Predicting driver behavior from road data using machine learning models

Authors: Tobias Hammarström[1], Chenglong Li[1] and Zhenyu Tang[1] **Supervisors: Daniel Noreland[2] and Gunnar Svenson[2]**

1, Department of Information Technology, Uppsala University

2. Skogforsk, https://www.skogforsk.se/

Introduction

Cutting down fuel costs can increase the competitiveness of Sweden's forest industry. Fuel consumption is influenced by truck weight, road conditions and the driver's behavior, here described by the truck's speed, acceleration, and usage of the brake and clutch.

Current models focus in city and heavy traffic environments, where stochastic events affect and define the driving more frequently. These are unsuitable for route planning on forest roads. A model predicting driver behavior along a selected route of mainly forest roads could assist future models to predict and lower fuel consumption.

One road data set and three truck data sets for empty, half-full and full load conditions, were gathered once in November and once in July, totalling eight data sets.

Models

Loss function and evaluation metric: Speed & acceleration used mean -squared error (MSE); brake & clutch used binary cross-entropy (B. C-E)

Random Forest (RF)
A tree-based algorithm utilizing a form of bagging.

XGBoost (XGB)
A tree-based algorithm utilizing gradient boosting.

Multilayer perceptron (MLP)
(MLP), with dense hidden layers, trained using backpropagation

1D Convolutional Neural Network (1D CNN)
Only convolutes in the time dimension. Some versions have pooling and dropout layers.

Gated Recurrent Unit (GRU)
Used gate recurrent units containing an update gate and a rest gate.

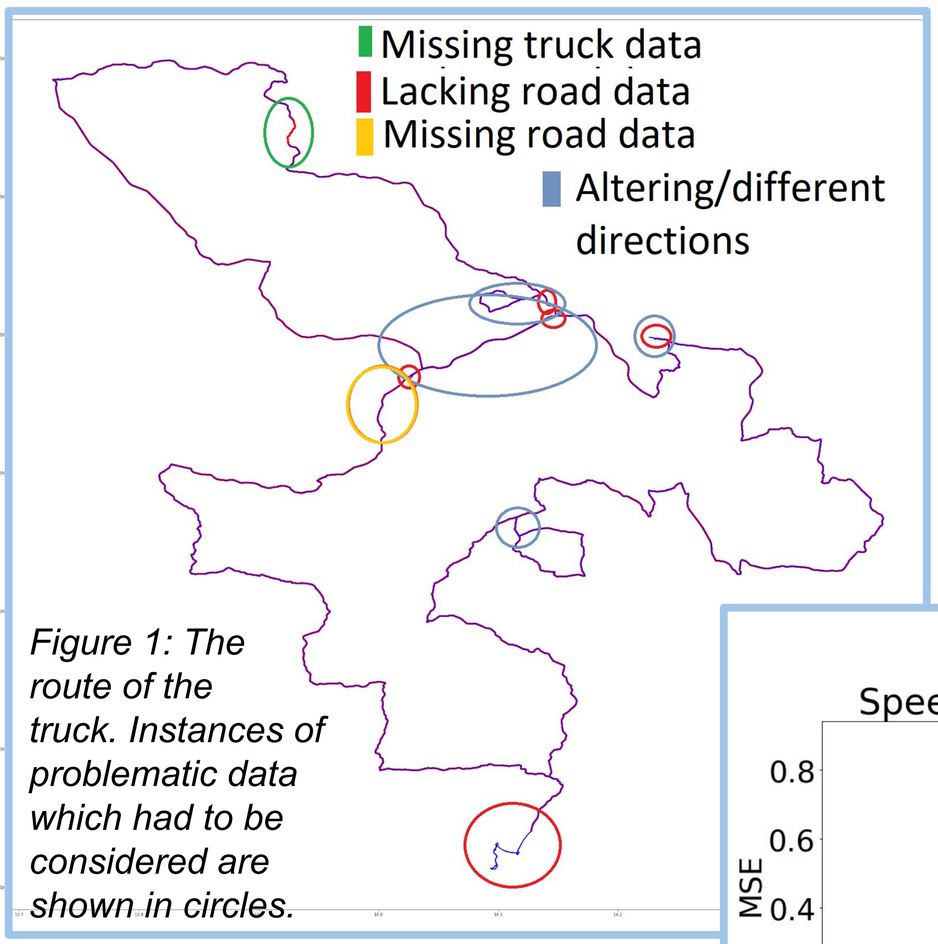

Figure 1: The route of the truck. Instances of problematic data which had to be considered are shown in circles.

Results

RF performed the best and was consequently used as the final model. Its error on test and validation data is similar, which indicate correctness. The mean absolute error of speed, 1.46 m/s, and acceleration, 0.1335 m/s², indicate a reasonable performance given the problem.

The predictions on the brake and clutch show learning difficulties as the performance is only slightly better than assuming neither are used (which would give 0.84 and 0.86 respectively)

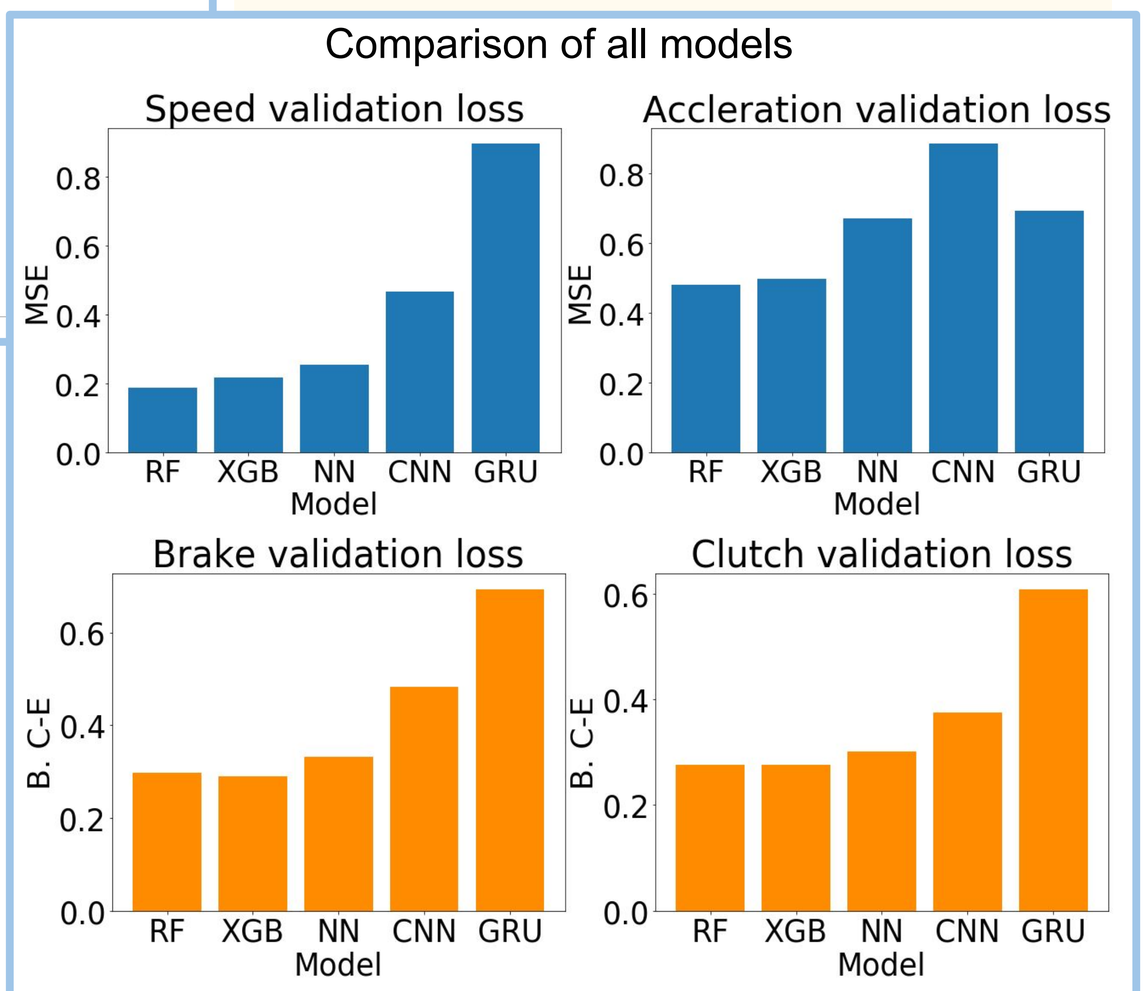

Figure 3: MSE (for speed and acceleration) and B. C-E (for brake and clutch) on validation data in standardized values for the best version of all model types.

Discussion

The preprocessing will affect the models' performance and should be reflected upon. The results show similar performance of the tree-based models, which have outperformed NN-based models on classifications tasks previously, but this result was unexpected.

The MLP has reasonable performance but the 1D CNN and GRU perform poorly, possibly due to interruptions in the sequentiality or their implementation. The 1D CNN could also possibly be unfit for the problem. As the GRU is meant for time-sequential problems, its poor performance was unanticipated. However, other recurrent models should be investigated in the future.

Braking and clutching might be more influenced by visibility, traffic or truck speed than road conditions. As using either is shown on the speed and acceleration indirectly, they seem unnecessary.

Data preprocessing

The merging of the data sets required handling or consideration of the problems in figure 1 as well as other issues:

- The November road data was mostly unusable but road conditions could have changed.
- Extreme values existed in the road data. Impossible and single improbable ones were removed.
- Instances of stopping were identified and removed.
- Acceleration had to be calculated using existing data, and the result's feasibility checked.
- The truck coordinates had to be interpolated to gain sufficient accuracy for matching.

After handling these issues, the data was divided according to figure 2 and normalized using z-score. Finally, correlations were plotted and based on these, an educated guess of which inputs to use was made.

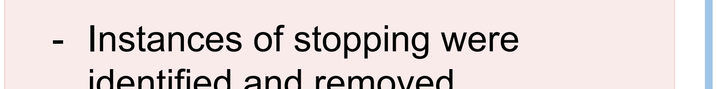

Figure 2: The splitting of training (64%), validation (16%) and testing (20%) data.

Conclusion

On this data, tree-based models performed better than neural network-based. The RF performs reasonably well, but the GRU, specialized in time-sequential problems, performs the worst with no obvious reason as to why.

The brake and clutch showed little correlation with the road data.

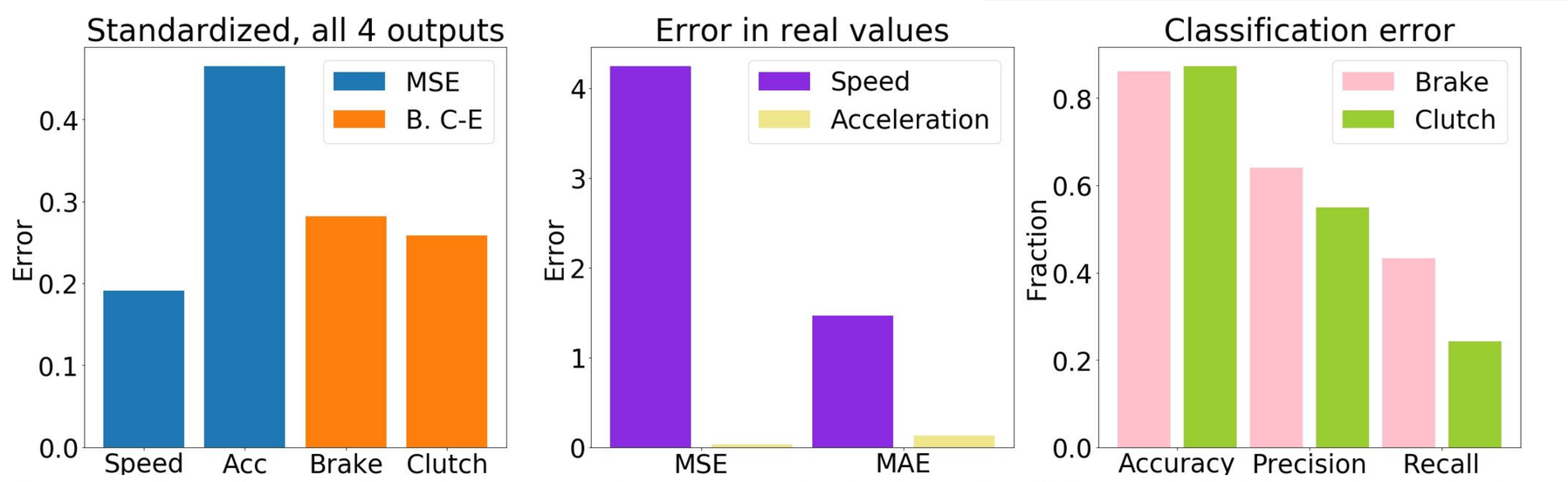

Figure 4: Final Random Forest model's performance on test data. Left: MSE and B. C-E for all four outputs in standardized values. Middle: MSE and mean absolute error (MAE) in real (non-standardized) values for speed and acceleration. Right: Accuracy, precision and recall for braking and clutching displayed as fractions of 1