Overcoming Imbalance to Predict Accident Severity Using Neural Networks and Random Forests

Joseph Lipa, Rui Ma, Andres Cambronero
{joelipa, mrui, cambrone}@umich.edu

April 15, 2018

Member Contributions

- Joseph Lipa: Neural networks section, final compilation and editing of report
- Rui Ma: Found dataset, tree-based classification methods, brainstorming
- Andres Cambronero: Data cleaning, data exploration, and writing of sections other than tree-based classification methods, neural networks, and conclusion

Introduction

For years, civil engineers, urban planners, and government officials have worked to develop accurate systems that can detect automobile accidents and predict the severity of collisions. Ideally, such systems would identify unsafe areas in highways and contribute to more efficient ambulance and police dispatching. Given that the estimated cost of accidents in 2016 in the UK alone rose to 20 billion pounds, researchers and governments have dedicated considerable efforts to achieve this goal.

With the advent of modern classification methods, researchers have moved closer to developing such systems. In particular, neural networks and random forests have allowed researchers to predict the severity of accidents with a higher degree of accuracy than previously possible. The results of studies implementing modern classification methods represent a remarkable improvement on the conclusions of previous studies.

Although the results are encouraging, accident prevention studies often do not acknowledge that classification methods do not perform optimally on imbalanced data. Common classification methods minimize the overall error rate rather than attempt to correctly predict the "rare" class label. In accident prevention, this behavior is problematic because correctly predicting the rare case of fatal accidents is more important than predicting the much more numerous case of minor incidents. By failing to address this feature of classification methods, researchers inadvertently ignore the difference in economic and societal costs of accident severity.

This study attempts to correct for the misuse of classification methods in the accident prevention literature. By employing randomized undersampling of the majority class, implementing cost-sensitive learning approaches, and using recall of serious accidents as the metric of evaluation, we found that researchers in traffic prevention can achieve better prediction of severe accidents. Specifically, the study relies on UK accident reports from 2014 to examine the ability of neural networks and random forest methods to identify the severity of accidents when presented with extremely imbalanced data as compared with balanced data and weighted class labels. These methods were chosen because they are often used in the accident prevention literature.

Literature Review

As mentioned previously, neural networks are common in studies of accident prediction, but the effect of class imbalance on performance is often overlooked. For example, Mehmet et al. (2011) used a neural network to predict the severity of accidents in Iran. While the neural network outperformed the generic algorithm to which they compared it, the authors based their results on a dataset that contained 14 percent of observations indicating fatal injury and 47.6 percent of observations indicating no injury. In their discussion, the authors fail to address the effect of class imbalance on their network's classification performance.

Similarly, Miao et al. (2005) employed a neural network, decision trees, and support vector machines to predict the severity of accidents in the United States. Their reported overall classification accuracy for the neural network ranged from 57.68 to 63.36 percent, which trailed behind the decision tree by only a few percentage points. The authors based their findings on data with 50 percent of observations indicating no injury and only 1 percent indicating fatal injury, while the rest of observations were distributed among other classes. Once again, the authors do not address the effect of class imbalance on their results.

Researchers using other classification methods also fail to discuss the effect of class imbalance on their results. Iranitalab and Khattak (2017) compare the overall performance of multinomial logit, nearest neighbor classifier, support vector machines, and random forests. The data used in the study contained a response variable with 64 percent of the data in the lowest severity category and only 4 percent in the highest severity. As with previous studies mentioned, Iranitalab and Khattak fail to address the effect of imbalance on their results.

The literature on learning methods from imbalanced data suggests two general measures to improve the performance of classifiers: sampling and cost-sensitive learning. Breiman et al. (2004) investigate the effect of undersampling from the majority class, oversampling from the minority class, and using random forest on imbalanced data sets to increase the cost of misclassifying minority class observations. Using performance metrics appropriate for imbalanced datasets rather than overall prediction accuracy, their results suggest that both sampling approaches lead to a larger improvement in performance than weighted random forests.

Mazuroski et al. (2008) find very similar results when employing a neural network to classify observations in medical settings. As with our accident setting, the authors view false negatives as carrying a higher cost than false positives. As a result, employing a neural network on the imbalanced data and assessing its performance with overall prediction accuracy is inappropriate. Using Area Under the Curve (AUC) as a performance metric, the study demonstrates that employing some sampling method to balance the classes improves the performance of the classifier.

Methodology

This section briefly describes the methods used in this paper to improve the performance of neural networks and random forests.

Random Undersampling: Using this method, observations with the class label having the largest proportion in the original data are randomly selected until they represent a specified percent of the data. According to Breiman et al. (2004), undersampling changes the prior probabilities imposed on the majority and minority class and as a result allows for greater degree of separability between classes. In our study, the majority classes were undersampled until each represented one third of the training data.

Cost-Sensitive Learning: This method imposes a heavier penalty on missclassifying observations from the minority class compared to the majority class. Zadrozny et al. (2003) explain that imposing different costs on observations of different classes allows for better overall prediction than if the classifier was applied on the data without weights.

In this paper, cost-sensitive learning was applied only to random forests; we did not find a suitable R implementation allowing the incorporation of weights into neural networks.

Recall: Rather that use overall error rate or accuracy as performance metric, this paper evaluates performance using recall (sometimes called sensitivity) for fatal and serious accidents. For each class, recall is defined as $\frac{TP}{TP+FN}$, where TP is number of true positive predictions and FN is the number of false negative predictions.

Data Processing

The data for this study was downloaded from Kaggle. The data contained accident reports from the UK for the years 2005 to 2014. Because the project intends to develop a classifier that accurately detects serious car accidents given today's traffic patterns, only observations from the most recent year were used. After removing all NA data cells, 2400 complete cases were selected. These observations preserved the original proportions in levels of "accident severity." The "accident severity" has 0.07 percent fatal, 12 percent serious and 87 percent slight. Variables that contained all NAs were removed, and categorical variables that had many levels were also dismissed. The dataset also retained variables that previous studies on the topic have found to be significant.

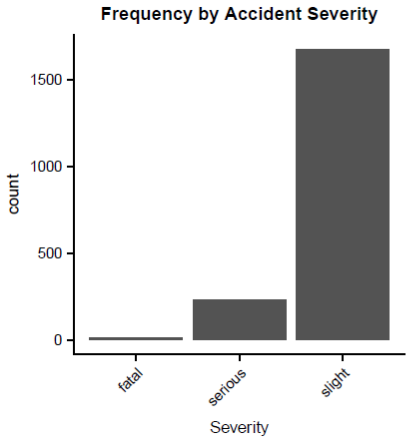
The final dataset contained 2400 observations and 30 columns. The predictors included variables that previous studies in the field have found significant, such as sex of driver, urban/rural, weather conditions, and age of driver. This dataset was split using an 8:2 ratio into a training set and a test set containing 1919 and 481 observations, respectively. The test set retained the original class proportions present in the pre-split dataset. To create the balanced training set, random observations for each class were selected until each class label represented one third of the data.

Data Exploration

This section presents general features of the data used in this study and explores characteristics of predictors that previous studies have identified as important variables for predicting accident severity.

Figure 1 below shows the class imbalance existing in the training set. The training data contained only 14 (0.7 percent) observations of class "fatal," 232 (12 percent) observations of class "serious," and 1673 (87 percent) observations of class "slight." This distribution of accidents is similar to that of data used in the papers mentioned in the literature review.

Figure 1: Frequencies and Proportions by Accident Severity



Accident Severity	Count	Prop.
fatal	14	0.00730
serious	232	0.12090
slight	1673	0.87181

Table 1 and Figure 2 below present visual and numeric relationships between the continuous predictors in the data. Table 1 shows that several observations have unusually low values for age of driver. Since the data source codebook indicates that these values are acceptable, the observations

were included in the analysis. The density plots in Figure 2 show that all three variables are skewed right. The correlation between the variables is almost non-existent, except for the relation between age of driver and age of the casualty, which is 0.432.

Table 1: Summary of Numeric Predictors

Predictor	Min	Mean	Median	Max
Age of Driver	3	28.73476	27	81
Age of Casualty	2	38.19958	35	98
Age of Vehicle	1	7.99531	8	47

Figure 2: Distributions and Correlations of Numeric Variables

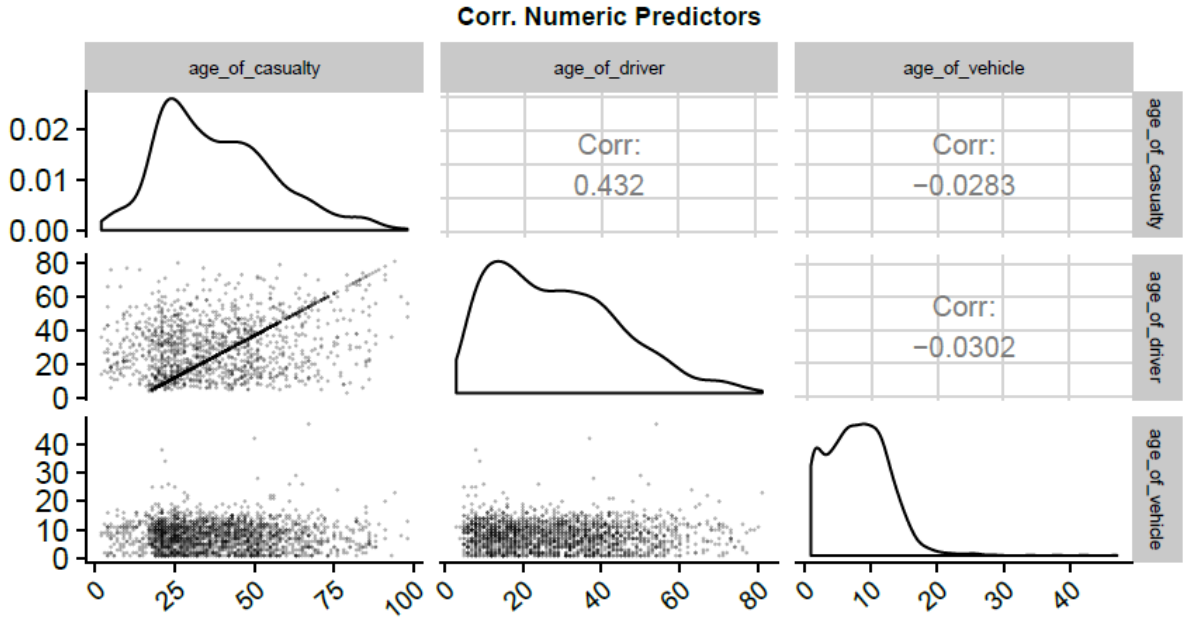
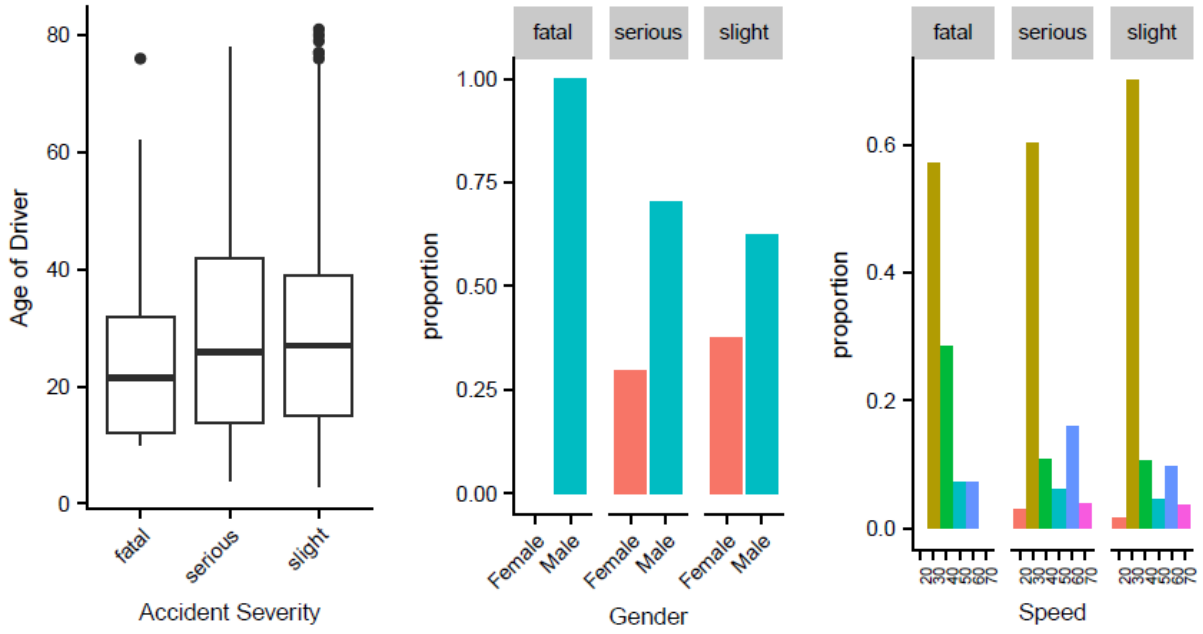


Figure 3 below shows the distribution of observations with class label fatal, serious, and slight for predictors that previous studies identified as significant predictors of accident severity. As previous research has concluded, the leftmost plot shows that the mean age of observations with class fatal is lower than that of the other two classes. While this result might suggest young drivers are more likely to be involved in fatal accidents, the difference between groups is marginal at best. In line with previous studies, a larger proportion of men are involved in accidents of all classes compared to women, as shown in the middle graph. In contrast to results of other studies, the graph on the right does not suggest that accidents of any class are more likely to occur at higher speeds. These graphs also indicate that within each class the number of observations for each level of several predictors was also imbalanced.

Figure 3: Significant Predictors in Previous Studies



Finally, Tables 2, 3, and 4 below show three variables that are commonly associated with traffic accidents and the corresponding level with the highest proportion in each class. The highest proportion of fatal accidents occur on Sundays and Tuesdays; for serious accidents, on Saturdays; and for slight accidents, on Tuesdays. This result is consistent with our expectations. Perhaps surprisingly, for all three levels of accident severity the most common weather condition "Fine no winds," and the most common road condition was "Dry." The count by class for each variable reinforces the heavy extent to which the data is imbalanced between fatal, serious, and slight accidents.

Table 2: Level of 'Day' with Highest Proportion in Each Class

Acc. Severity	Day	Count	Proportion
fatal	Sunday	4	0.2857143
fatal	Tuesday	4	0.2857143
serious	Saturday	45	0.1939655
slight	Tuesday	265	0.1583981

Table 3: Level of 'Weather' with Highest Proportion in Each Class

Acc. Severity	Weather	Count	Proportion
fatal	Fine no high winds	13	0.9285714
serious	Fine no high winds	186	0.8017241
slight	Fine no high winds	1375	0.8218769

Table 4: Level of 'Road Condition' with Highest Proportion in Each Class

Acc. Severity	Road Cond.	Count	Proportion
fatal	Dry	12	0.8571429
serious	Dry	158	0.6810345
slight	Dry	1202	0.7184698

Neural Network Classification Methods

In this section, we will consider neural networks, perhaps the most common method in the literature for classifying accident severity. Unlike previous studies, we do not wish to neglect the effect of class imbalance on performance. Hence, we will build neural network classifiers based on both the original training data, which is heavily imbalanced, and on training data that has been artificially balanced by undersampling from the majority class. We will test the models on the original, heavily imbalanced test data and examine recall of fatal and severe accidents in each case.

When building a neural network, there are two primary tuning variables to specify: the number of hidden layers and the number of hidden nodes per layer. We chose a single hidden layer for practical reasons, as the number of variables in our dataset made multi-layer networks computationally infeasible. Indeed, we were unable to fit even a two-layer network on the balanced training data without radically altering the algorithm's convergence criteria. The number of hidden nodes per layer also affects model complexity. More hidden nodes makes the model more complex, but also increases the chance of overfitting the data.

Using the imbalanced training set, we fit single-level neural networks with 10 and 20 hidden nodes. The classification results are given below in Table 5 and Table 6, respectively. Throughout this paper, we present classification results in the form of a confusion matrix, in which the row labels represent the true class and the column labels represent the predicted class.

Table 5: Neural Network Trained on Imbalanced Data, 10 Hidden Nodes

	1	2	3
1	0	1	3
2	1	14	43
3	12	60	347

Table 6: Neural Network Trained on Imbalanced Data, 20 Hidden Nodes

	1	2	3
1	0	2	3
2	0	11	45
3	3	47	317

We can calculate overall prediction error rates from these tables by dividing the sum of the off-diagonal numbers by the sum of all the numbers. Doing so gives modest error rates of 0.1384 and 0.2172, respectively. However, the effect of the class imbalance is glaring. In both cases, the recall for fatal accidents (class 1) is zero; that is, no fatal accidents were predicted correctly. The recall for severe

accidents (class 2) is $\frac{14}{1+14+43} = 0.2414$ for the model with 10 hidden nodes and $\frac{11}{11+47} = 0.1897$ for the model with 20 hidden nodes. In contrast, the recall for slight accidents (class 3) is $\frac{347}{419} = 0.8282$ and $\frac{371}{419} = 0.8854$ for the respective models. It is clear that, due to the disproportionately high number of slight accidents in the training data, the models have great difficulty predicting fatal and severe accidents. However, these are precisely the two accident types we are interested in predicting. We therefore have little confidence in these neural network models trained on such highly imbalanced data. Hence, we will turn our attention to models trained on balanced data.

We will obtain this balanced data by undersampling from the largest classes so that each accident severity class—slight, severe, and fatal—composes a third of the observations in the training data. We will use this data to again fit single-layer neural networks with 10 and 20 hidden nodes. The test results are given below in Table 7 and Table 8, respectively.

Table 7: Neural Network Trained on Balanced Data, 10 Hidden Nodes

	1	2	3
1	4	0	0
2	13	31	14
3	64	70	285

Table 8: Neural Network Trained on Balanced Data, 20 Hidden Nodes

	1	2	3
1	3	1	0
2	7	48	3
3	66	117	236

While the overall prediction error rates calculated from these tables have risen to 0.2363 and 0.3150, respectively, the improvement in recall is dramatic. The model with 10 hidden nodes correctly predicts all four fatal accidents, for a recall of 1. Its recall for severe accidents is $\frac{31}{58} = 0.5345$, a substantial improvement over the corresponding model trained on imbalanced data. The model with 20 hidden nodes has a recall of $\frac{3}{4} = 0.75$ for fatal accidents and $\frac{48}{58} = 0.8276$ for severe accidents, also dramatically better than its counterpart trained on imbalanced data.

We have seen that artificially balancing the class distribution in the training data can greatly improve the ability of neural networks to correctly predict rare classes. This is especially helpful when the rare classes are of the most interest, such as in our accident data. Nevertheless, this approach also has its disadvantages. For example, undersampling from the majority class may make it difficult to retain a healthy sample size, especially if the data set is small to begin with. Another drawback appears to be the increased computation time for neural networks trained on balanced data. Our model fit with 20 hidden nodes on the imbalanced data converged in just over 13 seconds; with the balanced data, it took almost 13 minutes! Hence, while neural networks are popular in the accident severity literature, it is worthwhile to also explore other methods, as we do in the next section.

Tree-Based Classification Methods

In this section, we will explore the difference in the classification results using tree-based classifiers. We will train them on both the original imbalanced data set and the balanced data set that we undersampled from the majority class.

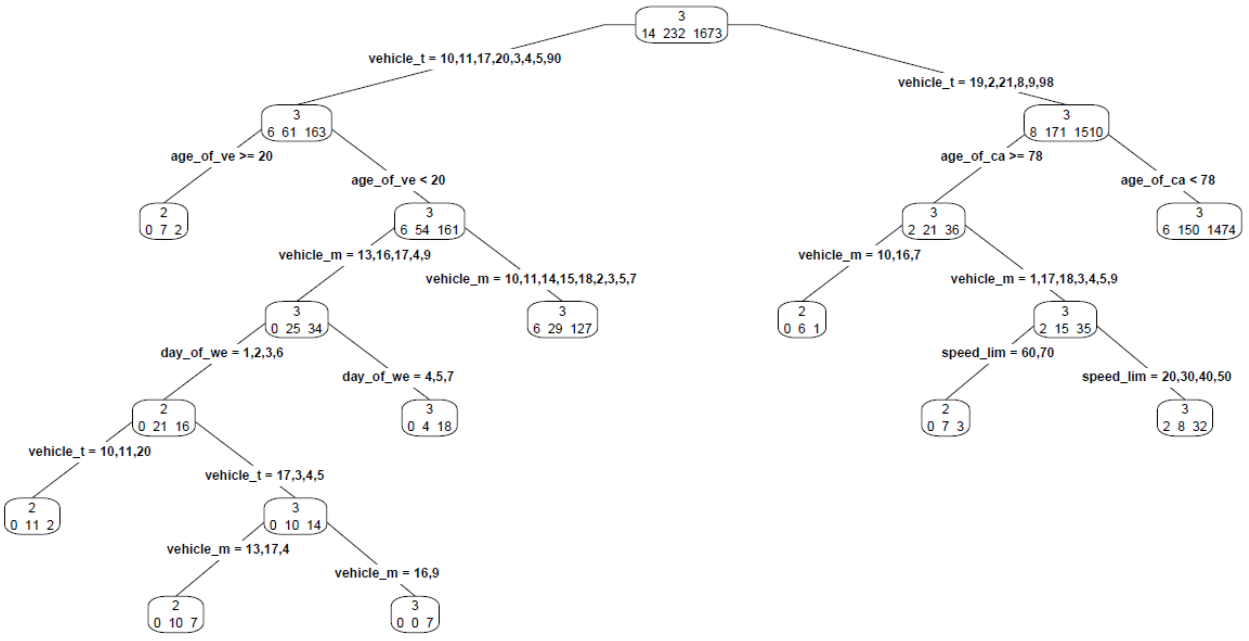
For classification problems, tree-based approaches are generally known as decision tree methods. These methods involve segmenting the predictor space into a number of regions based on simple decision rules (e.g., if the categorical predictor $a = 1$, etc.). Therefore, a big advantage of the decision tree classifier is its high interpretability and its ability to handle both quantitative and qualitative variables. However, typically a single decision tree classifier is not competitive with the best supervised learning approaches due to its tendency to bias towards the training set (i.e., to overfit). Thus, we will also focus on the random forest approach, which involves constructing many decision trees under certain constraints and combining them to yield a single consensus prediction result.

After cleaning up our data, we will first train a single decision tree on the original imbalanced data set, and then use this classifier to predict the accident severity on the test data set. The results are shown below in Table 9, and Figure 4 further below gives a visual representation of the tree.

Table 9: Single Tree Trained on Imbalanced Data

	1	2	3
1	0	0	4
2	0	3	55
3	0	5	414

Figure 4: Single Decision Tree on Imbalanced Data



The first split is based on the "vehicle type" variable, the second split on the left is based on the "age of vehicle" variable, and so on. From Table 9, we can calculate that the overall prediction error rate is 0.1331. While this error is quite low, the recall for "fatal" (class 1) and "serious" (class 2) are all close to zero. In other words, the accident severity of almost all of our test observations was predicted to be "slight" (class 3). Since over 84% of the test set observations are "slight" on the accident severity scale, we would obtain a prediction error of 0.16 if we were to blindly predict every accident to be slight. This seemingly low overall error rate is likely the result of our highly imbalanced data set.

Now, we will train a random forest (RF) classifier and see if there will be any improvement over the single decision tree approach. Random forest is a technique that constructs many individual decision trees in order to reduce variance, similar to bagging. However, while building the decision trees, the algorithm selects a random sample of m predictors as split candidates from the p total predictors. This is to reduce correlation between the trees. Hence, we first need to tune the random forest to determine the optimal m (*mtry* is the name of the argument in R).

From the plot below, we see that when $mtry = 5$, the out-of-bag (OOB) error is the lowest. We will then use this value to train our random forest classifier on the original imbalanced data set:

Figure 5: Out-of-Bag Error vs. Number of Predictors Considered at Each Split

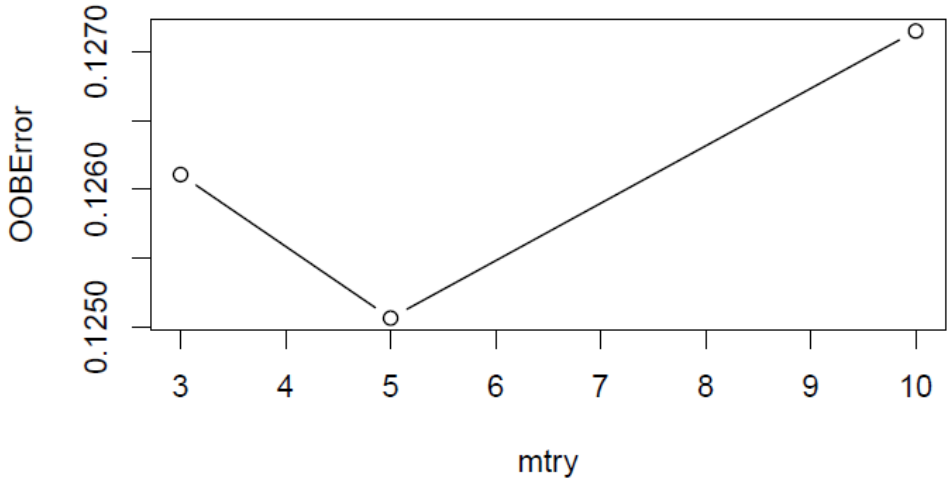


Table 10 below shows six of the most important predictors (sorted by importance) included in the training data set. This table was acquired using the *importance* function included in the `randomForest` package in R. From Table 11 below, we see that the overall error rate is now 0.1227, which is lower than the result using the single decision tree classifier. However, the same issue persists: none of the "fatal" accidents were predicted correctly due to the class imbalance.

Table 10: Variables by Importance

var	mean_dec_gini
age_of_casualty	55.80927
age_of_driver	46.15217
vehicle_manoeuvre	39.82866
age_of_vehicle	38.32333
day_of_week	37.33854
vehicle_type	30.36478

Table 11: Random Forest Trained on Imbalanced Data

	1	2	3
1	0	0	4
2	0	3	55
3	0	0	419

In order to address this problem, we will consider three approaches. The first option is to continue to train our random forest classifier on the imbalanced set, but adjusting the weights (or priors) of the three classes based on their distribution. Disappointingly, the resulting confusion matrix, given below in Table 12, is essentially the same as Table 11. Only one additional serious accident is predicted correctly, yielding a marginally better overall error rate of 0.1206.

Table 12: RF Trained on Imbalanced Data with Class Weights

	1	2	3
1	0	0	4
2	0	4	54
3	0	0	419

Hence, we will move on to a second approach, which is to continue to use the imbalanced training set, but employing stratified sampling based on the three classes of accident severity. In so doing, we are essentially undersampling within the imbalanced training set. Since there are only 14 observations in the training set of class "fatal", we will draw 14 samples from each class and train our random forest classifier based on this subset. The results on the test set are given in Table 13 below.

Table 13: RF Trained on Imbalanced Data with Stratified Sample

	1	2	3
1	1	3	0
2	8	25	25
3	36	96	287

The overall error rate is now 0.3493, which is substantially higher than the error we obtained earlier when the model was trained on the original imbalanced set. In addition, we finally predicted one case of fatal accident and twenty-five cases of serious accidents correctly. Specifically, now we have a recall of $\frac{1}{1+3} = 0.25$ for class "fatal" (compared to 0 previously), and a recall of $\frac{25}{58} = 0.431$ for class "serious" (compared to $\frac{3}{58} = 0.052$ previously).

In our problem, misclassifying fatal or serious crashes is deadly, while misclassifying minor bumps is relatively less important. Thus, despite it having a much higher overall error rate, we actually prefer

this stratified sampling approach because it yields much higher recall rates for fatal and serious accidents.

We will now see if we can further improve performance by examining a third approach, which is to train our classifier on a balanced data set. In other words, we will undersample from the majority class until there are an equal number of observations from each of the three accident severity classes.

First, using the default settings in R, we will train a single decision tree classifier on this balanced data set. The results are given in Table 14 below.

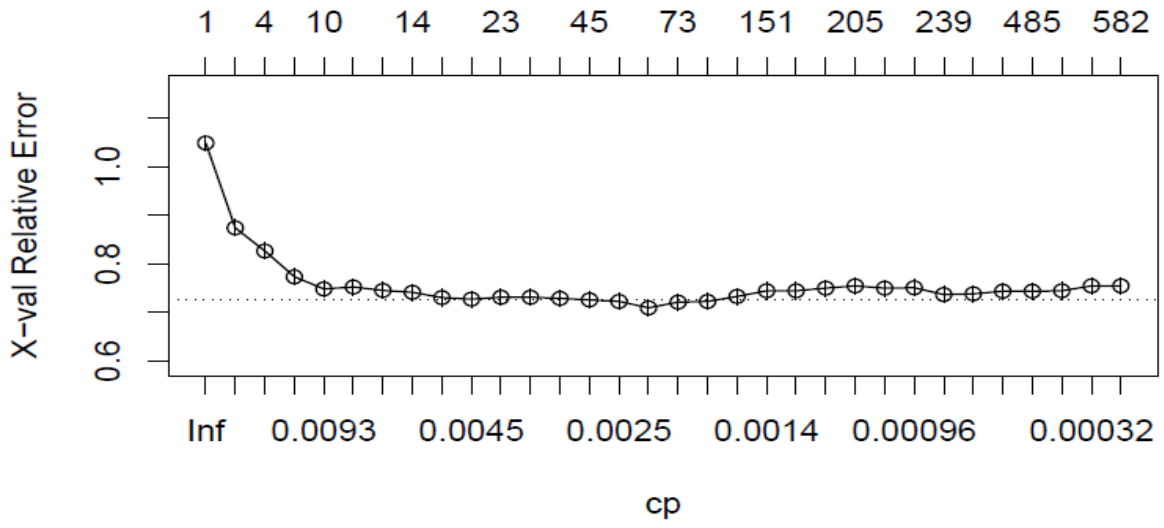
Figure 14: Single Tree Trained on Balanced Data

	1	2	3
1	3	1	0
2	20	16	22
3	57	63	299

The table shows some improvement in the number of correctly predicted fatal accidents. The recall for class "fatal" is $\frac{3}{3+1} = 0.75$, the highest of all approaches so far. The overall error rate is 0.3389, slightly better than the second approach.

In an attempt to further improve this approach, we can use cross-validation (CV) for the selection of the complexity parameter (cp) and the size of the tree (nsplit), with 10 CV folds. Figure 6 shows that the CV relative error at first decreases sharply as the size of the tree grows, but the change in the error is very small once the size of the tree is bigger than 10.

Figure 6: Cross Validation of Complexity Parameter/Tree Size



The complexity of the tree that gives the lowest CV error is 0.0021. Building a single pruned tree of this complexity gives the results shown in Table 15 below.

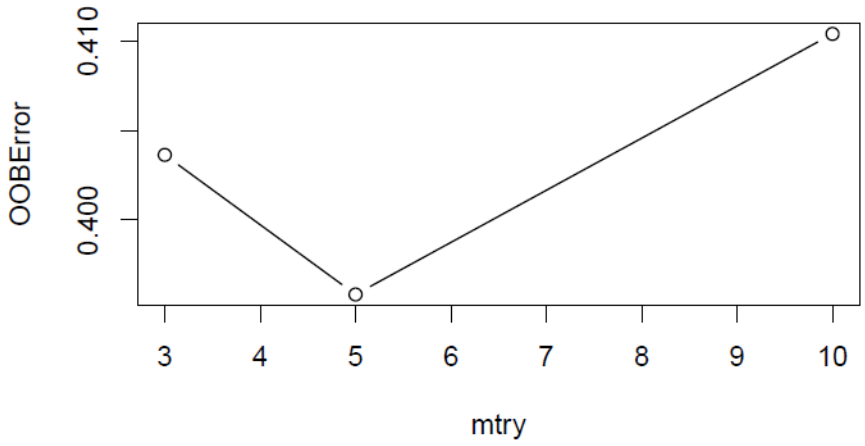
Table 15: Pruned Single Tree Trained on Balanced Data

	1	2	3
1	3	1	0
2	17	29	12
3	60	109	250

This pruned tree shows better performance in our metric of interest. While the overall error rate rises to 0.4137, the recall for serious accidents improves to $\frac{29}{58} = 0.5$ from $\frac{16}{58} = 0.276$. The recall for fatal accidents remains the same.

Next, we will train a random forest classifier on the balanced training set to see if there is any improvement over the single decision tree approach above. Similar to when we trained random forest on the imbalanced set, we will first tune the algorithm for the optimal *mtry*, as shown below in Figure 7.

Figure 7: Out-of-Bag Error vs. Number of Predictors Considered at Each Split



From the plot, we see that again when *mtry* = 5, the out-of-bag error (OOB) is the lowest. We will then use this value to train our random forest classifier on the balanced data set. The performance is given in Table 16 below.

Table 16: RF Trained on Balanced Data

	1	2	3
1	4	0	0
2	5	50	3
3	43	93	283

We see that for the first time, all four fatal accidents are predicted correctly! Additionally, the recall for serious accidents is $\frac{50}{58} = 0.862$, also the highest from all tree-based approaches. Though the

overall error rate with this approach (0.2994) is still not great, as explained earlier, we are much more concerned about correctly predicting fatal and serious accidents than slight accidents.

In summary, Table 17 compares the performances of the tree-based classifiers discussed above.

Table 17: Comparison of Classifiers

Classifiers	'Fatal' Recall	'Serious' Recall	Pred. Err.
Imbalanced:			
Single Tree	0	3/58	0.1331
RF	0	3/58	0.1227
RF w/ classwt	0	4/58	0.1206
RF w/ strata	1/4	25/58	0.3493
&			
Balanced:			
Single Tree	3/4	16/58	0.3389
CV Single Tree	3/4	29/58	0.4137
RF	1	50/58	0.2994

Conclusion

Overall, classification methods applied to the original imbalanced training set yielded a lower prediction error compared to methods trained on the balanced set. However, since we place greater value on correctly predicting the relatively few cases of fatal and serious accidents, recall for these two minority classes replaces overall prediction error as the most important metric of evaluation. Compared with methods trained on the original imbalanced data, those trained on the balanced set yielded higher recall, with the random forest classifier achieving perfect recall for fatal accidents and over 85 percent recall for serious accidents.

The results of this study suggest that, when working with imbalanced data, commonly used classifiers in accident prediction studies tend to misidentify the minority class despite obtaining relatively high overall prediction accuracy. Given that many applied researchers are interested in correctly classifying the minority class, it is important to raise caution that direct application of traditional classification methods and metrics may be inappropriate. By undersampling from the majority class to balance the data and shifting the performance metric from overall accuracy to recall of severe accidents, we found that classifications methods can be more effectively utilized to classify observations from the minority class. These results are consistent with those obtained by other researchers investigating performance improvement of classifiers in the presence of imbalanced data generally, but we believe this to be the first study to apply such techniques to accident data specifically.

References

Benf. "UK Accidents 10 Years History with Many Variables." Kaggle. 2018. Accessed 2018. https://www.kaggle.com/benoit72/uk-accidents-10-years-history-with-many-variables.

Chen, Chao, Andy Liaw, and Leo Breiman. "Using Random Forest to Learn Imbalanced Data." (2004)

Chong, Miao, Ajith Abraham, and Marcin Paprzyncki. "Traffic Accident Analysis Using Machine Learning Paradigms." Information 29 (2005): 89-98.

Garrido, Rui, Ana Almeida, and Jose Paulo Elvas. "Prediction of Road Accident Severity Using the Ordered Probit Model." Transportation Research Procedia 3 (July 2, 2014): 214-23.

Iranitalab, Amirfarrok, and Aemal Khattak. "Comparison of Four Statistical and Machine Learning Methods for Crash Severity." Accident Analysis and Prevention. 108 (2017): 27-36.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibsharani. An Introduction to Statistical Learning with Applications in R. Vol. 1. 2013

Kunt, Mehmet, Iman Aghayan, and Nima Noii. "Prediction for Traffic Accident Severity: Comparing the Artificial Neural Network, Genetic Algorithm, Combined Genetic Algorithm and Pattern Search Methods." Transport. 26, no. 4 (2011): 353-66.

Mazuroski, Maciej, Piotr Habas, Jace Zurada, Joseph Lo, Jay Baker, and Giorgia Tourassi. "Training Neural Network Classifiers Fro Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance." Neural Networks 21 (2008): 427-236

Pozzolo, Andrea, Oliver Caelen, Gianluca Bontempi. "When is undersampling effective in unbalanced classification tasks?". European Conference on Machine Learning. (2015).

Ren, Honglei, You Song, Jingxin Liu, Yucheng Hu, and Jinzhi Lei. "A Deep Learning Apprach to the Prediction of Short Term Traffic Accident Risk.

Zong, Fan, Huiyong Zhang, Hongguo Xu, Xiumei Zhu, and Lu Wang. "Predicting Severity and Duration of Road Traffic Accident." Mathematical Problems in Engineering, 2013.