Machine Learning: Group A Assignment

Submitted by: Maksim Gusev, Davin Mc Gowan, Senan Stanley, Rabin Sapkota

Submitted to: Donny Hurley

1 Introduction

This report is a group-based collation of research as part of the Machine Learning module taken at ATU Sligo. The objective was to present supervised analysis of the vehicle dataset (Peugeot and Opel-Corsa provided) based on their features and compare any two data classifier algorithms with that information.
The dataset was explored, features were recognised, data was processed, then formatted to remove inconsistency and two possible models identified (Linear Regression and Support Machine Learning). These models were chosen among several data algorithm techniques with the expectation of achieving best results. They were compared using metrics such as F1-score, accuracy and balanced accuracy. Metric results were analysed comparatively across models and chosen features.
The finalised version of the project was drafted by individual coding efforts that has been put together in a GitHub Repository [ML Group Code, 2022]. The project was facilitated and managed with weekly meetings, forum discussion, and improvisation through the Microsoft Team application.

2 State of the Art

An aspect of this assignment was to compare two multinomial classifier methods. The methods implemented in this study were support vector machine (SVM) and logistic regression.

2.1 Classification Methods

SVM is a supervised machine learning (ML) algorithm used normally for classification or regression. The SVM algorithm works by separating input data into classes using a maximum-margin hyperplane [Shmilovici, 2005]. Although SVMs were designed for binary classifications, they can be used in multinomial classification, by breaking down the multinomial categories into many binary classification problems [Wu et al., 2004].
Logistic Regression is another classification method and is an extension of Linear Regression. It is used to predict the probability of an input's class membership. Similarly to SVM, Logistic Regression was designed as a binary classifier, but can be extended for use in multinomial classification. Multinomial logistic regression uses maximum likelihood estimation to predict class membership [Starkweather et al., 2011].
Regularisation is a technique used to reduce a learning algorithm's generalization error but not it's training error [Goodfellow et al., 2017]. Regularisation penalises a complex model and can thus be used to reduce overfitting and improve a model's performance [Nagpal, 2022].

2.2 Metrics

To rigorously evaluate and compare the classification results of SVM and Logistic Regression probabilistic metrics must be employed. As part of this study, we implemented accuracy, balanced accuracy & F1 Score.
Accuracy is the ratio between the correct predictions and the total number of predictions. In other words, the true negative

& true positive predictions, divided by the true positive (TP), negative (TN), and false positive (FP), negative (FN) predictions [Metz, 1978]. Accuracy is a widely used metric in classification and prediction applications.

$$Accuracy = \frac{Correct\ Classifications}{Total\ Classifications} = \frac{TP + TN}{TP + FP + TN + FN}$$

Accuracy is not a perfect measurement; it returns misleadingly high results when minor classes are misclassified. As such, it works best when there an equal number of samples from each class [Haixiang et al., 2017].
A different accuracy metric is balanced accuracy. Balanced accuracy is a metric which is good at highlighting a class imbalance in the data, it is an average of recall and true negative rate [Brodeson et al., 2010].

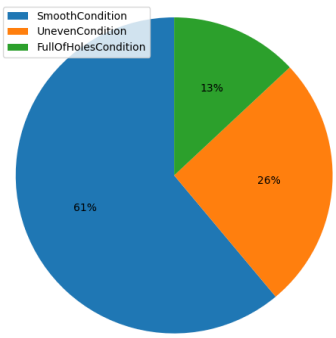
$$Balanced\ Accuracy = \frac{1}{2} \cdot \left(\frac{TP}{TP + FP} + \frac{TN}{TN + FN} \right)$$

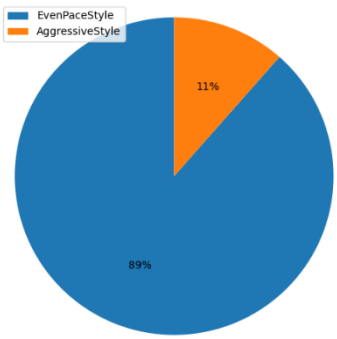
F_1 score is the harmonic mean of precision and recall [Sasaki, 2007]. This metric has a maximum value of 1, perfect recall & precision, and a minimum of 0, worst precision & recall. F_1 score is an appropriate metric to track when attempting to minimise both false positives and false negatives, as such it is widely used in ML applications.

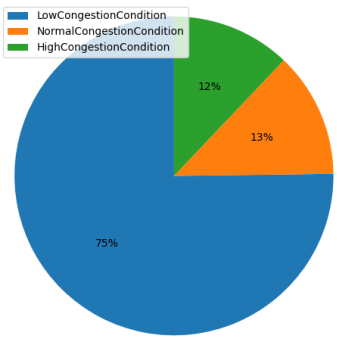
$$F_1\ Score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

3 Data Preparation

Data obtained from the real world can be vague and imbalanced. Data of this nature should be ‘sanitised’ before use, to ensure valid and sensical input. During this study, the raw data was cleared of commas to convert floating point numbers into decimal point numbers. Similarly, the non-numeric (NaNs) were replaced by the most frequent value in the column using scikit-learn “SimpleImputer” module [Simple Imputer]. Standardisation of the columns was accomplished by fitting and transforming a standard scalar to scale the variances, which streamlines the data for analysis.
The data target features; traffic, driving style and road surface, were assigned a numeric category by the “OrdinalEncoder” from the scikit-learn package [Ordinal Encoder].
Each of three features were independently, highly biased to a single element, as can be seen in the pie charts below. Bias of this sort occurs when a single element dominates the results, for example the “EvenPaceStyle” category accounts for 89% of all driving style data. This feature imbalance will be considered when processing the results.







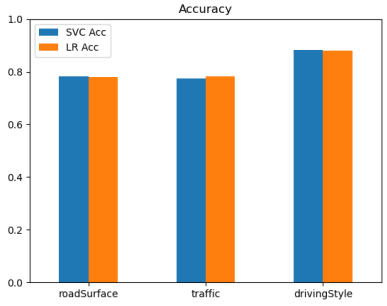
4 Implementations

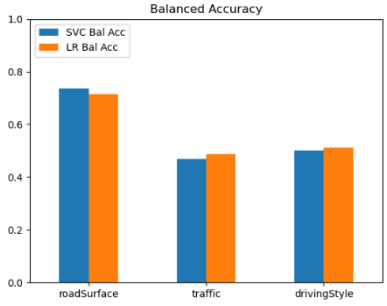
The implementation of the Logistic Regression and SVM was completed using the Machine Learning "scikit-learn" package within Python [Sci-Kit Learn Package, 2022]. The implementation is publicly available [ML Group Code, 2022].

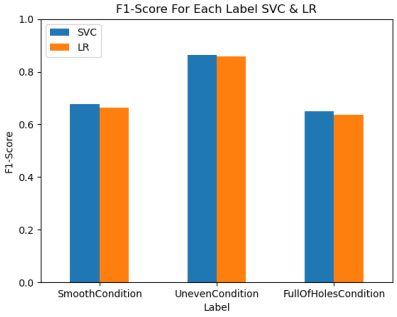
SVM implementation was relatively simple. A linear model kernel was chosen, which was then trained on the processed "training" dataset, iteratively, fitting the input data to the supplied output. When the model was trained, it was used to predict the output classes of the "test" dataset. These results were compared against the known test output classes.
The Logistic Regression implementation was also streamlined by using the "scikit-learn" package. Similarly to the SVM implementation, the model was trained on the "training" dataset and used to predict the "test" data, though the Logistic Regression had an extra step: Hyperparameter tuning. The multinomial Logistic Regression function is modified by the solver supplied (sag, saga, lbfgs, etc.) and most importantly by the regularisation parameters.
The regularisation possibilities in scikit-learn are: no regularisation, l1 regularisation & l2 regularisation; the difference between l1 & l2 regularisation is adding an absolute magnitude vs. adding a squared magnitude of the coefficient as a penalty term to the loss function [Nagpal, 2022].
The data splitting and modelling were implemented in a loop, so hyperparameters could be iteratively tested, and the best metrics achieved. This process was a crude example of iterative hyperparameter tuning, which allowed for optimisation of the regularisation solvers and methods.
This looped implementation also allowed for running models with each of the three possible target features (Road Surface, Driving Style & Traffic Conditions). Thus, it was possible to compare the same model to itself and see how it generalises.
Metrics were gathered by comparing the test dataset to the known target feature results. Accuracy, Balanced Accuracy, and F_1 Score were deployed in code using premade functions.

5 Results

	SVC Acc.	SVC Bal. Acc.	SVC F1 Mean	LR Acc.	LR Bal. Acc.	LR F1 Mean
Road Surface	0.782652	0.737606	0.728651	0.778446	0.715456	0.719583
Traffic	0.775441	0.468239	0.455450	0.782051	0.487309	0.499387
Driving Style	0.881611	0.500000	0.468540	0.880409	0.512502	0.496514







6 Discussion and Conclusions

This discussion will consider the possible reasons, analyses and conclusions for the applied models.
As can be seen in the *Results* graphs above, the imbalance of different labels in the target variables effected the classifiers accuracy and balanced accuracy. While the more unbalanced target variables showed higher accuracy, they also showed lower balanced accuracy indicating that the classifiers were doing a poor job of classifying less frequently encountered labels. F1 score more starkly reflects the poor precision and recall with the imbalanced data.
The SVM and Logistic Regression approaches are remarkably similar, and it's normal that the metrics gap is minimal. Low gap of metrics is a consequence of different approaches to loss function or that the SVM considers only points near the local boundary, while logistic regression considers global ones. In a situation with more balanced data, it could show bigger gaps in 2 methods metrics.
Improvements of models strongly depend on the quality and size of data. The results of the Linear Regression and Support Vector Machine models above, show a capacity for deployment in non-safety related functions. Further research into more complex models and hyperparameter optimisation could yield a model more suited to compensate for imbalances in training data.

References

- [Goodfellow et al., 2017] Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. MIT Press, ISBN: 9780262337434. http://www.deeplearningbook.org.
- [Metz, 1978], CE (October 1978). "Basic principles of ROC analysis" (PDF). Semin Nucl Med. 8 (4): 283–98. doi:10.1016/s0001-2998(78)80014-2. PMID 112681.
- [Haixiang et al., 2017] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73:220–239, 2017. https://doi.org/10.1016/j.eswa.2016.12.035
- [Brodeson et al., 2010] Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. (2010). The balanced accuracy and its posterior distribution. Proceedings of the 20th International Conference on Pattern Recognition, 3121-24.
- [Sasaki, 2007] Yutaka Sasaki. The truth of the F-measure. Teach Tutor Mater, 01 2007, https://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf
- [Shmilovici, 2005] Shmilovici, A. (2005). Support Vector Machines. In: Maimon, O., Rokach, L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. https://doi.org/10.1007/0-387-25465-X_12
- [Wu et al., 2004] Wu, Lin and Weng, "Probability estimates for multi-class classification by pairwise coupling", JMLR 5:975-1005, 2004.
- [Starkweather et al., 2011] Jon Starkweather and Amanda Kay Moske; Multinomial Logistic Regression, August 2011
- [Nagpal, 2022] Anuja Nagpal; L1 and L2 Regularization Methods, Explained, Jan. 05, 2022
- [Sci-Kit Learn Package, 2022] Sci-Kit Learn Package, scikit. Available at: https://scikit-learn.org/stable/ (Accessed: November 8, 2022).
- [Simple Imputer] *Sklearn.impute.SimpleImputer, scikit*. Available at: sklearn.impute.SimpleImputer (Accessed: November 10, 2022).
- [Ordinal Encoder] *Sklearn.preprocessing.OrdinalEncoder,scikit*. Available at: sklearn.preprocessing.OrdinalEncoder (Accessed: November 10, 2022).
- [ML Group Code, 2022] Stanley, S., McGovern, D. ML Group Code, GitHub. Available at: https://github.com/DavinMcGowan/Machine-Learning-Group-Assignment-/blob/main/ML_Assignment_Group_A.ipynb (Accessed: November 8, 2022).