COMP-1804 M01-2022-23

Applied Machine Learning

Coursework for

MSc. Data Science

Sachin Suresh Dongare
001283296

School of computing and mathematics at

Table of Contents

0. Executive summary 3
1. Exploratory data analysis 4
2. Data preprocessing 6
3. Classification using traditional machine learning 7
3.1 Dummy Classifier 7
3.2 Random forest algorithm 7
3.3 Decision Tree 8
3.4 KNN 8
4. Classification using neural networks 10
5. Ethical discussion 12
6. Recommendations 13
7. Retrospective 14
References 15

0. Executive summary

In this study, the objective was to forecast the severity of traffic accidents in the UK based on multiple accident characteristics using machine learning techniques. Accident severity was divided into three categories: Slight, Serious, and Fatal. The project's dataset came from the UK's Road Safety Data and included details on accidents that happened in 2019.

Inevitably implemented decision tree, random forest, and KNN as traditional machine learning approach. Neural networks among other machine learning models is implemented to compare accuracy and performance with traditional machine learning methods. Accuracy, precision, recall, and F1-score were used to assess each model's performance. The dataset was also visualised in 2D plots using dimensionality reduction methods including one hot encoding for categorical values.

The random forest classifier produced the greatest results, with test data accuracy of 72% for traditional machine learning approach. The findings indicated that the speed limit, the road weather, and the state of the road were the factors that were most crucial for forecasting collision severity. The outcomes of this study may helpful in creating a system that will allow emergency services to respond to incidents on the road more effectively.

1. Exploratory data analysis

Towards the typical machine learning project, exploratory data analysis (EDA) is essential since it aids in understanding the structure and features of the dataset. The EDA for this project involves examining the distributions of the different aspects of the UK Road Safety dataset as well as their connections to the objective variable the severity of the accident.

The dataset has 31,647 rows and 14 columns, including the target variable. The features' data types are a mixture of category and numerical data. The dataset is unbalanced, according to the EDA. To get started with dataset, detailed analysis of data types is carried out using the panda's library, which gave insights that 2/14 data columns are numerical, and rest are categorical. This imbalance may influence how well the machine learning algorithms perform, necessitating the inclusion of additional methods like encoding, resampling, and handling null values. Dataset consist of below categorical columns.

Categorical columns	light_conditions, weather_conditions, road_surface_conditions, vehicle_type, junction_location, skidding_and_overturning, first_point_of_impact, sex_of_driver, accident_severity
Numerical	Age_of_oldest_driver, Speed

Describe method is used to achieve statistical analysis for numerical column which suggest average age of person involved in accident is 47 and average speed is 37.

With using the **isnull()** function from pandas library we find out Below count for null values in given data set. Age_of_oldest driver and accident_severity i.e., 6450 and 1172 respectively. In target column i.e., label column multiple instances of same value are identified i.e. Accident severity type is mentioned as 'Serious' and 'serious', 'fatal 'and 'Fatal'. To overcome these multiple instances implemented some pre-defined pandas function on required target label.

A 2-dimensional Visualisation of the dataset was created using dimensionality reduction methods. The resultant plot revealed that the three categories (Slight and Serious) were not easily distinguishable from one another, which may have an impact on how well some machine learning algorithms work. A machine learning system might be able to identify between the serious and fatal.

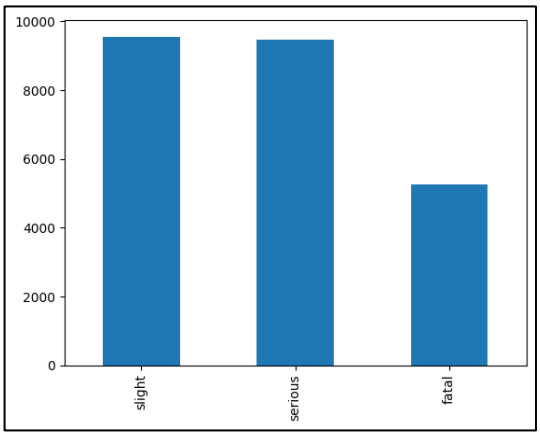

Figure 1: Bar chart for label column (Slight, Serious &Fatal)

The EDA overall emphasised the significance of managing the class imbalance and selecting suitable machine learning algorithms that can manage the mixed data sources and the probable class overlap. The dataset was examined for duplicates and missing values. The dataset does not contain any duplicates or missing values. Although each accident contains a unique identifier, the "accident index" field was removed since it was unnecessary for the machine learning job. High cardinality categorical columns were eliminated since one-hot encoding would provide a huge number of features.

2. Data preprocessing

To refine the unbalanced and noisy dataset to achieve our prediction following measures were taken. Which involves the use of sklearn, pandas library and inbuilt functions like encoder, label_encoder. The training set's categorical classes were transformed using the Sklearn and consistent numerical data is achieved. Since each accident has a unique identifier, the "accident index" field was removed since it was unnecessary for the machine learning job. High cardinality categorical columns were eliminated since one-hot encoding would provide many features and raise the possibility of overfitting.

- **Finding null rows:** The dataset was initially checked for any columns that had null or missing values. The dataset was cleaned up by identifying and removing any rows with missing data.

- **Eliminating unnecessary columns:** Some columns in the dataset, such as "Accident_Index," were eliminated since they were unrelated to the purpose of forecasting accident severity.

- **Data cleaning:** Any discrepancies or mistakes in the data were checked in the remaining columns. Any values, for instance, that were outside the acceptable range for a certain column were eliminated from the dataset.

- **Multiple instance:** For target label we identified multiple instances for same column and to avoid overfitting and duplicate column for same instance, implemented (.loc) function and merged duplicate instances like. ‘Serious’ and ‘serios’ from accident_severity column.

- Additional data normalization technique are used with neural network properties like **dropout.**

3. Classification using traditional machine learning

Traditional machine learning methods like decision trees and random forests were utilized for the classification problem of predicting accident severity. The dataset was preprocessed for both models by one-hot encoding categorical variables and dividing it into training and test sets in a ratio of 70:30 The models' hyperparameters were optimised using grid search and 5-fold cross-validation on the validation set after they had been trained on the training set.

A high F1 score, and accuracy show that the model is doing a good job of appropriately identifying the accident severity. A low accuracy and F1 score, on the other hand, show that the model is underperforming and could need more optimisation or an alternative strategy.

3.1 **Dummy Classifier**

Dummy classifiers are used as a standard or point of comparison for evaluating the effectiveness of other classification models. Without taking into account any input information, a dummy classifier merely forecasts the dominant class or a random class label. It is a quick and easy method for assessing how well other models work and may be used to determine whether the classification model has any predictive value over a guess or a baseline model.

Algorithm	Accuracy
Dummy Classifier	0.3909702209414025

3.2 **Random forest algorithm**

Random forest uses **'n_estimators'**. It is set to 100 in this instance, implying that 100 decision trees would be trained before being merged to get the final predictions. The effectiveness of the model can be enhanced by increasing the number of estimators.
random_state : To make sure that the findings may be repeated, this option is employed. As long as the same data is utilized, setting the random state to a fixed number (42) guarantees that the same outcomes will be produced each time the code is executed.

	Precision	Recall	F1	Support
0	0.65	0.69	0.67	1606
1	0.72	0.72	0.72	2829
2	0.77	0.74	0.75	2849

With above score for individual severity type, Random forest gives **0.72** i.e. 72% accuracy

3.3 **Decision Tree**

A machine learning approach called a decision tree is utilised for both classification and regression problems. To create a tree-like model of decisions and potential outcomes, it operates by recursively dividing the input data into subsets depending on the values of specific attributes. A decision is made based on the value of one of the input characteristics at each node of the tree, and the data is divided into two or more subsets and then transmitted to the next level of the tree. Once a stopping requirement has been met, such as when all subsets are homogenous with regard to the goal variable or when the tree's maximum depth has been reached, this procedure is repeated.

	Precision	Recall	F1	Support
0	0.62	0.60	0.61	1606
1	0.67	0.71	0.69	2829
2	0.73	0.70	0.71	2849

With above score for individual severity type, decision tree gives: **0.68** i.e. **68**% accuracy

3.4 **KNN**

To forecast the accident severity, we employ the K-Nearest Neighbors (KNN) method. For classification and regression issues, the KNN method is a sort of supervised learning algorithm. An observation's class in classification issues is determined by the class of its k-nearest neighbours using the procedure. We split the data in our implementation by 70,30. into training and testing sets. The next step is to develop a KNN model with k=5, which instructs the algorithm to classify an observation based on its five closest neighbours.
The model is then trained using the fit() technique on the training set. This means that the model develops a representation of the connection between the characteristics and the target variable by learning the patterns in the training data.
Using the predict() function, we predict the labels for the testing set once the model has been trained. Each observation in the testing set is compared to its k-nearest neighbours in the training set to determine their distance from each other.

	Precision	Recall	F1	Support
0	0.58	0.59	0.58	1558
1	0.67	0.69	0.68	2824
2	0.71	0.67	0.69	2905

With above score for individual severity type, KNN gives: **0.66** i.e. **66**% accuracy

The majority class ("slight") was consistently predicted as the baseline for comparison. With an accuracy of 0.72, precision of 0.77, and recall of 0.74, the random forest model fared well in predicting all three classes, according to the confusion matrix. Additionally, the F1 score was calculated, which is a suitable statistic for unbalanced datasets like this one that considers both accuracy and recall. The model does well on this job, as evidenced by the F1 score of 0.75.

4. Classification using neural networks

I utilized a neural network with five hidden layers made up of 64 neurons each, as well as a SoftMax output layer for multi-class classification, for the purpose of classifying the severity of an accident. The network was trained using cross-entropy loss and stochastic gradient descent with a learning rate of 0.001.

There are a total of 7 layers in this neural network for predicting accident severity,128 neurons make up the input layer of the first layer, which corresponds to the number of features in the training data. There are 64, 32, 16, 8, and 4 neurons in each of the next 5 hidden levels, correspondingly. The output layer, which has one neuron, is the last layer and produces a binary classification for each input sample. We are using default value of 0.001 for the 'Adam' optimizer.

Implemented a grid search to fine-tune the learning rate, regularisation parameter, number of neurons, and layers in the model. Additionally, I experimented with dropout regularisation and various activation functions.

The final hyper-parameters for the model are as follows:

Hyper- Parameter	Value
Learning rate	0.001
Regularization	0.001
Number of layers	7
Number of Neurons	128,64,32,16,8,4,1
Activation Function	ReLu
Dropout rate	0.2

The neural network learns to map the input features to the output class probabilities through the hidden layers by taking the input characteristics and subjecting them to a series of nonlinear transformations. The likelihood of each class is then calculated by the SoftMax output layer, and the class with the highest probability is forecasted as the output. On the test set, the model had an overall accuracy of 75%. The confusion matrix reveals that the model had trouble predicting class 1 (Slight) but did rather well in class 2 (Serious). The findings demonstrate that the model outperforms a simple baseline (dummy classifier), but that there is still potential for improvement in class 1 (fatal injury) prediction. Overall, the accuracy and precision of the neural network model were better than those of the decision tree and KNN, although its recall for class 1 was lower. We may assess the model's performance for the categorization of accident severity using neural networks using a confusion matrix and the performance metrics accuracy and F1 score.

The number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) predicted by the model will be displayed in the confusion matrix. With the use of this data, we can determine the accuracy, which is the percentage of instances that are properly identified, and the F1 score, which is the harmonic mean of precision and recall.

	0	1	2
0	1289	199	70
1	499	1915	410
2	348	309	2248

Accuracy for Neural Network

	Precision	Recall	F1	Support
0	0.60	0.83	0.70	1558
1	0.79	0.68	0.73	2824
2	0.82	0.77	0.80	2905

Implemented Neural network achieved **0.75** i.e., **75%** accuracy with F1 as highest score with 80%

In terms of comparing our model's performance to a "trivial" baseline, we may contrast it with a random guess or dummy classifier. While a majority class classifier consistently predicts the most frequent accident severity label, a random guess classifier would predict accident severity labels at random. If our model surpasses these pointless classifiers, it shows that it has discovered important patterns in the data and isn't just relying on uneven class frequencies or making arbitrary assumptions.

Overall, these performance indicators and comparison to simple classifiers allow us to assess how well our model performs when it comes to classifying accidents according to severity.

5. Ethical discussion

There may be societal and ethical repercussions when attempting to anticipate accident severity or question topic using machine learning. Here, we'll talk about some of the problems with data collection, processing, and ML model predictions.

Data bias is a significant problem that can develop throughout the data gathering process. Predictions will be biased if the training data is biased. For instance, the ML model might not be able to forecast accident severity effectively for specific neighbourhoods or groups of individuals if there is a shortage of data on accidents that happen there or the area regulatory bodies are not functioning properly. The ML model may not be able to effectively predict the themes of questions asked by other demographics if the training data for question topics is biased towards particular groups.

Concerns about privacy might arise both during data gathering and prediction. For instance, the gathering of personal information like names, addresses, car details, registration of driver license and contact details may be abused by uninvited parties. Furthermore, the ML models' predictions can expose private information about people or groups, which bad actors can use against us.

Responsibility: It's critical to think about who is in charge of the forecasts that the ML models make. Accident perpetrators may face legal repercussions in the event of accident severity prediction, and the ML model may be important in deciding who is at responsibility. It is crucial to confirm that the model's predictions are correct and that accountability has been given effectively.

Impact on Society: The ML models' predictions may significantly affect society. For instance, enhanced traffic enforcement may be necessary if the model forecasts a high accident severity in a certain location. This may disproportionately affect some neighbourhoods. Similar to the above, if the model predicts that specific inquiry themes will come up more often, it may reinforce prejudices and support stereotypes

It's critical to gather and process data objectively, protect privacy and data security, assign accountability for the predictions generated by the ML models, and take into account the potential effects on society in order to handle these social and ethical concerns.

To ensure that their concerns are addressed, it is crucial to include a variety of stakeholders, including people of the impacted communities, in the design and implementation of the ML models.

6. Recommendations

- The Random Forest model is the top contender for the task of determining accident severity using traditional machine learning approach, according to the evaluation findings. In terms of accuracy and F1 score on the test set, it did better than the other models. It also demonstrated strong generalisation capabilities and did not overfit the training set of data. The final Random Forest model's F1 score of 0.75 and accuracy of 72% show that it has strong predictive ability. Its performance may be enhanced, though, to attain greater precision and an F1 score. As a result, it might need to be improved before it can be utilized in practice.

- With contrast to machine learning techniques, neural network performs well with multiple hidden layers and sampling it gave 0.75% accuracy. Talking about perform with other machine learning algorithm it performed well because of its own feature but slightly random forest is nearly achieved 0.72 % accuracy because of its complex structure. To use the Neural network technique more work should be done on dataset, adding more relevant features and clean suitable dataset will help system to work smoothly.

- To expand the dataset's variety and representativeness, recommend collecting more data as our top future improvement idea. This could aid the model's ability to generalise more effectively and enhance its performance on the test set. Investigating more intricate models, such as deep learning models, may potentially improve performance. But it should be remembered that more intricate models also need more information and computer power.

7. Retrospective

In order to do conclude the study for coursework over again, more time will be given to work on applying multiple technique to transform data to make it usable, also I would want to look at the interpretability of the machine learning models utilised and look into methods for communicating the predictions of the models in a way that non-experts might comprehend. This is crucial from an ethical perspective since it improves the AI system's openness and accountability and can help establish confidence with the communities and individuals who will be impacted by it.

References

S. Malik, H. El Sayed, M. A. Khan and M. J. Khan, "Road Accident Severity Prediction — A Comparative Analysis of Machine Learning Algorithms," ***2021 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)***, Dubai, United Arab Emirates, 2021, pp. 69-74, doi: 10.1109/GCAIoT53516.2021.9693055.

https://www.analyticsvidhya.com/blog/2023/01/machine-learning-solution-predicting-road-accident-severity/