing metrics such as accuracy score, precision score and recall score, the KNN algorithm is found to be the appropriate algorithm according to the collected data and thus fulfilling the aim of this project.