Abstract

Ab

**Abstract**

In this project, we were asked to experiment with a real-world dataset and to explore how machine learning algorithms can be used to find the patterns in data. We were expected to gain experience using common data-mining and machine learning libraries such as scikit-learn and pandas. We were also expected to submit a report about the dataset and the algorithms used. After performing the required tasks on a dataset of my choice, herein lies my final report.

*Keywords:* Machine Learning, Classification, Supervised learning, Linear Regression, Regression, Machine Learning Model.

**CE802\_Assignment\_2021**

**Introduction:**

Machine learning is a sub-domain of computer science that evolved from the study of pattern recognition in data, and also from the computational learning theory in artificial intelligence.

Machine Learning can be thought of as the study of a list of sub-problems, viz: decision making, clustering, classification, forecasting, deep-learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic algorithms, sparse dictionary learning, etc. Supervised learning or classification is the machine learning task of inferring a function from a labeled data.

In this assignment, we are going to focus on Supervised learning. A wide array of supervised machine learning algorithms are available, for example, Neural Networks, Decision Trees, Support Vector Machines, Random Forest, Naïve Bayes Classifier, Bayes Net, Majority Classifier, etc., and each algorithm has its own merits and demerits. There is no single algorithm that works for all cases, as merited by the No free lunch theorem. In this project, we try and find patterns in datasets. Our project is divided into two parts, one problem is a classification problem and the other is a regression problem. Train and test dataset is provided for both parts. We first train various models on the given data, and see what sticks.

**Datasets:**

**For Classification Problem:**

The dataset used in the classification part is the customers' data collected by Tosco & Spency over time. The dataset has the following columns:

|  |  |
| --- | --- |
| **Columns** | **Data types** |
| F1 | int64 |
| F2 | float64 |
| F3 | float64 |
| F4 | float64 |
| F5 | float64 |
| F6 | float64 |
| F7 | float64 |
| F8 | float64 |
| F9 | float64 |
| F10 | float64 |
| F11 | float64 |
| F12 | float64 |
| F13 | float64 |
| F14 | int64 |
| F15 | float64 |
| Class | bool |

In the dataset, there are 1500 rows, all columns are of numerical type except the one with the name Class. The class column is the label that we are going to predict. The feature F15 has some missing values, so we will have to deal with them before we pass them to machine learning models.

**For Regression Problem:**

The dataset used in the regression part is the customers' data collected by Sunsbory’s. The dataset has the following columns:

|  |  |
| --- | --- |
| **Columns** | **Data types** |
| F1 | float64 |
| F2 | float64 |
| F3 | float64 |
| F4 | object |
| F5 | object |
| F6 | float64 |
| F7 | float64 |
| F8 | float64 |
| F9 | float64 |
| F10 | int64 |
| F11 | float64 |
| F12 | int64 |
| F13 | float64 |
| F14 | float64 |
| F15 | float64 |
| F16 | float64 |
| Target | float64 |

In the dataset, there are 1500 rows, all columns are of mixed types some are numerical and some are object types. Object types can be assumed as string types. The Target column is the label that we are going to predict. The features that are of an object type can be converted into numerical types using labeling methods. We cannot pass object types in our machine learning models; the models only accept numerical values.

**Performed Experiments:**

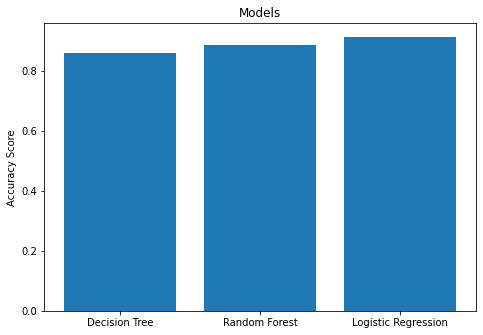
**Classification Part:**

As we know that column F15 had some missing values, we cannot pass them as they are. We choose to remove those rows which had missing values.

We choose three different procedures/machine learning models to evaluate which one is better. Following are the models we tested:

1. Decision Tree Classifier
2. Random Forest Classifier
3. Logistic Regression Classifier

After testing each model, we evaluated each model on the accuracy matric, here is the graph plot where we can see each model's accuracy on the validation set:



As we can see all models performed fairly well, we can’t see a big difference here. But even a little difference matter. After evaluating, we found the Logistic regression Model better among all we tested with 91% accuracy. Therefore, we concluded the Logistic regression model as the final model for the predictions.

**Regression Part:**

We know that columns F4 and F5 are of string type, we will have to convert them to numerical types before passing them into the machine learning models. We converted them into numerical values using the Label encoding method.

We choose three different procedures/machine learning models to evaluate which one is better. Following are the models we tested:

1. Linear Regression
2. Ridge Regression
3. Lasso Regression

After testing each model, we evaluated each model on the Mean squared error matric, which is the most popular and best among all metrics. If the Mean squared error is less then this means our model performed well, so here we will select the model which has the minimum mean squared error. Here is the graph plot where we can see each model's Mean Squared error on the validation set:



Here if we see there is not much of a difference between the mean squared error, but we will choose the one with the least mean squared error score. We choose Lasso Regression because it had 649.93 as compared to other models which had a slightly higher score.

**Results:**

**Classification Part:**

For this problem, we selected Logistic Regression because it had the highest accuracy among those models we tested. We then predicted our test data on Logistic regression.

****

This is our test dataset before the prediction, we can see that the class label has no values because we have not predicted these values yet.

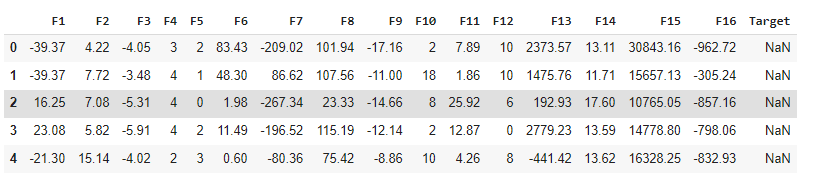
In the next step we will predict these values using logistic regression model. Below we can see the class predicted values after we applied logistic regression.

****

We then exported this predicted dataset to a CSV file.

**Regression Part:**

For this problem, we selected Lasso Regression because it had the lowest Mean squared error among those models we tested. We then predicted our test data on Lasso regression.



This is our test dataset before the prediction, we can see that the target label has no values because we have not predicted these values yet.

In the next step we will predict these values using Lasso regression model. Below we can see the target predicted values after we applied Lasso regression.



We then exported this predicted dataset to a CSV file.

**Conclusion:**

Hence, we conclude that the datasets for the classification part were complete they helped us to predict the customer classes with an accuracy score of 91%. But the datasets provided for the regression part were not enough to predict the target prices, it would have been better if we were provided with a much accurate dataset for this part, our Mean squared error could have been much lower.