HackStat-2.0 Report

Team Leader:

- Kavindu Gayantha

Team Members:

- Nimesha Dilini
- Madushanka Kahawa

University:

- University of Kelaniya

Registered Team Name:

- Exterminators

Kaggle user name:

- exterminators

Kaggle display name:

- Team-exterminators

1. Introduction

This report explains the way we have solved the solution of the first round competition for hackStat-2.0. We had to handle a challenging dataset consisting of data of visitors of a website and predict the class of the type of customer as to whether the customer would be a revenue generating customer or not, by using the revenue variable as the dependent variable and rest of the variables as independent variables. We had to upload the predicted outcome of the test set according to the format provided, to obtain the accuracy of the prediction.

2. Methodology

These are the steps we have followed:

- Clean and Preprocess the dataset
- Identify the best suit classifier
- Final code for the given problem

CLEAN AND PREPROCESS THE DATASET

We used the python libraries ***numpy*** and ***pandas***.

1. Load the dataset csv files. Identify its values and shape of the database.
2. Identify the null values
3. Remove the rows with null values / Fill the null values with mean
4. Identify the data types of the data fields. To identify non numeric (categorical, Boolean) data fields.
5. For categorical data values get the frequency of its data
6. Map the month data field values to a numeric dictionary
7. VisitorType data fields encoded with the binary encoding technique
8. Boolean data type Weekend convert to integer format
9. Save the cleaned dataset to a new file.
   Link to the code: 1.Making Dataset code

IDENTIFY SUITABLE CLASSIFIER

1. Load the trainset dataset
2. Make arrays with given dataset

3. Whole dataset split as train dataset and test dataset
4. Train with some machine learning classifiers and checked the accuracy for each classifier

   Classifiers used:
   - Linear Regression Classifier
   - Decision tree classifier
   - KNN classifier
   - Support Vector Classifier

Link to the Code: 2.Identify the classifier

FINAL CODE FOR THE GIVEN CHALLENGE

1. Load the cleaned datasets
2. Allocate data to arrays
3. Train the dataset with the selected classifier
4. Predict the results for the test dataset
5. Results are taken to a numpy array
6. Save to the submission file as the given format

Link to the code: 3.FinalCode

3. Results

ACCURACY FOR PREDICTED VALUES FROM THE TRAIN DATASET:

1. Linear Regression Classifier - Accuracy: 0.8764312977099237
2. Decision tree classifier - Accuracy: 0.8616412213740458
3. KNN classifier(n_neighbors=7) - Accuracy: 0.8568702290076335
4. Support Vector Classifier - Accuracy: 0.8520992366412213

4. Conclusion

We used the Linear Regression classifier to solve the challenge

Maximum Accuracy that we have achieved: *0·87567*