Cow Activity Prediction using ML

By

P Raja Vamsi Krishna

Objective of the Project : Classify Cow's activities into 9 categories based on Data collected from IMU SENSORS

Importance of the study :- Predicting the activity of the cow over a period of time helps in better livestock management

Data Source for the project: IMU Data (Accelerometer, Gyroscope, Magnetometer) in the form of nine CSV files pertaining to each activity.

The Process

Basic Processes – Importing the basic and required libraries, loading the csv file from Google drive , renaming and reading the csv file.

1. The nine csv files were first loaded onto google drive , the drive was mounted onto google collab notebook, each file was renamed and read.
2. All the files were concatenated into a single dataframe called df. This dataframe had 12263524 records and 11 features.

Analyzing and Visualizing the data set

1. Checked the top 5 rows of the dataset.
2. Checked the number of rows and columns of the dataset.
3. Checked the type of data under each column of the dataset
4. Checked for null values. There were no null or missing values.
5. Checked the description (count, minimum and maximum values, 25-50-75percentile, standard deviation of the data under each column having int64 or Float64 data (the columns were 'SeniorCitizen', 'tenure', 'MonthlyCharges', 'TotalCharges') and transposed the same data.
6. Created a heatmap to visualize the correlation between each numeric variable.
7. Created distribution plot to visualize the skewness of distribution of each numeric variable.
8. Input vairables and output variable were separated into 2 different datasets
9. Created count plot for the output variable ' label ' to visualize the imbalance in the class.
10. The percentage of each output variable class was found out.

Outlier Identification using Box Plot and Treatment using IQR Method

1. Creating a common function to identify outliers for each feature.
2. Plotted box plots for each feature to identify outliers.

3. Createdcommon function to replace outlier values with the median value and print boxplot thereafter to recheck the removal of outliers and print upper and lower values of each Feature.

Data Normalization

1. Normalization of input variables were done using MinMaxScaler.
2. New dataframe was created appending normalized x values and y values.
3. Observed the distribution of target variable using pie chart.

Feature Selection

1. Those input features which impacts the output variable (outcome) the most were selected using Ch2 test . Insignificant features gyr_x, gyr_y, gyr_z , acc_y and mag_z were identified.

Dataset Random Sampling

1. As dataset is large, we took different sizes of dataframes by random sampling: df1 with 30000 rows, df2 with 40000 rows, df3 with 5000 rows and df4 with 60000 rows.
2. Checked the distribution of target classes in each of these samples and compared these with original dataframe and observed that the distributions are closely matching, we just went ahead with samples for ML models building.

Data Split

1. Data Split was done to split the data first into Training and Testing Data in a 70/30 ratio.

Model Building

1. For dataset df1, model built was Logistic Regression.
2. For dataset df2, model built was Decision Tree along with oversampling technique SMOTE.(explained below separately)
3. For dataset df3, model built was Random Forest along with oversampling technique SMOTE. (explained below separately)
4. For dataset df4, model built was SVC. Also Cross Validation is done by GridSearchCV. (explained below separately)

Model Evaluation

1. Evaluation of all the models were done by using Confusion Matrix, Classification Report and accuracy_score.
2. Logistic Regression, Decision Tree, Random Forest and SVM Classifiers returned accuracy scores of 43%, 54%, 91% and 74% respectively
3. For Decision Tree, Criterion taken was ‘Gini’ and Maximum Depth taken was 5.
4. For Random Forest , n_estimators was 10 with Criterion as ‘Entropy’

SVC – Grid Search – Cross Validation

1. For SVC, model parameters and grid parameters were obtained.
2. Cross Validation was done by importing GridSearchCV.
3. The grid was fitted into the training dataset.
4. Inspected the best parameters using GridSearchCV in the best_params attribute and best estimator in the best_estimator_attribute
5. The predictions were re-run and confusion matrix was re-obtained using Y_test and grid_predictions.
6. The re-estimation returned an accuracy score of 68%

SMOTE -

There are imbalance observed in the target classes of the output variable, hence balanced these target classes using over sampling technique called SMOTE

1. Imblearn was pip installed
2. SMOTE was fitted to X_train and Y_train data in Decision tree and Random Forest models.
3. Printed the count of each output class before and after the SMOTE process

Conclusion

Classification models using Random Forest & SVC were the best suited for prediction of cow positions in this project followed by Decision Tree and Logistic Regression.