Daniel Sandoz

Final Part 1

Step 1: Develop an understanding of the data mining project

In this project, we have a data set that is obtained from the department of transportation and we are trying to perform a data mining analysisthat predicts the number of fatalities that will result from a given accident. There are many variables in this data set, so identifying the main predictors will be important.

Step 2: Obtain the dataset to be used in the analysis

The data set is found here https://www.kaggle.com/usdot/nhtsa-traffic-fatalities

Step 3: Explore, clean, and preprocess the data & Step 4: Reduce the data dimension

I combined these two steps. I viewed the structure of the data and the number of variables involved in this data set. I then went on to delete the variables that were either irrelevant to the business question, had a large amount of missing values or could be explained better by other more significant variables. In order to run a confusion matrix, I also turned the variables into factors, so I would be able to gage the accuracy and compare across different models using a confusion matrix. The random forest model cannot run with variables that have factors of more than 53 levels. I thought that time of notification and time of arrival were important predictors when modeling fatalities. To combat this I turned the time of notification and time of arrival into dummy variables’ grouped the variable minutes of notification and min of arrival into dummy groups with a 10 min span to allow for less than 53 levels(6 levels each). I then turned all the variables that I thought were relevant into factors so that they could run on the confusion matrix. It turns out that running a CM after dropping the time of notification and time of arrival from the data set resulted in the same accuracy as dummying these variables into time segments, however I thought that these variables were relevant and would help increase the accuracy of my random forest model, but they did not. For Random Forest, I brought the target variable: number of fatalities to the front.

Step 5: Determine the data mining task

Predict the number of fatalities that will result from a given accident.

Step 6: Partition the data

For Random Forest I partitioned the data into an 80/20 split as standard for that algorithm. For LDA, I partitioned the data into a 70/30 split as standard for that algorithm.

Step 7: Choose the data mining techniques to be used

I decided that Random Forest would be a good technique to allow for me to understand what variables contribute to the ability to predict fatalities, I then revisited step 4 and eliminated the variables that had a small or no impact on the model to allow for the least amount of processing resources to be used. I chose LDA because it is suitable for large data sets. In the scope of the fatalities data set, each data point had at least one death. I thought it would be interesting to see what variables contributed heavily to an outcome of more than one death. LDA works well for this situation because it wants observations in each class to be as close as possible, but it also wants each class to be as different as possible. With a large data set like we have, LDA is able to point out these differences and allow us to understand what variables contribute to more than one death.

Step 8: Use algorithms to perform the task

I used Random Forest and Linear Discriminate Analysis.

Step 9: Interpret the results

LDA split the data to reference the number of fatalities in groups from 1-6 to help us understand what variables are important to predict the increase of fatalities. The probabilities of the amount of fatalities that occurred can be seen here:

1 2 3 4 5 6

0.9249214959 0.0636597202 0.0074222095 0.0022837568 0.0014273480 0.0002854696

We can see that with National highway system, (being a binary variable) all accidents that occurred with 6 fatalities were on the highway as seen here:

national\_highway\_system

1 0.3712963

2 0.4080717

3 0.4230769

4 0.5000000

5 0.6000000

6 1.0000000

LDA confusion matrix:

Accuracy : 0.9098

95% CI : (0.8942, 0.9238)

No Information Rate : 0.9272

P-Value [Acc > NIR] : 0.9948

Kappa : 0.0235

Mcnemar's Test P-Value : NA

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6

Sensitivity 0.98055 0.010526 0.000000 0.000000 0.000000 NA

Specificity 0.04587 0.997147 0.990585 0.999331 0.995989 0.995992

Pos Pred Value 0.92901 0.200000 0.000000 0.000000 0.000000 NA

Neg Pred Value 0.15625 0.936997 0.993257 0.997995 0.999329 NA

Prevalence 0.92719 0.063460 0.006680 0.002004 0.000668 0.000000

Detection Rate 0.90915 0.000668 0.000000 0.000000 0.000000 0.000000

Detection Prevalence 0.97862 0.003340 0.009352 0.000668 0.004008 0.004008

Balanced Accuracy 0.51321 0.503837 0.495293 0.499665 0.497995 NA

This is a very accurate confusion matrix with an accuracy of .9080.

The rest of the LDA results are attached at the bottom.

For Random Forest, we can see the mean decrease in accuracy for each variable which allows us to see what variables have a significant impact on the model, and ones that do not. This allows for me to go back and delete these from the model to allow for a decrease in processing resources. It seems that the time of notification and arrival have a small impact, however when they are broken into dummys this takes away from their total impact. However, it is apparent in this Random Forest graph that these are still important predictors when they are combined as this can be inferred from the hour of arrival and notification. If the hour of notification and arrival are important, then obviously the minutes add statistical significance as well, however with the nature of the Random Forest Algorithm, it is difficult to model this accurately.

drunk\_drivers 10.0550025

crash\_level\_3 0.7099655

crash\_level\_2 1.9397471

crash\_level\_1 8.4899986

hour\_of\_arrival 30.8652883

hour\_notif 28.8808067

atmospheric 18.1711594

light 10.9568108

trafficway 5.7549524

junction\_specific\_location 5.9412157

mancollision 15.9880205

firstevent 18.6956684

route 19.0005110

owner 15.8256619

highway 5.6759975

hourcrash 29.4320915

day 27.0224306

persons\_in\_motor\_vehicles\_in\_transport\_mvit 47.0106233

persons\_not\_in\_motor\_vehicles\_in\_transport\_mvit 7.8038494

parked\_working\_vehicles 2.8534293

motor\_vehicles\_in\_transport\_mvit 11.2790976

fatalities.df.clean.school\_bus\_related 0.9452007

fatalities.df.clean.work\_zone 1.7487856

fatalities.df.clean.type\_of\_intersection 7.8467631

W\_in\_10min\_notif 0.7760585

W\_in\_20min\_notif 2.0774810

W\_in\_30min\_notif 1.5491660

W\_in\_40min\_notif 1.2675186

W\_in\_50min\_notif 0.5786308

w\_in\_60min\_notif 1.4090074

W\_in\_10min\_arr 1.6701174

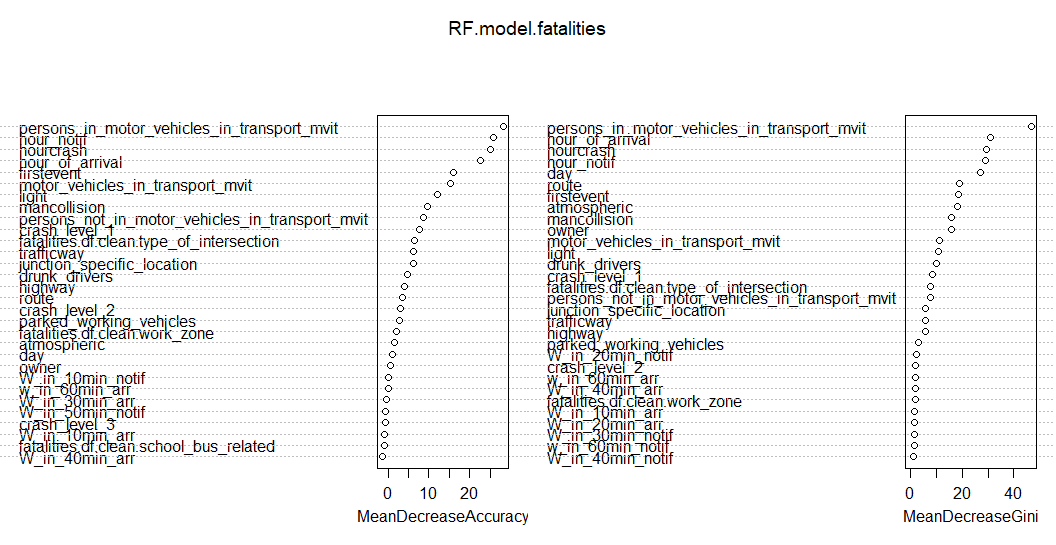
W\_in\_20min\_arr 1.6048740

W\_in\_30min\_arr 1.0814645

W\_in\_40min\_arr 1.8447942

W\_in\_50min\_arr 1.1580563

w\_in\_60min\_arr 1.8544743



# Step 10: Deploy the model:

|  |
| --- |
| Random Forest:  Accuracy : 0.9269  95% CI : (0.9089, 0.9422)  No Information Rate : 0.9269  P-Value [Acc > NIR] : 0.5311    Kappa : 0  Mcnemar's Test P-Value : NA  Statistics by Class:  Class: 1 Class: 2 Class: 3 Class: 4 Class: 5  Sensitivity 1.0000 0.00000 0.000000 0.000000 0.000000  Specificity 0.0000 1.00000 1.000000 1.000000 1.000000  Pos Pred Value 0.9269 NaN NaN NaN NaN  Neg Pred Value NaN 0.93687 0.992986 0.997996 0.998998  Prevalence 0.9269 0.06313 0.007014 0.002004 0.001002  Detection Rate 0.9269 0.00000 0.000000 0.000000 0.000000  Detection Prevalence 1.0000 0.00000 0.000000 0.000000 0.000000  Balanced Accuracy 0.5000 0.50000 0.500000 0.500000 0.500000 |
| LDA:  Accuracy : 0.9098  95% CI : (0.8942, 0.9238)  No Information Rate : 0.9272  P-Value [Acc > NIR] : 0.9948    Kappa : 0.0235  Mcnemar's Test P-Value : NA  Statistics by Class:  Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6  Sensitivity 0.98055 0.010526 0.000000 0.000000 0.000000 NA  Specificity 0.04587 0.997147 0.990585 0.999331 0.995989 0.995992  Pos Pred Value 0.92901 0.200000 0.000000 0.000000 0.000000 NA  Neg Pred Value 0.15625 0.936997 0.993257 0.997995 0.999329 NA  Prevalence 0.92719 0.063460 0.006680 0.002004 0.000668 0.000000  Detection Rate 0.90915 0.000668 0.000000 0.000000 0.000000 0.000000  Detection Prevalence 0.97862 0.003340 0.009352 0.000668 0.004008 0.004008  Balanced Accuracy 0.51321 0.503837 0.495293 0.499665 0.497995 NA |
| |  | | --- | |  | |

Here we can see that Random Forest has a higher accuracy rate, however the sensitivity and specificity results can eliminate any consideration of deploying this model. LDA has a high accuracy, and overall LDA is better because it helps us to determine what variables are key predictors when trying to predict what circumstances involve more than one fatality.

LDA results:

Group means:

number\_of\_motor\_vehicles\_in\_transport\_mvit number\_of\_parked\_working\_vehicles

1 1.519136 0.03858025

2 1.860987 0.03139013

3 2.076923 0.23076923

4 2.250000 0.00000000

5 1.600000 0.00000000

6 2.000000 0.00000000

number\_of\_persons\_not\_in\_motor\_vehicles\_in\_transport\_mvit

1 0.1953704

2 0.1255605

3 0.1923077

4 0.0000000

5 0.0000000

6 0.0000000

number\_of\_persons\_in\_motor\_vehicles\_in\_transport\_mvit day\_of\_week hour\_of\_crash

1 2.207099 4.170370 13.39568

2 3.524664 4.493274 13.25561

3 4.730769 4.692308 13.15385

4 5.125000 4.250000 13.37500

5 6.000000 1.800000 9.00000

6 12.000000 4.000000 2.00000

national\_highway\_system ownership route\_signing first\_harmful\_event manner\_of\_collision

1 0.3712963 17.54691 3.486420 18.40772 1.823765

2 0.4080717 15.16592 3.076233 17.04933 2.816143

3 0.4230769 24.69231 3.038462 15.84615 1.923077

4 0.5000000 13.25000 2.375000 12.00000 3.250000

5 0.6000000 20.80000 2.600000 15.40000 3.000000

6 1.0000000 1.00000 1.000000 12.00000 1.000000

relation\_to\_junction\_specific\_location relation\_to\_trafficway light\_condition

1 1.969753 2.290432 1.798457

2 1.744395 1.901345 1.717489

3 2.000000 2.038462 1.692308

4 3.125000 1.000000 1.625000

5 1.400000 2.400000 1.800000

6 1.000000 1.000000 2.000000

atmospheric\_conditions hour\_of\_notification minute\_of\_notification hour\_of\_arrival\_at\_scene

1 8.472222 12.93519 29.57716 13.01605

2 6.327354 13.34081 30.13004 13.32287

3 2.538462 13.15385 32.03846 13.30769

4 2.125000 13.62500 24.87500 13.75000

5 41.800000 9.00000 29.00000 9.20000

6 1.000000 2.00000 27.00000 2.00000

minute\_of\_arrival\_at\_scene related\_factors\_crash\_level\_1 related\_factors\_crash\_level\_2

1 29.55278 1.538272 0.5824074

2 29.74439 2.403587 1.0896861

3 33.07692 2.730769 0.0000000

4 27.12500 1.750000 0.0000000

5 26.20000 0.000000 0.0000000

6 42.00000 0.000000 0.0000000

related\_factors\_crash\_level\_3 number\_of\_drunk\_drivers

1 0.4277778 0.2685185

2 0.8878924 0.3946188

3 0.0000000 0.4615385

4 0.0000000 0.2500000

5 0.0000000 0.4000000

6 0.0000000 0.0000000

Coefficients of linear discriminants:

LD1 LD2 LD3

number\_of\_motor\_vehicles\_in\_transport\_mvit -0.5417908451 -1.2052580598 0.039224519

number\_of\_parked\_working\_vehicles 0.2841141952 -0.3755442383 -2.059195416

number\_of\_persons\_not\_in\_motor\_vehicles\_in\_transport\_mvit 0.1724669768 -0.3444150901 -0.343344498

number\_of\_persons\_in\_motor\_vehicles\_in\_transport\_mvit 0.7276162607 0.3003608619 -0.090410844

day\_of\_week 0.0285756994 -0.2235715014 -0.005323319

hour\_of\_crash -0.0042850650 -0.0012290184 -0.018555690

national\_highway\_system -0.0872986363 0.2629731232 -0.051063227

ownership -0.0002877750 0.0006422682 -0.007050525

route\_signing -0.1181617558 0.0312990539 -0.078769863

first\_harmful\_event 0.0006180421 -0.0102034662 0.001028302

manner\_of\_collision -0.0058663856 0.0030379397 0.079147033

relation\_to\_junction\_specific\_location -0.0165131824 0.0117984647 -0.057625049

relation\_to\_trafficway -0.0214931308 -0.0218287294 -0.088227604

light\_condition -0.0296921053 0.1099241117 -0.010469071

atmospheric\_conditions -0.0004894492 0.0201901519 0.007826144

hour\_of\_notification 0.0092977079 -0.0578349566 0.150733529

minute\_of\_notification -0.0003548036 -0.0031198022 0.001136925

hour\_of\_arrival\_at\_scene -0.0112320392 0.0150532574 -0.121300013

minute\_of\_arrival\_at\_scene 0.0019254684 -0.0017201601 -0.013853916

related\_factors\_crash\_level\_1 0.0046495348 -0.0253992312 -0.020872104

related\_factors\_crash\_level\_2 -0.0600521528 0.0198333433 0.113936176

related\_factors\_crash\_level\_3 0.0595130406 -0.0018215818 -0.062365833

number\_of\_drunk\_drivers 0.5331517338 -0.6689539582 0.355956944

LD4 LD5

number\_of\_motor\_vehicles\_in\_transport\_mvit 0.3907643047 -0.499498012

number\_of\_parked\_working\_vehicles 0.8273441824 -0.240274479

number\_of\_persons\_not\_in\_motor\_vehicles\_in\_transport\_mvit 0.3698739224 0.630909906

number\_of\_persons\_in\_motor\_vehicles\_in\_transport\_mvit -0.1051341922 0.098255258

day\_of\_week -0.1221307977 0.152615976

hour\_of\_crash -0.0035149453 -0.003372612

national\_highway\_system -0.0084826931 0.283283181

ownership 0.0068326731 -0.005209837

route\_signing 0.0396333343 0.121196204

first\_harmful\_event 0.0005214861 0.016626577

manner\_of\_collision 0.0379176328 0.028435974

relation\_to\_junction\_specific\_location -0.0530668228 -0.164980843

relation\_to\_trafficway 0.1093419642 0.033435244

light\_condition -0.1953129298 0.027406934

atmospheric\_conditions 0.0245370717 -0.004764367

hour\_of\_notification 0.0251782938 0.088547527

minute\_of\_notification 0.0135503805 0.014217499

hour\_of\_arrival\_at\_scene 0.0035891511 -0.123494506

minute\_of\_arrival\_at\_scene -0.0073467285 0.013854065

related\_factors\_crash\_level\_1 0.0010470515 -0.015699221

related\_factors\_crash\_level\_2 -0.0347462080 0.087699243

related\_factors\_crash\_level\_3 0.0313427171 -0.045980894

number\_of\_drunk\_drivers 1.1713967481 0.293132962