Report on the Investigation of the customer details of supermarkets.

Word count : 1134

Registration number: 2010557

This report consists of brief synopsis of the procedures used to analyze the class of the customers who buy a few expensive products and who buy many cheap products in the supermarket Tosco and Spency. An experiment on prediction of the expenditure of a new customer at Sunsbory's based on the data of other customers is also carried out.

DATA PREPROCESSING:

TOSCO AND SPENCY

Initially the necessary libraries are imported and the data is read into a data frame called df. The beginning rows of the data are inspected to check the structure of the data and get an overview of the features. The data types of the features are inspected to check if there are any categorical features. The data does not have any categorical features and hence the data is checked for missing values. The column “F15” consists of 750 missing values. Therefore, the missing values have been replaced with the mean of the column for efficient analysis.

The data is divided into input and target variables to classify the Class variable which is the target variable and the rest of them are input variables.

Similar pre processing is performed on the test data given in which the class of the customer is to be predicted and the input, target features are separated as well in the test data set.

SUNSBORY’S

Similar analysis is performed for Sunsbory’s data but, the data consists of two categorical features(F4,F5). The data is One Hot Encoded to convert categorical data into a binary vector representation that can be used in machine learning techniques using pd.get_dummies(). The encoded data frame is merged with the original data frame to perform the analysis. The input features do not contain any missing values.

Similar operations are performed on the test data set given in which the expenditure of a new customer is empty and must be predicted.

When the target amount is inspected and a histogram is plotted, it is observed that majority of the customers spent below 500 units of money in the month of May 2021 at sunsbory’s. The results can be viewed below in fig1.

The correlation of the variables is checked and it is observed that the variables are not much correlated. Hence all the variables are used for analysis. The correlation of the variables is shown in fig2 below.

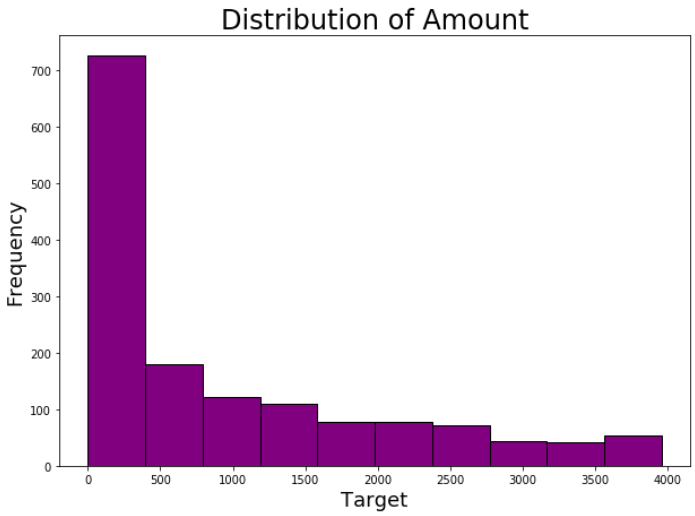

Fig1

	F1	F2	F3	F6	F7	F8	F9	F10	F11	F12	...	F15	F16	Target
F1	1.000000	0.017902	-0.004979	-0.006548	0.002309	-0.012594	-0.012025	0.064451	-0.050148	-0.000822	...	-0.013438	0.061850	-0.015412
F2	0.017902	1.000000	0.018861	0.059413	0.026478	0.046542	0.037862	-0.013474	-0.022325	-0.028201	...	0.060349	0.012850	0.357137
F3	-0.004979	0.018861	1.000000	-0.014330	-0.008484	0.012457	-0.017546	-0.006276	-0.021228	-0.026931	...	-0.042207	0.039825	-0.011330
F6	-0.006548	0.059413	-0.014330	1.000000	-0.027544	0.011987	0.032332	-0.013561	-0.005763	0.046442	...	0.066430	-0.040372	0.027713
F7	0.002309	0.026478	-0.008484	-0.027544	1.000000	-0.009240	-0.031183	0.007359	0.014407	0.016949	...	-0.012488	0.016704	0.031407
F8	-0.012594	0.046542	0.012457	0.011987	-0.009240	1.000000	0.000302	-0.011457	0.000099	-0.018839	...	0.007560	0.051243	0.430990
F9	-0.012025	0.037862	-0.017546	0.032332	-0.031183	0.000302	1.000000	-0.004793	0.009609	-0.002310	...	-0.006157	0.004091	0.020443
F10	0.064451	-0.013474	-0.006276	-0.013561	0.007359	-0.011457	-0.004793	1.000000	0.005589	0.006516	...	0.004620	0.002825	-0.236200
F11	-0.050148	-0.022325	-0.021228	-0.005763	0.014407	0.000099	0.009609	0.005589	1.000000	0.003647	...	-0.007184	0.041882	0.346373
F12	-0.000822	-0.028201	-0.026931	0.046442	0.016949	-0.018839	-0.002310	0.006516	0.003647	1.000000	...	0.004651	-0.007670	0.193704
F13	0.001042	0.005397	-0.019297	0.006771	0.072085	0.006295	-0.023763	-0.013712	-0.001317	0.014783	...	-0.010391	-0.007113	0.300890
F14	-0.035125	-0.032430	0.025476	-0.001758	0.006407	-0.027557	0.020489	-0.049907	-0.036053	-0.022900	...	-0.024939	0.000184	-0.018677
F15	-0.013438	0.060349	-0.042207	0.066430	-0.012488	0.007560	-0.006157	0.004620	-0.007184	0.004651	...	1.000000	0.007832	0.029194
F16	0.061850	0.012850	0.039825	-0.040372	0.016704	0.051243	0.004091	0.002825	0.041882	-0.007670	...	0.007832	1.000000	0.263022
Target	-0.015412	0.357137	-0.011330	0.027713	0.031407	0.430990	0.020443	-0.236200	0.346373	0.193704	...	0.029194	0.263022	1.000000
F4_Low	0.020880	0.042147	0.032147	-0.026531	0.027511	-0.024045	0.066768	-0.027504	-0.024725	-0.051716	...	0.053573	0.022652	-0.084860
F4_Medium	0.026437	0.005945	-0.028772	-0.015921	0.004579	-0.031939	-0.022820	-0.040788	-0.013344	-0.006355	...	-0.042912	-0.028305	-0.015035
F4_Very high	0.002943	0.011823	0.008418	0.013100	0.006224	-0.000138	-0.040894	-0.001281	0.030851	-0.028251	...	-0.040524	-0.001507	0.154642
F4_Very low	0.024656	-0.023717	0.022267	0.052144	-0.012380	-0.012555	-0.028270	0.030762	-0.002469	0.014700	...	-0.005206	0.003444	-0.165111
F5_Rest	-0.013766	-0.029161	0.019092	-0.027263	-0.003016	0.010444	-0.007310	-0.025410	0.026506	-0.007810	...	0.006720	0.010587	0.271030
F5_UK	0.002369	0.014971	-0.036029	-0.023508	0.003473	-0.013182	-0.029612	-0.005872	-0.010787	0.006498	...	-0.014913	-0.006636	-0.037504
F5_USA	0.052827	-0.018869	0.023105	0.050749	0.025547	0.000654	0.068829	-0.013735	-0.035406	-0.012512	...	0.000020	-0.028358	-0.006765

Fig2

The below boxplot (fig3) demonstrates the amount spent by the customer and the level of expenditure(high,low,medium).

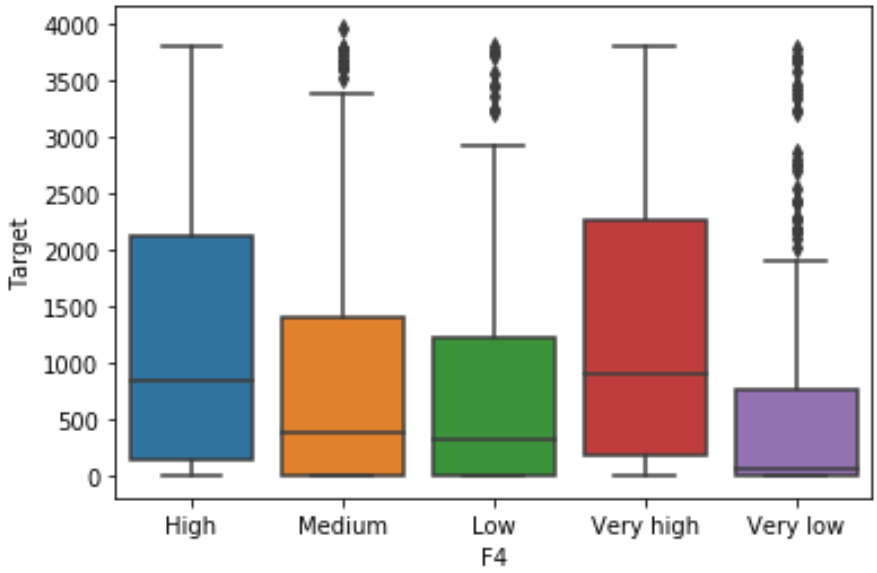

Fig3

The training data of both the supermarket chains is divided into training and test sets using sklearn's train test split function in order to evaluate the built model on the test set before making predictions on new(unseen) data.

In the case of Tosco and Spency a new customer is classified as TRUE or FALSE using three classification algorithms.

- Decision Tree Classifier
- Random Forest classifier
- Naïve Bayes Classifier

DECISION TREE CLASSIFIER:

A decision tree classifier object Is created and the object is used to fit the classifier on the training data (X_train, y_train). The classifier object is then used to predict the outcomes on the test data(X_test). When the accuracy of training and test data is evaluated, the scores are 100% and 80.22% respectively. To further improve the model's performance, grid search cross validation is implemented to find the best parameters. The model is again fit on the training data using the best parameters obtained in the grid search. The predictions are made using the best fit model and the accuracy on test set is noticed as 80.44%. The accuracy of the model did not improve much after the grid search.

RANDOM FOREST CLASSIFIER

A random forest classifier object is created and is fit on the training data with a random state 40. Random state is set to obtain the same results every time the model is executed. The accuracy of predictions on the test set using the model is 86.22% which is better than the decision tree classifier. Randomized search cross validation is implemented to tune the hyper parameter parameters. The best results obtained in the randomized search are passed to the random forest classifier and the model thus obtained is used to train the training data. The same model is used for predictions on the test set and given unseen data set(new

data set). The test accuracy after hyper parameter tuning is noticed as 86.44%. Thus the accuracy improved after performing randomized search cross validation.

NAÏVE BAYES CLASSIFIER.

The data is trained using naïve bayes classifier and predictions are made on the test dataset. The accuracy of predictions is observed as 60.66%. A grid search is performed to improve the accuracy and a new value for var_smoothing is obtained as the best parameter. The value is passed to the naïve bayes classifier and again trained on the training data. When the classifier is used for predictions on the test data, the accuracy remained the same (60.66%) which implies the grid search did not improve the results.

On comparing all the three classification techniques, Random Forest classifier has given better accuracy and is consequently used to make predictions on the new customers. The predictions are transferred to the CE802_P2_Test file.

In the case of Sunsbory's, the expenditure of a new customer is predicted using using three regression algorithms.

- Linear Regressor
- Random Forest Regressor
- XGBoost Regressor

LINEAR REGRESSOR

A linear Regression model is fit on the training data using a linear regression model object and predictions are made on the test data. The r squared value of the predictions and true values is obtained as 0.80 which implies that the model fits 80% of the data. Fig4 indicates the distribution of true values and the predicted values.

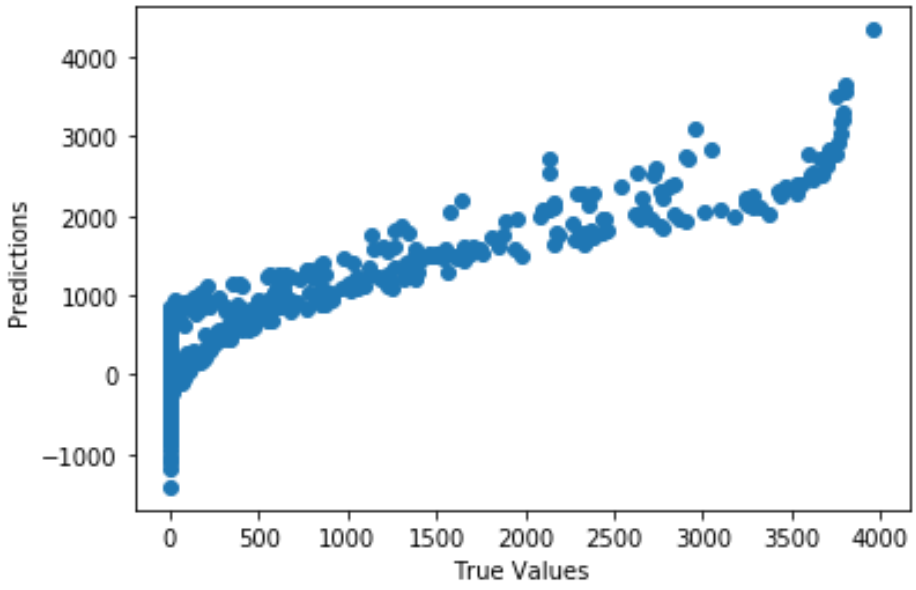

Fig 4

The predictions are made on unseen data and the Mean squared error of linear regression model is 256174.53006926703.

RANDOM FOREST REGRESSION

A random forest Regressor object is built and fitted with a random state 42 on the training data. The hyp er parameter parameters are tuned via randomized search cross validation and parameters like number of trees, number of splits are adjusted. The model is trained on the training data and predictions are mad e on the unseen data after the model is evaluated on the test data. The Mean Squared Error of Random F orest Regressor is observed as 434225.4025427875.

XGBOOST REGRESSOR

The XGBoost regressor is built on the training data using KFold cross validation. The model is used to train the data and is used to predict on the test data and unseen data. The mean squared error of the pr edictions made using xgboost is 303769.33951614687.

When compared to Random forest regressor and Xgboot regressor, Linear regressor has better mean sq uared error and therefore is used to make predictions on the unseen data set.
The expenditure of a new customer is predicted using linear regression and the predictions are transferr ed to the file CE802_P3_Test.csv.