Report On Investigation

Assignment: *Design and Application of a Machine Learning System for a Practical Problem*

CE802-Machine Learning And Data Mining

Name: Thara Jeni Samala Anthony Raj

Reg. No: 2200940

MSC Data Science with Professional Placement

Word Count: 1185

Introduction:

A manager from AENERGY would like to investigate the feasibility of using machine learning to predict whether a customer is going to suffer from the increasing energy prices or not. Two tasks have been given to predict the best machine learning model for predictions.

Short Description:

The given two tasks are:

- Comparative study
- Additional Comparative Study

1) Comparative Study

In this comparative study, "CE802_P2_Data.csv" labeled data set was given to investigate the machine learning procedures, mainly by classifiers (binary classification type) and asked to find the best predictive model.

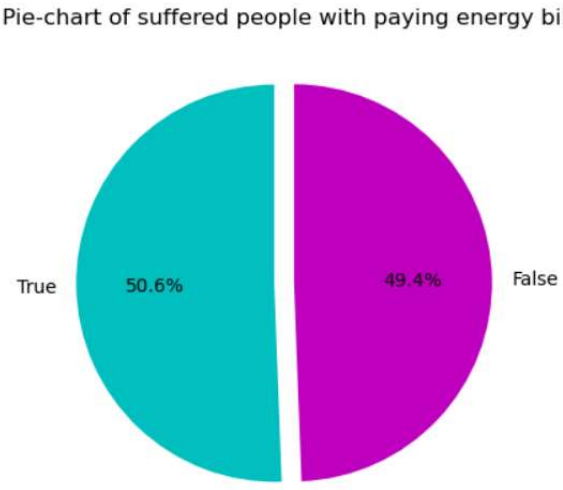
Data Analysis:

Analyzed the shape, feature information and value count of the data set. The given data has 1000 rows and 22 columns.

Data Visualization:

Visualizing the Class of the given data, where

- ✓ *True* indicates the Customers who were suffered from the increasing energy prices &
- ✓ *False* indicates the Customers who weren't suffered from the increasing energy prices.



Pre-processing:

While in pre-processing, we found 500 null values in F21 of the given historical data. Handling those missing values in two ways, one is by drop method & another one is by imputation method.

Finally we have two null-free cleaned data sets,

- **Dropped_Historical_Data** - after deleting the column which had null values
- **Filled_Historical_Data** - after filling the null values with mean

Furthermore, proceeding machine learning procedures with these two data sets to find that which method is good to handle the missing values and also to investigate the best predictive model. Next is the data transformation, splitting data into training and testing set.

Independent Variables: F1 to F21
Dependent Variable: Class

Investigating the performance of ML Techniques:

Here we are investigating the performance of five Machine Learning Procedures such are:

- Decision Tree Classifier
- K-Nearest Neighbor Classifier
- Support Vector Machine Classifier
- Random Forest Classifier
- Stochastic Gradient Descent Classifier

To predict the best model, these five learning procedures are performed. And to find the best accuracy, each model is comparing with Dropped_Historical_Data and Filled_Historical_Data.

Predictive Analysis:

To predict best Machine Learning procedure, comparing the accuracy value & confusion matrix of five performed procedures with Dropped_Historical_Data and Filled_Historical_Data.
Note:

- Accuracy_1 and Confusion Matrix_1 refers Dropped_Historical_Data
- Accuracy_2 and Confusion Matrix_2 refers Filled_Historical_Data

Table.1: Accuracy value of Five Classifiers

Classifiers	Accuracy_1	Accuracy_2
DTC	0.829167	0.8375
KNNC	0.629167	0.629167
SVMC	0.6375	0.6625
RFC	0.854167	0.825
SGDC	0.5875	0.604167

Table 2: Configuration matrix of Five Classifiers

Classifiers	Confusion Matrix_1	Confusion Matrix_2
DTC	[[88 35] [6 111]]	[[95 28] [11 106]]
KNNC	[[76 47] [42 75]]	[[76 47] [42 75]]
SVMC	[[79 44] [43 74]]	[[81 42] [39 78]]
RFC	[[99 24] [11 106]]	[[99 24] [18 99]]
SGDC	[[69 54] [45 72]]	[[88 35] [60 57]]

Table 3: Score values of Five Classifiers with Dropped_Historical_Data

	Classifier	Score
0	Decision Tree Classifier	0.841667
1	K-Nearest Neighbor Classifier	0.629167
2	Support Vector Machine Classifier	0.637500
3	Random Forest Classifier	0.854167
4	Stochastic Gradient Descent(SGD) Classifier	0.520833

Table 4: Score values of Five Classifiers with Filled_Historical_Data

	Classifier	Score
0	Decision Tree Classifier	0.854167
1	K-Nearest Neighbor Classifier	0.629167
2	Support Vector Machine Classifier	0.662500
3	Random Forest Classifier	0.825000
4	Stochastic Gradient Descent(SGD) Classifier	0.658333

From the obtainted *Scores*, *Accuracy* & *Confusion Matrix* in the above tables, we got

- **Decision Tree Classifier** predicts **84%** (with Dropped_Historical_Data) and **85%** (with Filled_Historical_Data)
- **Random Forest Classifier** predicts **85%** (with Dropped_Historical_Data) and **83%** (with Filled_Historical_Data)

Result:

Since now we have investigated the five Machine Learning Procedures along with two data sets, we obtained similar results between Dropped_Historical_Data & Filled_Historical_Data and also if we have more than 75% null values in a column, would suggest best method to deal missing values is drop. But from our hypothesis, the given data set has only 50% null values. so here the best method to deal missing values is filling them with imputation methods.

Since we are considering Filled_Historical_Data as best, we conclude that the Decision Tree Classifier has best performance with 85% of predictions.

2) Additional Comparative Study

In this additional comparative study, "CE802_P3_Data.csv" labeled (numerical values) data set was given to investigate the machine learning procedures, mainly regression type problem and asked to find the best predictive model.

Data Analysis:

Analyzing the given training set of historical data which contains features of each customers & a numerical value representing the variation in the annual expenditure of a customer who had due to increase of the energy cost, where positive sign indicates the customer who spend more and negative sign indicates the customer who spend less.Analyzing the shape, feature information and value count of the data set. The given data has 1500 rows and 37 columns.

Pre-processing:

There is no null values in the given data set but we found categorical values in F5 column and F34 column.

Categorical Values:

In the given data, **F5** have an ordinal relationship, so we transform directly to numerical value respecting the order.

```
In [12]: Historical_Data['F5'].value_counts()

Out[12]: Very low     307
         High         307
         Very high    306
         Medium       294
         Low          286
         Name: F5, dtype: int64
```

Here Very low, Low, Medium, High and Very high are replaced by numerical values.

But **F34** have no ordinal relationship, so using one-hot encoding method, F34 replaced into numerical values.

```
In [14]: Historical_Data['F34'].value_counts()

Out[14]: Europe    393
         USA       378
         Rest      365
         UK        364
         Name: F34, dtype: int64
```

Finally, we got the cleaned data to start the investigation of machine learning procedures.

Table 5: Cleaned_Historical_Data

```
Cleaned_Historical_Data.head()
```

2	F3	F4	F5	F6	F7	F8	F9	F10	...	F31	F32	F33	F35	F36	Target	F34_Europe	F34_Rest	F34_UK	F34_USA
8	-177.02	137.97	2	4.08	-140.19	15882.39	-3728.44	-145.22	...	-2.16	4.84	20.70	-413.31	-1451.18	3179.17	0	0	0	1
4	-202.85	85.50	2	16.80	-248.01	10796.76	-2575.88	-233.30	...	-3.84	11.47	20.94	-242.97	-1583.66	2784.99	1	0	0	0
0	-185.33	61.49	-1	9.12	-151.20	8993.31	-4532.70	-176.56	...	-12.90	1.95	17.37	-257.25	-1360.91	1174.61	1	0	0	0
6	-206.10	114.58	2	19.23	-161.76	8527.65	-4896.54	-108.32	...	-9.26	16.53	16.41	-419.46	-744.03	453.84	1	0	0	0
0	-172.72	95.61	-2	12.33	-162.03	17019.57	-4590.98	-143.82	...	-11.68	13.95	18.87	-353.49	-1026.62	-402.80	0	0	1	0

Next is the data transformation, splitting data into training and testing set.

Independent Variables: F1 to F36
Dependent Variable: Target

Hyperparameter tuning on Multiple Models – Regression:

Before going to investigate the learning procedures, hyperparameter tuning on multiple regression models were done to know their Training time, Prediction time, R2_score & Mean absolute error, By this, it is easy to get the good models by their performance to investigate further with GridSearchCV.

```
RandomForestRegressor()
        Training time: 1.178s
        Prediction time: 0.000s
        Explained variance: 0.6651158737033516
        Mean absolute error: 592.7886803333333
        R2 score: 0.6395000909325298

DecisionTreeRegressor()
        Training time: 0.019s
        Prediction time: 0.000s
        Explained variance: 0.2536555590544297
        Mean absolute error: 788.6279
        R2 score: 0.22621830317082225
```

```
LinearRegression()
        Training time: 0.040s
        Prediction time: 0.000s
        Explained variance: 0.6845000159668364
        Mean absolute error: 570.6678333333333
        R2 score: 0.6827794462535388

SVR()
        Training time: 0.053s
        Prediction time: 0.029s
        Explained variance: 0.013406738244474248
        Mean absolute error: 993.7505852719011
        R2 score: 0.00019862151808713868
```

From the above process, three regression models are selected based on their score which are higher. From this step, we obtained Linear Regression (68%), Random Forest Regression (63%) and Decision tree regression (26%) as good procedures to proceed further for investigation.

Investigating performance of ML Techniques:¶

Here we are investigating the performance of three Machine Learning Procedures such are:

- Linear Regression
- Random Forest Regression
- Decision Tree regression

Linear Regression:

Here, the obtained best Parameters of Linear regression are the same as their default parameters. And also after GridSearchCV, we obtained the same score, so proceeding with the obtained Linear Regression score.

Table 6: Score of Linear Regression

Regression	Scores
LinearRegression	0.682779

From above score, we obtained the best Linear Regression Score(68%).

Random Forest Regression:

Table 7: Scores of Random Forest Regression

Random Forest Regressor	Scores
Before GridSearchCV	0.632836
After GridSearchCV	0.473712

From above scores, we obtained Random Forest Regression Score before GridSearchCV (63%) as best.

Decision Tree Regression:

Table 8: Scores of Decision Tree Regression

Decision Tree Regressor	Scores
Before GridSearchCV	0.273345
After GridSearchCV	0.369091

From above scores, we obtained Decision Tree Regression Score after GridSearchCV(36%) as best.

Predictive Analysis:

Here, comparing the above obtained best scores of performed Regression Models.

Table 9: Scores of Three Regressions

Regressors	Scores
Linear Regression	0.682779
Random Forest Regression	0.473712
Decision Tree Regressor	0.273345

From above comparison table, we conclude that the Linear Regression has best performance with 65% of predictions.

Conclusion:

Two tasks were completely investigated and best models were chosen. With the chosen best linear procedures CE802_P2_Test and CE802_P3_Test data sets were predicted by chosen approaches and missing class and target replaced with the output predictions.

References:

1) R. Saravanan and P. Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification," *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2018, pp. 945-949, doi: 10.1109/ICCONS.2018.8663155.
2) M. Somvanshi, P. Chavan, S. Tambade and S. V. Shinde, "A review of machine learning techniques using decision tree and support vector machine," *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, Pune, India, 2016, pp. 1-7, doi: 10.1109/ICCUBEA.2016.7860040.
3) Liang Zhang, Jin Wen, Yanfei Li, Jianli Chen, Yunyang Ye, Yangyang Fu, William Livingood, "A review of machine learning in building load predictions", Applied Energy,Volume285,2022,116452,ISSN03062619,doi:10.1016/j.apenergy.2021.11645.