Analysis of Road Fatalities in the UK

Comparison of Classification Algorithms

Saroop Kaur Samra
Halıcıoğlu Data Science Institute, UCSD
sksamra@ucsd.com

ABSTRACT

The purpose of this report is to analyze UK road fatalities in the year 2018 and build a prediction engine to estimate if the fatalities involved different age groups, people over or under the age of 38. The research implements a variety of different prediction models including neural networks, tree classifications, neighborhood and statistics-based classifiers to evaluate which provide the most accurate results as well as determine their execution performance.

CCS CONCEPTS

•Computing methodologies → Machine learning → Cross-validation

KEYWORDS

Road Fatalities, Classification, Hyper-Parameter Tuning, Bias Variance Tradeoff

1 INTRODUCTION

There were 1770 road deaths and 26,610 serious fatalities in the UK in 2018 alone. The overall number of casualties (death, serious or minor injury) was 165,100 that represented an increase of 3% year on year [1]. The popularity of the automobile has resulted in millions of deaths, from the 1950's to 2006 a total of over 300 thousand people were killed and 17.6 million were injured in accidents on British roads. Using the data from 2018, it is hoped that prediction models can be developed to help reduce fatalities by for example determining what additional road safety measure might be developed and using the model to estimate the impact of the as well as reducing the financial impact.

The data available from government agencies has been collected for long periods of time, for example, the Department of Transport maintains records for all road accidents that result in causalities which was first collected since 1926. However, more recently the processing and algorithms for sorting, processing and analyzing data has caught up significantly to allow researchers to analyze complex patterns in hundreds or thousands of gigabytes of data.

Previous work includes the Road Accidents in the UK (Analysis and Visualization) Anjul K. Tyagi [4]. This work also analyzed UK road accidents from 2005-2015 leveraging multiple variables (time, weather conditions, the age of driver etc.) with a Multiple Correspondence Analysis (MCA) approach. Although they did not develop a prediction engine, they did develop from their research the selection of the important features using hypothesis testing to predict the trend in accidents.

time series analysis. Other research includes a blog by Sheehan entitled - A Road Incident Model Analysis [5]. This research was a regression rather than a classification and used Auto Regressive Integrated Moving Average (ARIMA) to predict the number of road accidents in 2016.

1.1 Classification

The focus for this research is to survey 8 different classification algorithms to determine which show the best accuracy and have reasonable tradeoffs for performance. We will also develop our own "hybrid" model based on the three best individual classifiers. Our classification will predict if the driver is above or below a certain threshold age of 38 years old. One potential use for predicting the age of the driver involved in the accident is for auto insurance companies to determine the insurance rates. Traditionally young drivers have had higher premiums as they are considered high risk for accidents; however, Loughran et al highlight the dangers of older drivers in "What Risks Do Older Drivers Pose to Traffic Safety?" [3]. The rationale to determine our threshold age was based on the mean age of drivers in our population.

2 DATA

The data for this research as mentioned was gathered from the Department of Transportation. Our research leverages a large data set from 2018 which includes 226,409 vehicles, 122,635 accidents and 160,597 casualties recorded in these incidents [A.1]. However, this data did not include any economic or land usage data for the local authorities (the name for the UK's local government cities and towns responsible for local administration). Both the economic and land usage could be important parameters that determine the likelihood of accidents and their frequency based on the local authority where the incident took place. The land usage data was the most recent (2017) data from the Ministry of Housing, Communities & Local Government [A.2]. Finally, regional gross domestic product local authority data was from the Office of National Statistics from the year ending 2019 [A.3].

The data was in CSV and Microsoft Excel file formats and was approximately 40 MBs. The first step of processing was to merge the data sets together based on the local authority name. However, the accident data set recorded this using an integer code whereas the other data sets used the actual names. This required processing the data dictionary from the Department of Transportation prior to the merge step. The final merged data set had 76 parameters and 232,974 observations. Figure 1 shows the

parameters that had the highest number of missing values; for example, the age of the driver was missing in over 20 thousand incidents. As this parameter was used to derive our label (Age > 38), we removed all observations missing this data. Finally, new parameters were created, these include the month and hour which were extracted from the data and time parameters as these both would be important for our model (day and month were shown to have different characteristics of accidents). The 'Older_Driver' parameter was created based on our age cutoff of 38 and was used as our binary classification label.
The descriptive statistics was then generated after the process of cleaning, a small subset of this is shown in the appendix [A.4].

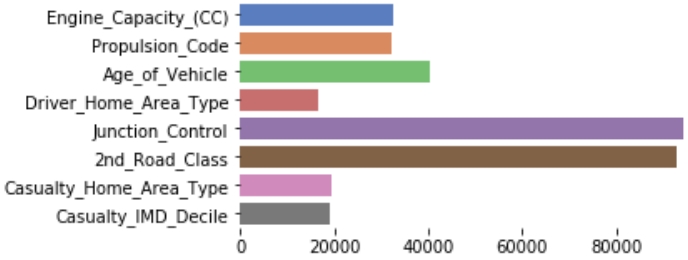

Figure 1: Most Missing Data

3 ANALYSIS

We first analyze the distribution of the ages of drivers which our label will be derived from. Figure 2 shows a distribution histogram of the male and female drivers with the cutoff age.

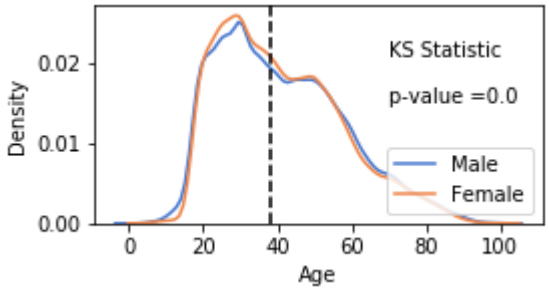

Figure 2: Age of Driver Distribution by Gender

Furthermore, we perform a KS Statistics on the male and female distributions and the resultant p-value is zero showing the distributions are approximately the same. However, in our EDA visualizations the gender differences for speed limit and the severity of accident.

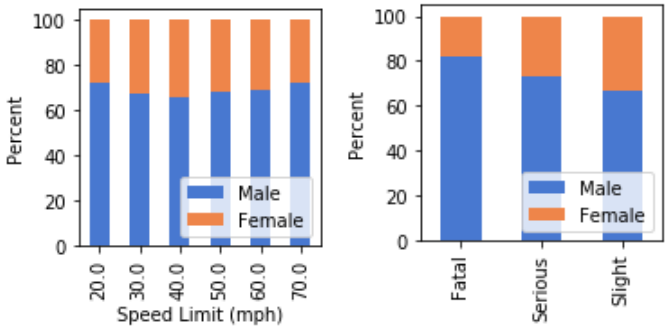

Figure 3: Gender Differences

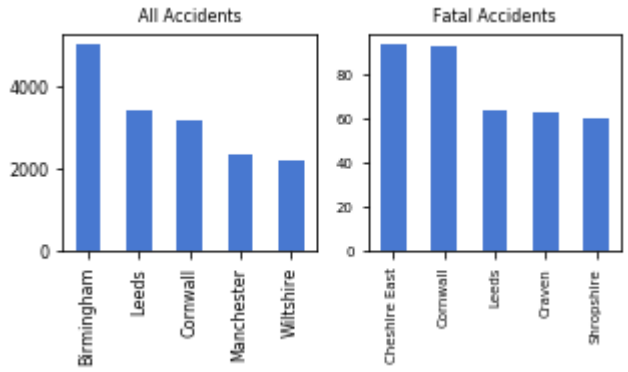

Figure 4: Top 5 Local Authorities

We further analyze the top local authorities with accidents, overall and fatal shown in Figure 4. The data highlights that the large metropolitan areas (Manchester and Birmingham) have more accidents, which is not surprising due to their population. However, Cornwall is a rural county but dominates in both overall and fatal accidents, which might be due to its lack of highways.

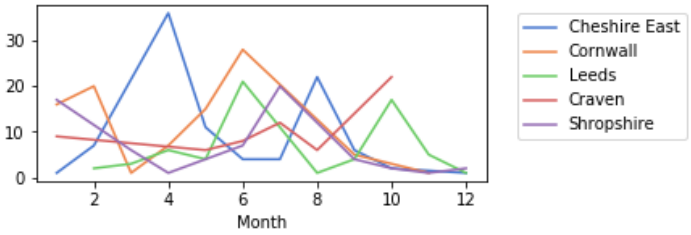

Figure 5: Monthly fatal accidents in Top 5 Regions

Figures 5 shows the same fatal accidents in the geographic regions over 12 months. These show a large degree of variance over the year. This does not show how this related to our target label.

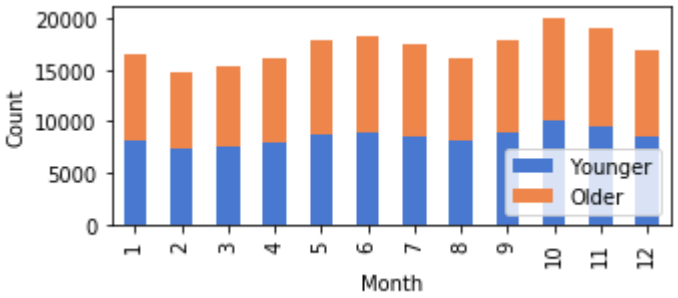

Figure 6: Classification of Age of Driver over Months

We also analyzed the distribution of the younger and older drivers over the course of the year in Figure 6. This again shows that the month derived feature shows differences in our label and is a suitable feature for the model. The month parameter was extracted from the year and was used in the model features.

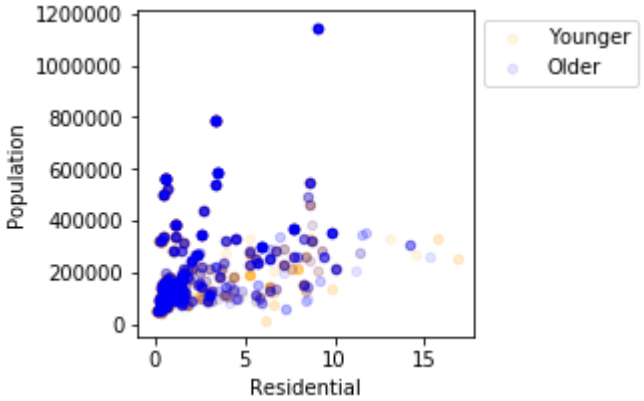

Figure 7: Population/Residential and Age

The scatterplot of financial data can be used to show some differences between the younger and older drivers, with more younger drivers in highly residential neighborhoods which presumably is towns and cities.

The data also includes weather and light conditions, in Figure 8 we show the non-normal light levels for fatal crashes which does show older drivers tended to be more affected in no lighting situations, for example 51.9% of the older drives are involved in no lighting accidents whereas younger drivers only are recorded at 49.2%.

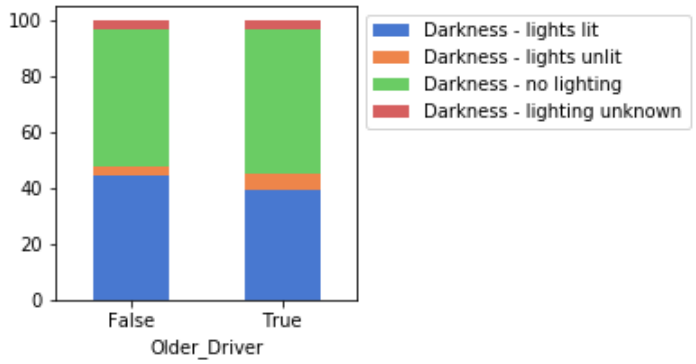

Figure 8: Light Levels by % and Age

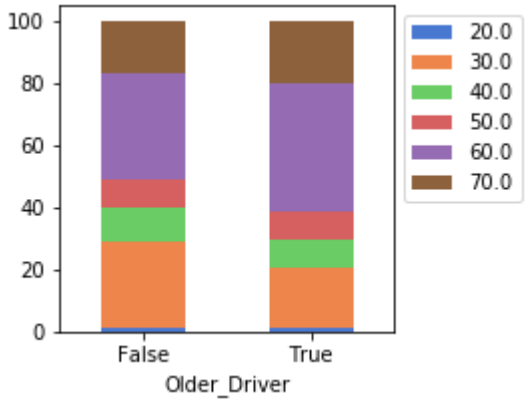

Figure 9: Speed Limit by % and Age

Each incident the speed limit of the road is also captured by the police and in Figure 9 we show the comparison against younger and older driver which shows older drivers are proportionally involved in more accidents on road over 60 miles per hour, for example 41.7% of the older drives are involved in road accidents at 60 mph whereas younger drivers only are recorded at 34.4%.

3.1 Feature Determination

Based on our exploratory analysis we concluded that 22 features could be used in our classification pipeline.

Data Set	Count
Vehicle	6
Casualty	2
Accident	8
Land Usage	2
Economic	3

Table 1: Data Set/Feature

Type	Count
Ordinal	1
One Hot	14
Min Max	4
Z Scale	1
Function	3

Table 2: Feature Type

The summary of the which data set the features originated from is shown in Table 1. For example, the vehicle data was used for light, weather and road conditions as well as the location and time and hour of the accident. The pipeline was developed using a variety of different processes on the features dominated by one hot encoding as the features were mostly categorical. However, the custom month was ordinal, and the various economic and land usage features used numerical manipulation (z-score, min max scaling).

4 Model Pipelines

The data was split into training, validation and testing based on a split of 60%, 20% and 20% respectively. The label we selected was a bool parameter for the age cutoff to determine young vs older drivers and was selected so that the there was an equal number of true and false cases in the data so our model would learn from balanced classes. For each model we measured the accuracy, precision recall and F1 scores. Additionally, we also recorded the prediction time. The first step was to predict based on fitting with the default parameters then afterwards we did a hyper-tuning parameter pass using 5 folds to determine the optimal settings and used to determine the results.

4.1 Decision Tree

The summary of the results for the decision tree is shown in Table 3. As expected, the training data set predictions received high scores, but dropped dramatically for the test data, F1 score of 67.75% for default settings. The parameters we did hyper-tuning were as follows: criterion, splitter and max depth. The resultant prediction on the optimized settings improved the F1 score to 70.28% and the confusion matrix is show in Figure 10.

Settings	Data	Accuracy %	Precision %	Recall %	F1 %	Prediction Time (s)
Default	Train	98.34	99.69	97.01	98.33	40.36
Default	Validate	67.85	68.98	66.35	67.64	17.78
Default	Test	68.01	68.57	66.96	67.75	16.18
Optimized	Train	72.02	74.65	67.25	70.75	38.96
Optimized	Validate	71.70	74.77	66.60	70.45	16.48
Optimized	Test	71.66	74.19	66.77	70.28	17.19

Table 3: *Decision* Tree Results

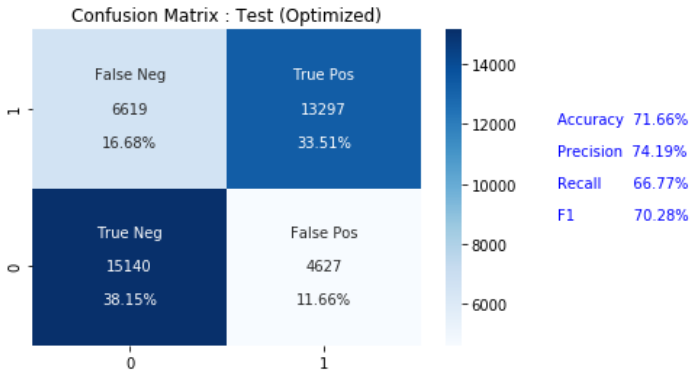

Figure 10: Decision Tree Confusion Matrix

4.2 Random Forest

The parameters we did hyper-tuning were as follows: criterion, number estimators and max depth. The resultant prediction on the optimized settings improved the F1 score to 74.64% as shown in Figure 11.

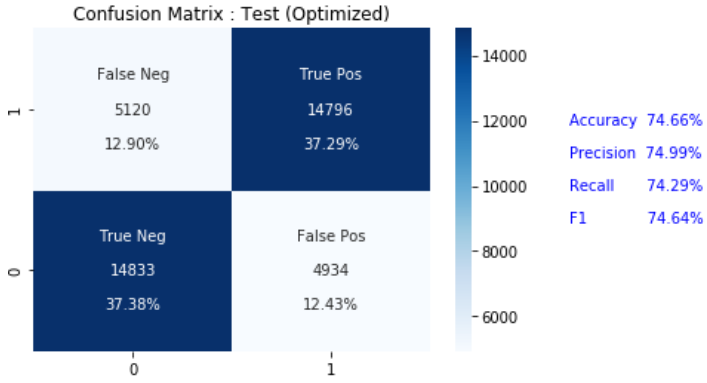

Figure 11: Random Forest Confusion Matrix

4.3 SVC

The parameters we did hyper-tuning were as follows: C regularization parameter. The resultant prediction on the optimized settings improved the F1 score to 55.19%. The SVC model proved to be the worst performing model in terms of accuracy as shown in Figure 12.

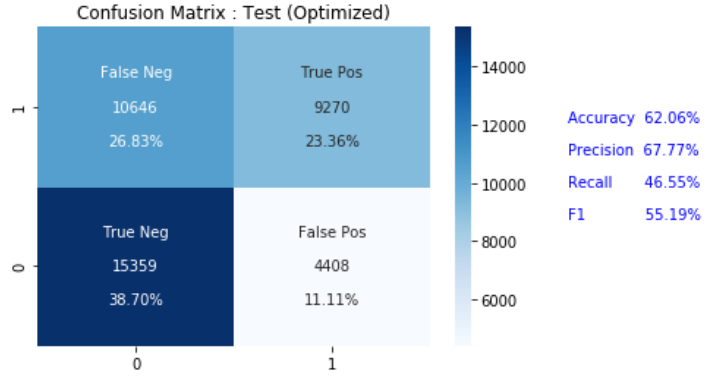

Figure 12: SVC Confusion Matrix

4.4 Gradient Boost

The parameters we did hyper-tuning were as follows: loss and number of estimators. The resultant prediction on the optimized settings improved the F1 score to 72.64% as shown in Figure 13.

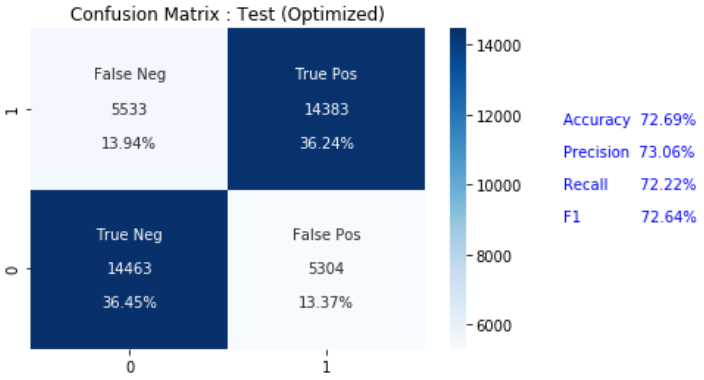

Figure 13: Gradient Boost Confusion Matrix

4.5 Bagging

The parameters we did hyper-tuning were as follows: max features and max samples. The resultant prediction on the optimized settings improved the F1 score to 71.14% as shown in Figure 14.

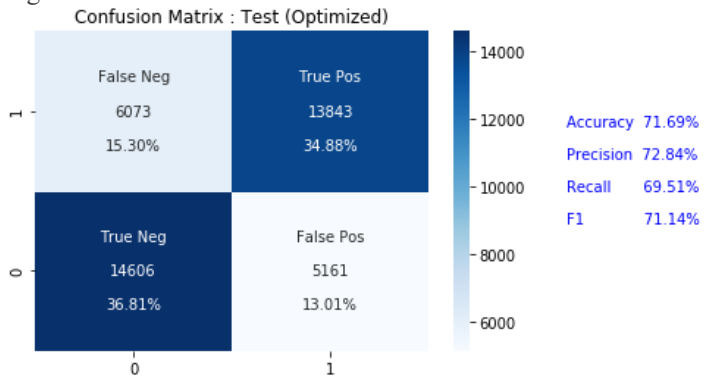

Figure 14: SVC Confusion Matrix

4.6 Naïve Bayes (Bernoulli)

The parameters we did hyper-tuning were as follows: alpha and binarize. The resultant prediction on the optimized settings improved the F1 score to 60.22% as shown in Figure 15. This was also a poorly performing model.

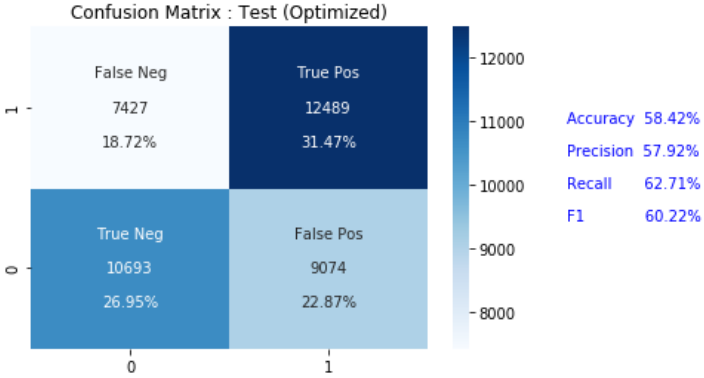

Figure 14: Naïve Bayes Confusion Matrix

4.7 Nearest Centroid

The parameters we did hyper-tuning were as follows: metric. The resultant prediction on the optimized settings improved the F1 score to 70.04% as shown in Figure 15.

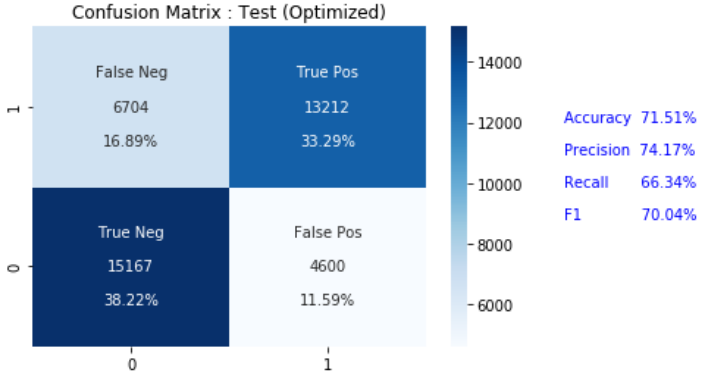

Figure 15: Nearest Centroid Confusion Matrix

4.8 Multi-Layer Perceptron

The parameters we did hyper-tuning were as follows: activation and solder. The resultant prediction on the optimized settings improved the F1 score to 71.83% as shown in Figure 16.

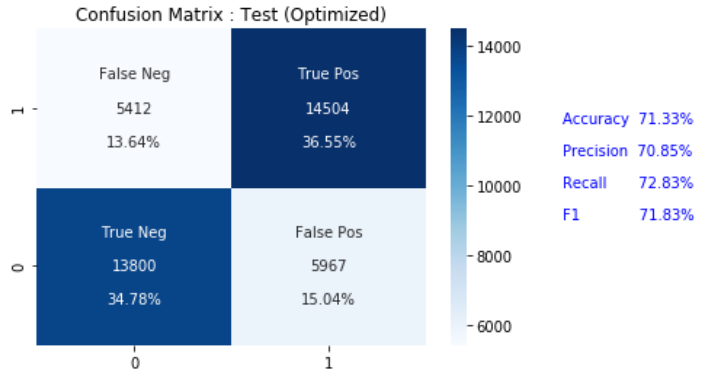

Figure 16: Multi-Layer Perceptron Confusion Matrix

4.9 Model Score Summary

The model score summary is shown in Table 4. The best performing model in terms of F1 score is random forest followed closely by gradient boost and multi-layer perceptron. The worst performing model is SVC and naïve bayes also performed poorly.

Table 4: Model Score Summary

Model	Accuracy %	Precision %	Recall %	F1 %
Decision Tree	71.66	74.19	66.77	70.28
Random Forest	74.66	74.99	74.29	74.64
SVC	62.06	67.77	46.55	55.19
Gradient Boosting	72.69	73.06	72.22	72.64
Bagging	71.69	72.84	69.51	71.14
Naive Bayes (Bernoulli)	58.42	57.92	62.71	60.22
Centroid	71.51	74.17	66.34	70.04
Multi Layer Perceptron	71.33	70.85	72.83	71.83

4.10 Hybrid Model

Our final experiment we combined three models together and used their combined results to see if we could generate a better overall score. The three models we selected were: random forest, bagging and gradient boost. These all had decent scores when predicting individually. Our initial approach was to use the idea of best of 3, i.e. if 2 predicted True then we would select True otherwise if 2 predicted False then we would select False. However, the results of this were unsatisfactory and resulted in worse overall performance. After experimenting we found that the best approach was to use bias it more to True, i.e. "Or", if any models were True then we would select True and only select False if all three were False. The resultant prediction on the optimized settings improved the F1 score to 75.67% as shown in Figure 17. This is a slight improvement over selecting one model.

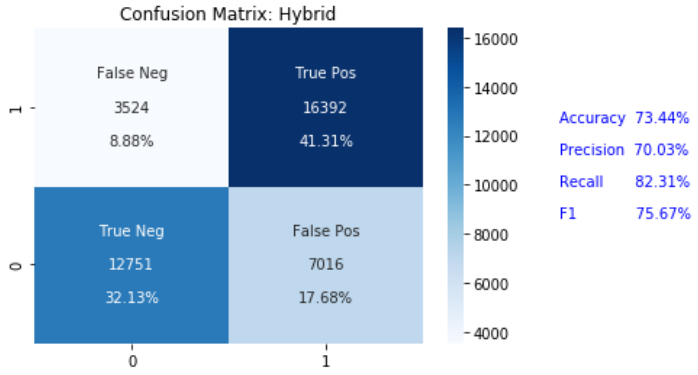

Figure 17: Hybrid Model Confusion Matrix

4.11 Development Summary

The development was done in Jupyter notebook and then an application demo, Figure 18, was developed with a front-end using a web browser and a backend python application that implemented a WebSocket to communicate with the browser.

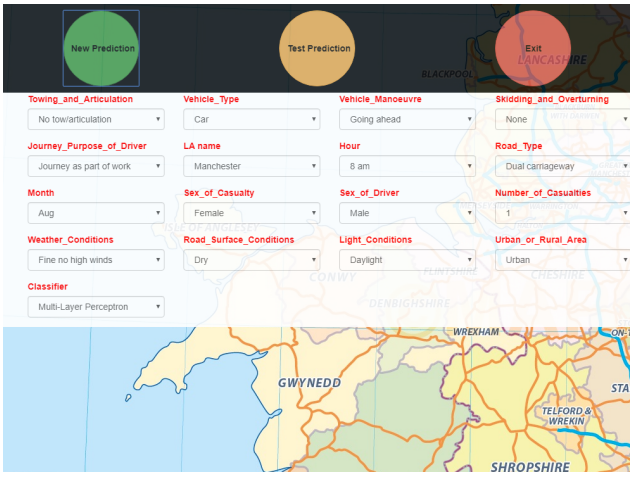

Figure 18: Demo User Interface

The demo allows the user to enter in parameters a make new prediction including selecting any of the models. In addition, the demo also predicted the entire test data set and generated the F1 score. Finally, the demo loaded the models rather than re-training from scratch, the models were initially saved from Jupyter notebook.

Model	Prediction Time (s)	Train Time (s)
Decision Tree	17.19	1.258532
Random Forest	32.27	29.15604
SVC	24.18	0.861355
Gradient Boosting	32.37	53.596719
Bagging	38.46	113.82743
Naive Bayes (Bernoulli)	13.69	0.497749
Centroid	14.78	0.488499
Multi Layer Perceptron	17.78	116.46408

Table 5: Model Speed Summary

The initial experiments were run on a Mac laptop; however, the performance was too slow when doing the training as the data set was large. The final experiments were performed on a desktop PC with 32 GB of system memory and an Intel Core i7-5960X CPU @ 3Ghz. The model speed summary is shown in Table 5. The fastest model for training was nearest centroid and naïve bayes but the slowest training performance were nearly 200 times slower: multi-layer perceptron and bagging classifiers. However, the training is a one-time cost, so we also reviewed the speed of prediction which shows that the variance of difference is substantially reduced, for example, the fastest versus slowest prediction times are only 2.3 times slower, implying prediction time is similar across these models but the training times are vastly different.

The file size of the model is shown in Table 6. The largest model was the random forest and bagging. Random forest was significantly larger than all the others at 30MBs. The other models were relatively small and smaller than 1 MB.

Model	File Size
Decision Tree	10 KB
Random Forest	30 MB
SVC	8 KB
Gradient Boosting	342 KB
Bagging	14 MB
Naïve Bayes (Bernoulli)	14 KB
Centroid	8 KB
Multi-Layer Perceptron	335 KB

Table 6: Model Space Summary

4.12 Application

5 Conclusion

We have optimized 8 different models for the to analyze UK road fatalities in the year 2018 and build a prediction engine to estimate if the fatalities involved different age groups, people over or under the age of 38. The models selected were broadly using different prediction algorithms including neural networks, tree classifications, neighborhood and statistics-based classifiers. We found that naïve bayes and SVC classifiers performed poorly and random forest, gradient boosting, bagging and multi-level perceptron produced the best results. Finally, we developed our own hybrid model that was based on the "max" of the three best models that produced a slight improvement using any single model.

ACKNOWLEDGMENTS

This report was written as the final project report for DSC190B – Introduction to Data Mining course. I wish to thank Professor Shang for his teaching in Spring 2020.

REFERENCES

[1] Reported road causalities in Great Britain 2018, Department for Transport, UK, Nov 2018, https://assets.publishing.service.gov.uk,

[2] Tomorrow's Roads: Safer for Everyone (Report). Department for Transport. 1999

[3] Loughran, David S, Seth A. Seabury, and Laura Zakaras, What Risks Do Older Drivers Pose to Traffic Safety?. Santa Monica, CA: RAND Corporation, 2007. https://www.rand.org/pubs/research_briefs/RB9272.html.

[4] Anjul K. Tyagi, et al, Road Accidents in the UK (Analysis and Visualization) Department of Computer Science, Stony Brook University, New York. https://www3.cs.stonybrook.edu/~anshul/vis18_poster.pdf

[5] Blog: A Road Incident Model Analysis , David Sheehan https://dashee87.github.io/data%20science/general/A-Road-Incident-Model-Analysis/

A APPENDICES

In the appendix section, three levels of Appendix headings are available.

A.1 Accident, Vehicle and Causality Data Sets

https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

A.2 Land Usage Data Set

https://www.gov.uk/government/statistics/land-use-in-england-2017

A.3 Regional gross domestic product local authorities Data Set

https://www.ons.gov.uk/economy/grossdomesticproductgdp/datasets/regionalgrossdomesticproductlocalauthorities

A.4 Key Descriptive Statistics

	Sex_of_Driver	Vehicle_Type	Light_Conditions	Hour	Population	GDP
count	232974.000000	232672.000000	232974.000000	232952.000000	2.329740e+05	2.329740e+05
mean	1.444848	10.333560	2.005636	13.635281	2.553424e+05	4.695051e+04
std	0.634352	10.898588	1.719359	5.020517	2.140972e+05	3.631176e+05
min	1.000000	1.000000	1.000000	0.000000	8.706000e+03	1.401000e+04
25%	1.000000	9.000000	1.000000	10.000000	1.215660e+05	2.196300e+04
50%	1.000000	9.000000	1.000000	14.000000	1.773520e+05	2.741600e+04
75%	2.000000	9.000000	4.000000	17.000000	3.172560e+05	3.503500e+04
max	3.000000	98.000000	7.000000	23.000000	1.141374e+06	8.235208e+06

Descriptive Statistics of key parameters