Machine learning Model to Predict Whether the Road is Safe or Not

Abishek

April 26, 2024

**Abstract**

This report presents a study focused on the machine learning (ML) to optimize travel performance in professional environments. The research centres around the development of a predictive model designed to assess driver eligibility for travel. The model incorporates in study to evaluate the effectiveness of the predictive model in identifying whether the road is safe or not, for traveling based on predefined criteria. Statistical analysis and machine learning algorithm evaluation are employed to assess the model's performance and predictive accuracy. The findings highlight the potential of AI-driven predictive models in enhancing decision-making processes.

**1. Introduction**

* 1. **Background**

with the advancement of technology, particularly artificial intelligence (AI) and machine learning (ML), new possibilities have emerged to reduce life risk of human being. The impetus behind this project using ML algorithm in addressing safe travel or not safe travel on road. The model will predict whether the road is safe for travel or not.

**1.2 Objectives of the Project**

The main objective of this project is to develop a ML model that predict whether the road is safe for traveling or not based predefined criteria.

**Implementation:**

**Tkinter for the GUI**

In the implementation phase, we leveraged Tkinter, a Python library for creating graphical user interfaces (GUIs), to develop the front-end interface of model. Tkinter was chosen for its versatility, ease of use, and seamless integration with Python, making it an ideal choice for rapid prototyping and development. Tkinter offers a wide range of built-in widgets and tools for designing intuitive and interactive user interfaces, including buttons, labels, text boxes, and menus, among others. Its simple and intuitive syntax allows developers to create complex GUIs with minimal coding effort, facilitating the development of user-friendly applications.

Furthermore, Tkinter's cross-platform compatibility ensures that our management tool can run seamlessly on various operating systems, including Windows, macOS, and Linux, without requiring extensive modifications. This universality enhances the accessibility and usability of our tool across diverse organizational environments, enabling widespread adoption and acceptance among users

**Files:**

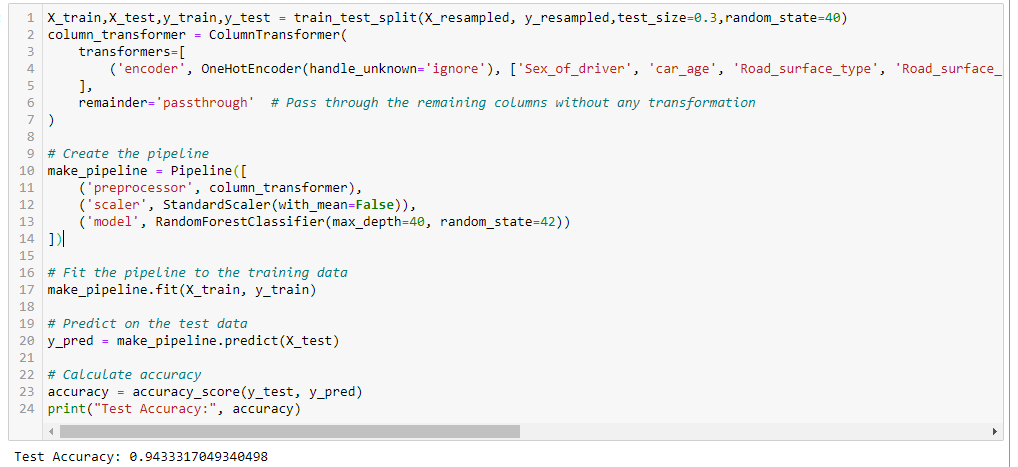
* In main\_model.ipynb files you will found the code of our GUI interface like {login page}
* In predictive\_model.ipynb file you will found a gui interface where user enter data as input to model, once user entered records then by pressing predict key model will predict whether the road is safe or not.

**Note:**

* User will select the source and destination model will fetch all the features of road like[“road\_surface\_type”,” road\_surface\_condition”,”cause\_of\_accident” and “history of accident”] between source and destination from road\_accident\_final.csv file.
* User just have to provide [“Gender”,”light condition”,”weather conditions” and “age of driver”] reaming above features between source and destination will be automatically fetched from road\_accident\_final.csv file.

**Model Traning Code:**

**You will found model training code safe\_road\_copy.ipynb file each line of code is explained with proper comments you can read it overtheir.**

****

* In model training code we used oneHotencoder to convert object type attributes in numerical attributes.
* Then we make a pipeline in which we are passing column transformer, Standard Scaler [to bring data in small interval for the purpose of reducing computational cost].
* And finally we passing our Random Forest model.
* Whenever we predicting unseen records, the records will pass through pipeline and then will predict the target variables.

**Preprocessing**Preprocessing is the techniques applied on data before it is used for analysis or modeling. The purpose of preprocessing is to prepare the data in a way that improves the performance and accuracy of machine learning algorithms.

Preprocessing can involve several tasks, including

**Data cleaning:**

**Duplicates record:**

* removing duplicates ensure data consistency
* Removing duplicates helps prevent over fitting
* It can occupy unnecessary space

In our case there is no any duplicates record

**Handling missing values:**

* It is important to handle missing values before training model
* Many statistical model can’t handle missing values
* Missing values lead model to bias
* It brings inconsistency in dataset

In our case there is no any missing record

**Handling Categorical column:**

Model can understand only numerical values not categorical values this how it mandatory to convert these column in numerical values.

['age\_of\_driver', 'Sex\_of\_driver', 'car\_age', 'Road\_surface\_type', 'Road\_surface\_conditions', 'Light\_conditions', 'Weather\_conditions', 'Cause\_of\_accident']

In above list you can see we have many categorical variables. But we have to convert it into numerical values

**Removing Outliers:**

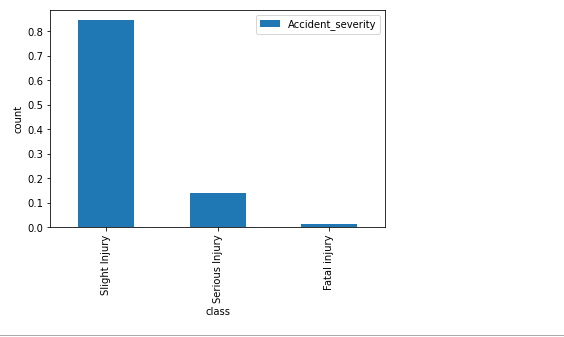
* Handling outliers in machine learning model is very important
* Here in classification outlier effect the decision boundary and it move decision boundary toward outlier class.

**In our case there is no outlier.**

**Unbalanced Distribution of the Target Class:**

* + We can’t train model on unbalance class it may lead the model to be biased
  + It is necessary to balance the class before training model
  + Here in our target class “slight injury” is the majority class, we have to apply a statistical technique to balance the class that does not change the context of original data.
  + Basically we have two main techniques (over sampling and under sampling) but we apply over sampling technique because already our dataset length is less and under sampling technique reduce further dataset so it is good to apply over sampling.

**Imbalance Class**

****

**Methodology**For this project, we trained Random forest classifier

**Classification:**

Predicting passenger survivalis a classification problem because the target variable is the Accident Severity which is a multiclass classification variable and according to model selection it is clear that for predicting categorical or discrete class label we always use classification modelclassification model is a statistical model which involves making statistical inferences based on data in our case we want to predict categorical variable (Accident Severity)

**Random Forest**

After rigorous evaluation and comparative analysis, our investigation culminated in the selection of the Random Forest algorithm as the optimal solution for our management tool. Random Forest combines the strengths of decision trees with ensemble learning, offering robust performance, scalability, and resistance to over fitting. By aggregating predictions from multiple decision trees trained on random subsets of the data, Random Forest achieves high accuracy while maintaining computational efficiency.

Moreover, Random Forest's inherent parallelism and ability to handle categorical and numerical features without extensive preprocessing align with the diverse nature of data encountered in managerial contexts. Its flexibility, interpretability, and resilience to outliers make it well-suited for real-world applications, enabling organizations to derive actionable insights from complex datasets efficiently.

**Analysis:**

**Algorithmic analysis for safe road Classification**

Main model for Passenger Survival Prediction using Random Forest classification

For making main model I am choosing random forest classifier. Because random forest classifier can capture nonlinear relationship and it is also robust to outlier although we minimize outliers in our dataset.

It also helps to reduce over fitting and improve model generalization because; it is an ensemble method that combines multiple decision tree.

It also can handle both binary classification tasks and multiclass classification tasks. it extend and combine multiple decision to build more robust and accurate models to achieve more accuracy..

Although, our problem is multiclass classification task, but we use random forest classifier for obtaining better accuracy, because it can have the ability of capturing nonlinear relationship.

Another importance of random forest is this, that it is useful for feature selections on the basis of Gini information values for each feature.

|  |  |
| --- | --- |
| **Training Accuracy** | **1** |
| **Test Accuracy** | **0.94** |

**Random forest Classifier Metrics**

|  |  |
| --- | --- |
| **Training Accuracy** | **1** |
| **Test Accuracy** | **0.94** |
| **Precision** | **0.95** |
| **Recall** | **0.94** |
| **F1 Score** | **0.94** |

Overall random forest classifier gives high accuracy score.

random forest tree classifier performing better in terms of recall and F1 score. The random forest tree classifier has a higher recall value, indicating its ability to correctly identify positive instances.

Thank you

END