1. Project Summary

Customer behaviour prediction and identifying their buying patterns is one of the challenging problems in the sales industry and Supermarkets. With the advancement in the field of machine learning and data science, the possibilities to classify whether the customer is prone to buy a few expensive products and predicting the amount of money they usually spend in a month has increased significantly. We have conducted our research to tackle these two problems separately in which we will create a classification model for the first task and a regression model for the second task. Our proposed methodology consists of six phases. In the first two phases, data pre-processing and feature analysis is performed. In the third phase, feature selection is taken into consideration using the lasso regression. Next, the data has been split into two parts: train and test set in the ratio of 80% and 20% respectively. In the prediction process, most popular predictive models have been applied, namely, logistic regression, support vector machine, random forest and decision trees for classifying whether the customer is prone to buy a few expensive products and linear regression, random forest regressor and decision tree regressor to predict the amount of money they usually spend in a month. Different learning algorithms and techniques are applied to see the effect on accuracy of models. In addition, K-fold cross validation has been used over train sets for hyperparameter tuning and to prevent overfitting of models. Finally, the obtained results on the test set have been evaluated using confusion matrix ,AUC score, Recall score and Precision score for classification task and root mean square and mean square for regression task. It was found that Random Forest Classifiers give the highest accuracy of 87.71% on classification tasks and Linear Regression outperforms other regression methods to give the lowest root mean square of around 530 respectively.

2. Problem Definition

The famous supermarket chain "Tosco & Spency" would like to differentiate its offer and marketing strategy on the base of two main classes of customers. Indeed there are two kinds of customers visiting the supermarket: the first ones are prone to buy a few expensive products and the latter usually buy many cheap products. They provide you with a training set of historical data containing features of each customer and a label representing whether the customers belong to the first class or not.

In the same way, the famous competitor "Sunsbory's" would like to design a similar system for them but, unlike the first system, they would like you to predict not only if the customer is prone to buy few expensive products but also the amount of money that a customer usually spend in a month. They provide you with a training set of historical data containing features of each customer and a numerical value representing the value of the amount of money they spent in May 2021.

3. Aims & Objectives

The main aims of this project are as follows:

i) to identify the appropriate data requirements needed for modelling the algorithm..

ii) to conduct an analysis on the provided datasets to select the appropriate features that have a relation with categorizing a potential buyer..

iii) to experiment with different classification models and compare their performance of the solution before deploying it.

5. Methodology

The proposed solution uses an extensive research and experimental approach to formulate two separate prediction models; one for classifying whether the customer is prone to buy a few expensive products and the other to predict the amount of money they usually spend in a month. However, we share a common workflow for conducting research on both these problems. The common workflow includes various phases, namely, Data pre-processing and feature selection, Splitting of Pre-processed Data into train and test set, training and testing of models respectively. The diagram showing the overall system workflow is shown below:

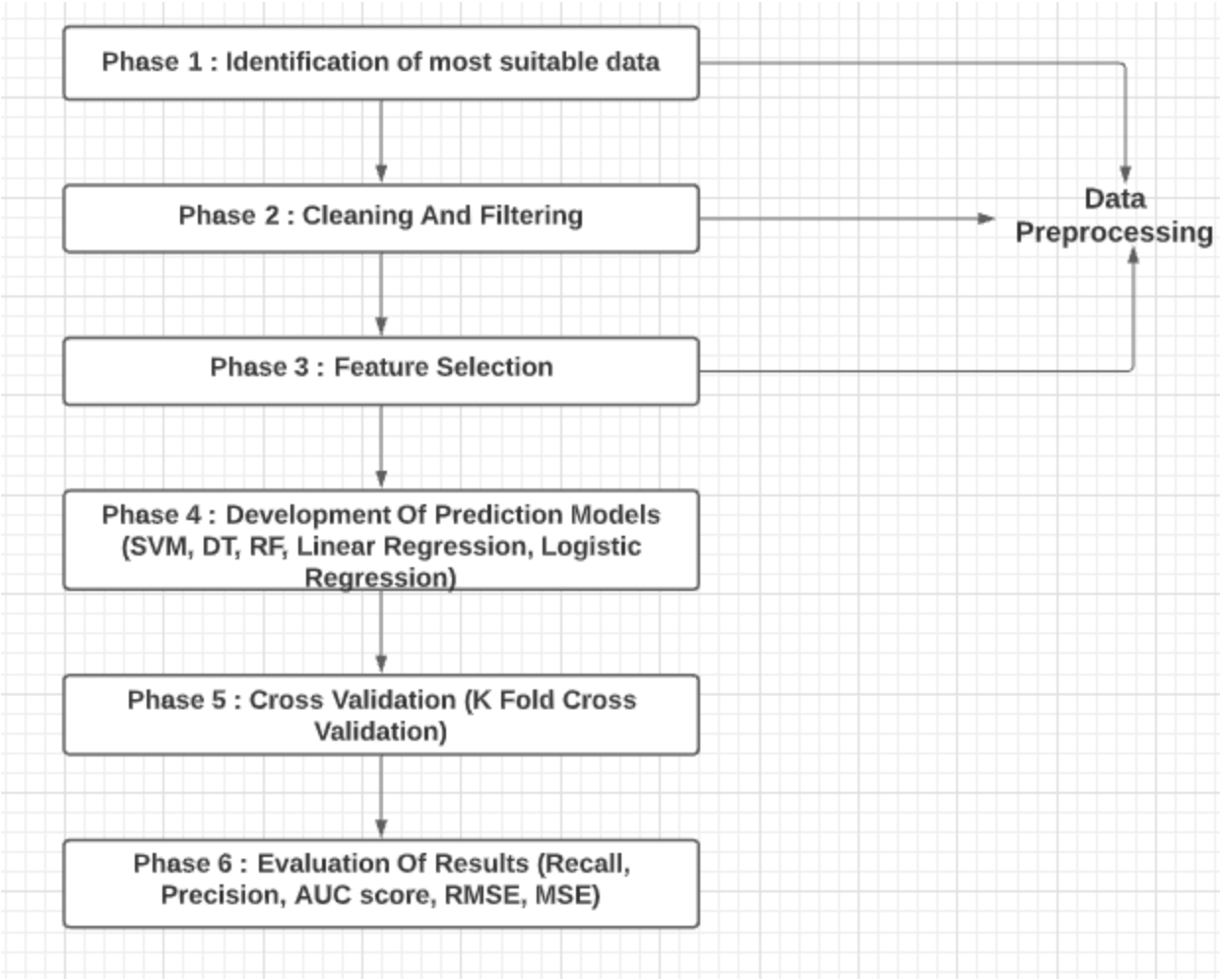

Figure 1 : Workflow Of the Proposed Solution

Phase 1 - Identification of most suitable data :

Task 1 :- For the first task, we are provided with a training set of historical data containing features of each customer and a label representing whether the customers belong to the first class or not. There are two datasets provided in 'csv' format one for training and one for testing out the prediction. There are 1500 data samples for training and 1500 data samples for testing the result. The datasets consist of 15 unique features and a target column out of which two of the features are of integer data type and remaining are of float data type. The target column is a boolean value indicating True for high class customers and False for low class customers. In the testing dataset, the target column has 'nan' values which are supposed to be filled with the prediction values.

Task 2 :- For the second task, we are provided with a training set of historical data containing features of each customer and a numerical value representing the value of the amount of money

they spent in May 2021. There are two datasets provided in ‘csv’ format one for training and one for testing out the prediction. There are 1500 data samples for training and 1500 data samples for testing the result. The datasets consist of 16 unique features and a target feature out of which two of the features are of integer data type, two of the features are categorical and remaining are of float data type. The target column is a continuous value indicating the amount of money the customer spent in May 2021. In the testing dataset, the target column has ‘nan’ values which are supposed to be filled with the prediction values.

Phase 2- Data Cleaning and Transformation:

Task 1 :- For the classification task, there are missing values in both training and testing set in one feature column. So, we use four different strategies to deal with the missing values : first we drop the feature column, second by filling the missing values in a column by a mean , third we fill the missing values in a column by median and fourth we fill the missing values using the knn imputation method. The training data is scaled using standardization in case we are using logistic regression and support vector machine algorithms but for decision trees and random forest, we do not apply any data transformation techniques because they are not sensitive to outliers, data distribution and scaling.

Task 2 :- For the regression task, we first encode the categorical variables into a numerical feature because we need to convert our features to numerical values before fitting the data into the model. Then, we look at each feature in the data and detect any outliers present in the data or not since linear regression algorithms are very sensitive to outliers. We use z-score value and IQR value to detect the outliers and find the lower range and higher range of values that we can accept. Finally, for dealing with the outliers, we replace those values which are greater than higher range with the value of higher range and lower than lower range with the value of lower range. We also check the distribution of data in each feature because linear regression assumes that the data must have a gaussian distribution. We found out that the features ‘F1’, ‘F3’, ‘F7’, ‘F9’, ‘F12’ and ‘F16’ are not following the gaussian distribution so we apply log transformation, reciprocal transformation and box-cox transformation techniques to transform these features to follow normal distribution.

Phase 3- Feature Selection

Task 1 :- For the classification task, we have used lasso regression for doing the feature selection when applying logistic regression and support vector machine algorithms. There are 16 features in the training datasets and we need to deal with feature selection when we feed the data to logistic regression and support vector machine algorithms.When we apply lasso regression, the non important features are discarded and not taken for modelling. When we apply lasso

regression in logistic regression, the four features ‘F7’, ‘F9’, ‘F10’ and ‘F12’ are removed from the total features because their coefficients shrank to 0. In this way, feature selection is done. But incase of random forest and decision trees, they automatically do the feature selection when splitting the decision nodes for classification.

Task 2 :- For the second task we didn't do any feature selection technique and fed all available features to model the algorithm.

Phase 4 - Development Of Prediction Models

Task 1 :- For task1, we experimented with four classification models. They are Decision Tree, Random Forest, Logistic Regression and Support Vector Machines. We apply different feature selection strategies and data transformation techniques before fitting the data into the model. The two tree based algorithms : Decision Tree and Random Forest are not sensitive to outliers, distribution of data and scale factor so we don’t apply any feature selection strategies and transformation techniques in our data for fitting our data in these algorithm models. However, logistic regression and support vector machine algorithms are sensitive to outliers, data distribution and scale factor, so we use lasso regression for doing feature selection and standardize our data before fitting it into these models for prediction. We use scikit learn library to develop and create these prediction models and use the available datasets to compare and test our predictions.

Task 2 :- For task2, we experimented with three regression models. They are Decision Tree Regressor, Random Forest Regressor and Linear Regression. Since, Tree based algorithms are not sensitive to outliers, scale factor and distribution, we can remove the burden of data preprocessing pipelines but for linear regression we must take care of outliers, understand data distribution and choose appropriate scale factor. So, we use z-score value and interquartile range for detecting and handling outliers, different transformation techniques for making the distribution normal and normalizing the inputs before feeding the data into the algorithm. We use scikit learn library to develop and create these prediction models and use the available datasets to compare and test our prediction

Phase 5 - Cross Validation and Data Splits

Task 1: - For task1, we have splitted the datasets into training and testing using 80:20 split ratio. We use a 5-fold cross validation strategy to validate our result and average the prediction accuracy to get the overall accuracy of the algorithm.

Task 2 :- For task2, we have also splitted the datasets into training and testing using 80:20 split ratio. However we did not do any cross validation strategy to further validate our result.

Phase 6 - Evaluation Of Results

Task 1 :- For the classification task, we have used accuracy score as an evaluation metric. We calculate the accuracy score of all four different classification models taking the missing values into consideration and compare their results based on the accuracy score. The model with higher accuracy score is selected as the model for doing the prediction on a hold-out test set.

Task 2 :- For the regression task, we have used mean squared error and root mean squared error as an evaluation metric.We calculate the 'rmse' and 'mse' score of all three different regression models and compare their results based on this evaluation metric. The model with low 'rmse' and 'mse' score is selected for doing the prediction on a hold-out test set.

6. Findings

Taks1 : Classifying whether the customer is prone to buy few expensive items

Removing missing column

Algorithms	Cross Validation Score (5 Fold %)
Decision Tree	77.06
Random Forest	83.33
Logistic Regression	79.39
Support Vector Machines	80.33

Table 1 : Algorithm Comparison table for data with missing column

Filling missing values with the mean of the overall distribution

Algorithms	Cross Validation Score (5 Fold %)
Decision Tree	78.4
Random Forest	87.2
Logistic Regression	84.26

Support Vector Machines	85.66

Table 2 : Algorithm Comparison table for data when missing values filled with mean

Filling missing values with the median of the overall distribution

Algorithms	Cross Validation Score (5 Fold %)
Decision Tree	78.26
Random Forest	87.26
Logistic Regression	84.33
Support Vector Machines	85.46

Table 3 : Algorithm Comparison table for data when missing values filled with median

Filling the missing values with the KNN imputation method

Algorithms	Cross Validation Score (5 Fold %)
Decision Tree	79.2
Random Forest	86.46
Logistic Regression	82.6
Support Vector Machines	85.2

Table 4 : Algorithm Comparison table for data when missing values filled with KNN

From the comparison table above, we found that the random forest algorithm outperforms all other models in predicting whether the customer buys a few expensive products or not. We get an accuracy score of about 87.2% when cross validation using 5 folds. The accuracy of the random forest algorithm is higher when we fill the missing values with the mean and median of the overall data. However when we apply the random forest algorithm by dropping the missing column, we get an 83.33% accuracy. The second best performing model is the Support Vector Machine algorithm with 85.66% accuracy when filling the missing values with mean.

Task 2 : Predicting the amount of money the customer usually spend in a month

Algorithms	Root mean square (RMSE)	Mean square error (MSE)
Linear Regression	533.2973399815393	284406.0528313855
Decision Tree Regressor	846.718291562537	716931.8652665815
Random Forest Regressor	605.3546466519803	366454.2482231439

Table 5 : Algorithm Comparison table for different regression algorithm

From the above table, we found that the linear regression algorithm is the best algorithm for predicting the amount of money the customers will spend in a month. We experimented with three different regression based models and linear regression outperforms two tree based regression models with rmse of 533.29 and mse of 284406.05. The second best performing model is Random Forest Regressor with rmse of 605.35 and mse of 366454.24.

7. Conclusion

From the research project, we conclude that the random forest algorithm can categorize whether a customer visiting the supermarket will end up purchasing expensive items or not with an accuracy of 87.2% which is a fairly good model. Similarly, linear regression outperforms tree based regression techniques for predicting the amount that a customer can spend in a month. The rmse score of the model is 533.29 which is better compared to other model's rmse score.