UNDERSTANDING PATTERNS OF ROAD SAFETY

Shuonan Zhang, Yash Gupta, Chethan Singh Mysore Jagadeesh,
Rae-Djamaal Wallace

Group 4 Mar 2019

EXECUTIVE SUMMARY

Predictive Features Based on Importance Level

- High Importance Features: Junction Detail, Light Conditions, Hours, and Speed Limit
- Medium Importance Features: Road Class and Road Type
- Low Importance Features: Policy Attendance, Urban or Rural Areas, Day of Week, and Road Surface Condition

Chosen Predictive Model

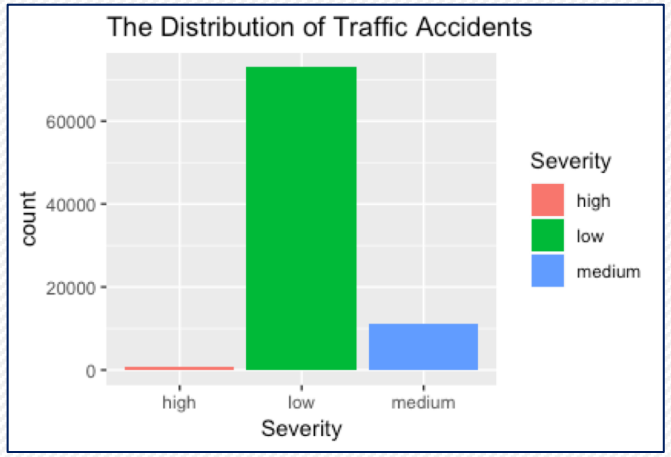
- Logistic Regression Model (Accuracy of 87 % - Slightly higher than the default classifier
- Using only important features did not decrease predicting accuracy

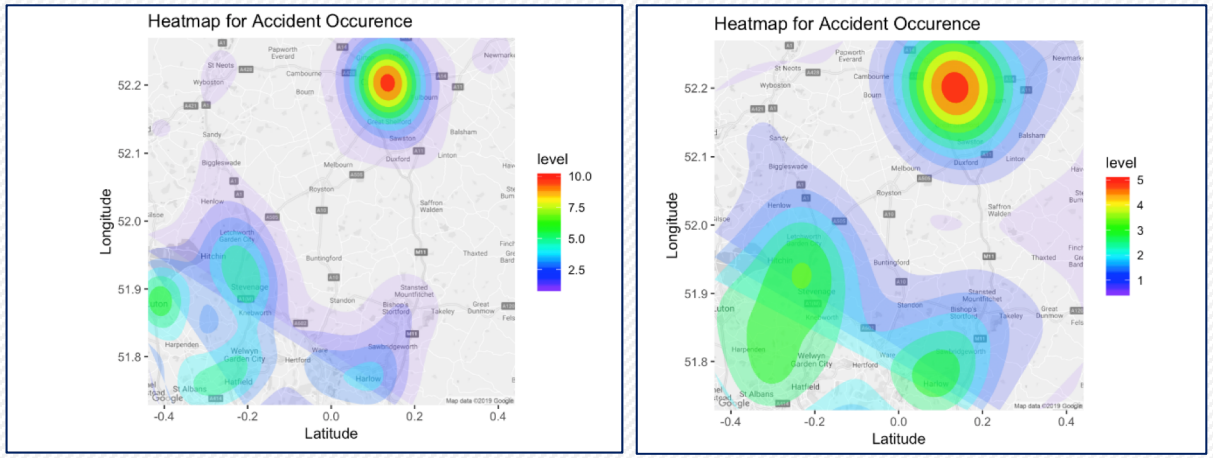
Policy Recommendations

- ✓ Preventing Severe Accidents
  - ✓ Adjust speed limit for roads with many accident records and Improve light condition
- ✓ Policy Jurisdiction
  - ✓ Increase police force in Staggered T junction, private entrance / drive way, and one way roads where severe accidents are more likely to happen

PROJECT OVERVIEW AND DATA PROCESSING

- Objectives
  - Select informative features regarding high-severity accidents
  - Provide policy recommendations to improve road safety
- Data
  - Subset of the public collection of data of the circumstances of personal injury road accidents in Great Britain in 2012.
  - Features
    - Extremely Unbalanced → Combined the medium and high severity group and upsampled the severity level, which did the following:
      - Created a balanced dataset
      - Reduced classification complexity
      - Avoided overly strong assumption on the distribution of the severity of traffic accident
    - All Categorical → Recoded the number to factors / Combined levels within certain features
    - Pre-selection
      - Removed features that vary at individual level (E.g. Road id)
      - Remove location related information: No identifiable differences in the geographical distribution





All Accidents (left) and Severe Accidents (right)
No significant difference in geographical distribution

FEATURE SELECTION - APPROACHES

Calculating Chi-squared statistics
”Pedestrian Control” Removed

CORRELATION ANALYSIS

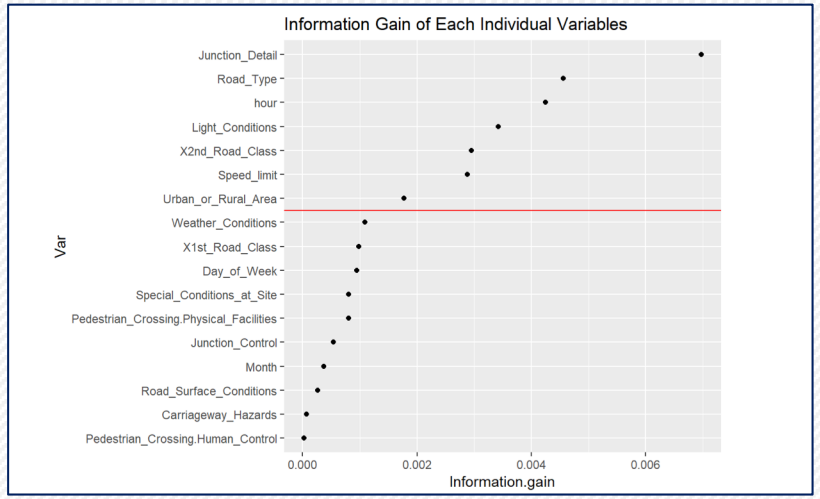
Step AIC() removed
“Month” indicator

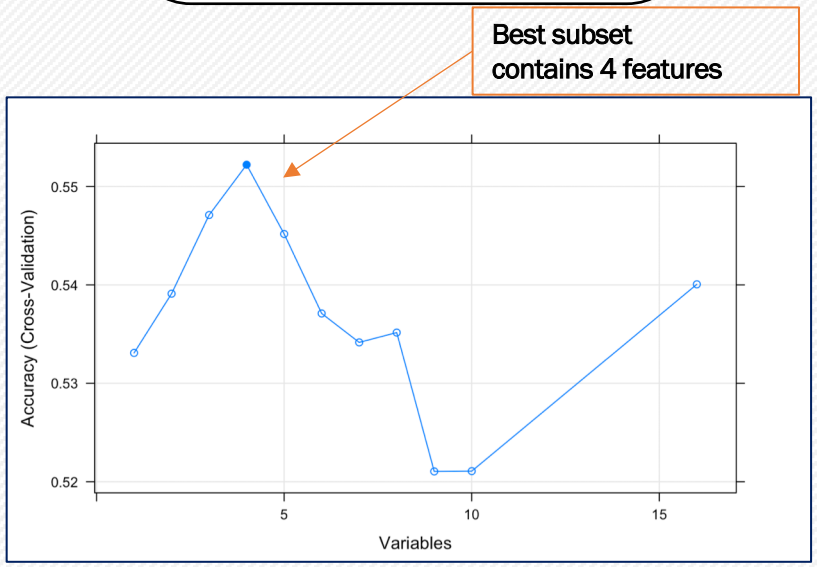
STEP AIC WITH LOGISTIC REGRESSION

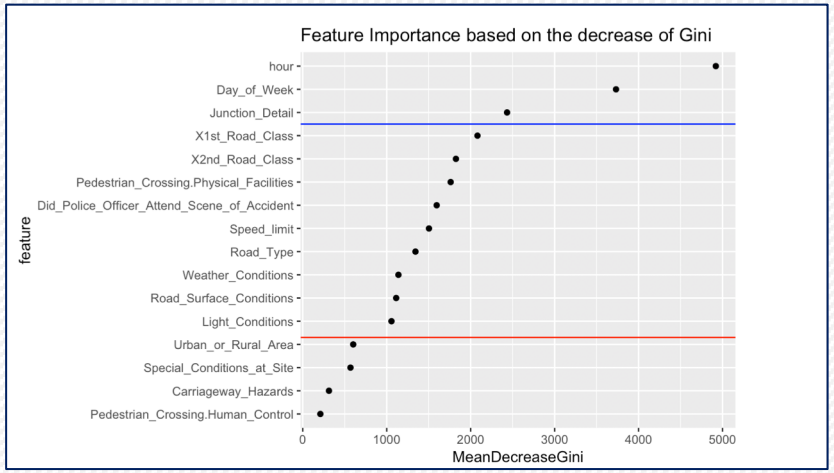
INFORMATION GAIN

RANDOM FOREST (1)

RANDOM FOREST (2)








```
[1] "Light_Conditions"                    "Carriageway_Hazards"
[3] "Did_Police_Officer_Attend_Scene_of_Accident" "Junction_Detail"
```

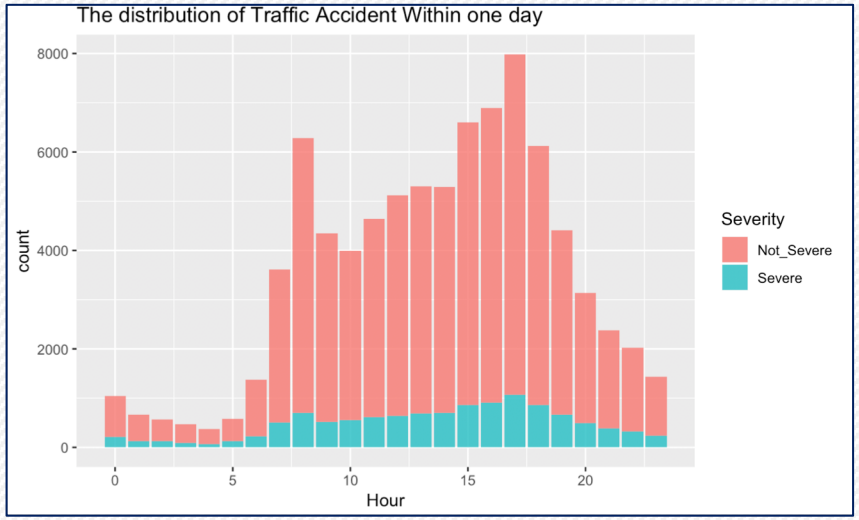
FEATURE SELECTION - RESULTS

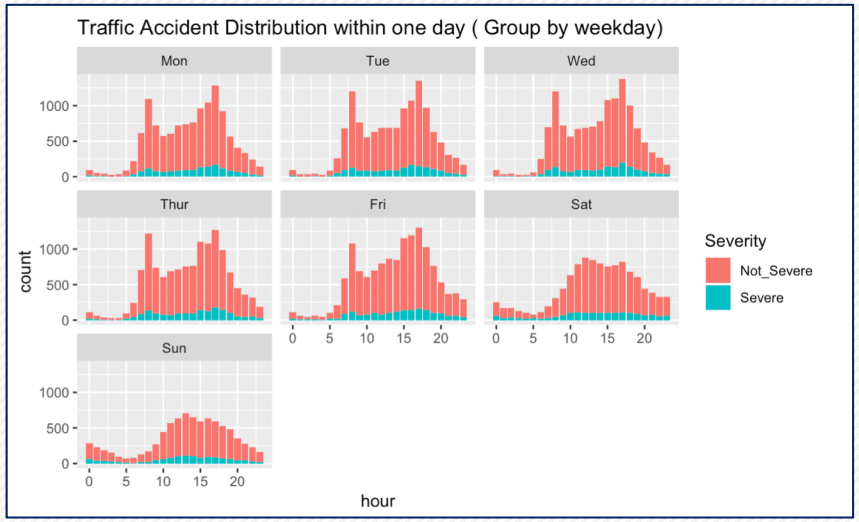
All Fetures	Correlation Analysis	Step AIC	Info gain	Random forest	Random forest 2	Judged Based on Intuition	Importance Level
Junction_Detail						1	1
Light_Conditions						1	1
hour						1	1
Road_Type						3	2
X2nd_Road_Class						3	2
Speed_limit						1	1
1st_Road_Class						3	2
Did_Police_Officer_At						3	3
Urban_or_Rural_Area					Removed	3	3
Day_of_Week						2	3
Pedestrian_Crossing.H						0	0
Road_Surface_Conditi						2	3
Pedestrian_Crossing.Pl						Uncontrollable	0
Weather_Conditions						0	0
Carriageway_Hazards					Removed	Uncontrollable	0
Junction_Control	Romoved					0	0
Month		Removed				0	0
Special_Conditions_at					Rmoved	Uncontrollable	0

Table 1: Feature Selection Outcomes
(Blue colored cell indicates features selected by certain methods)

- The final selection result is based both on algorithms and intuition
- Features are divided into three importance levels where “1” represents the highest importance

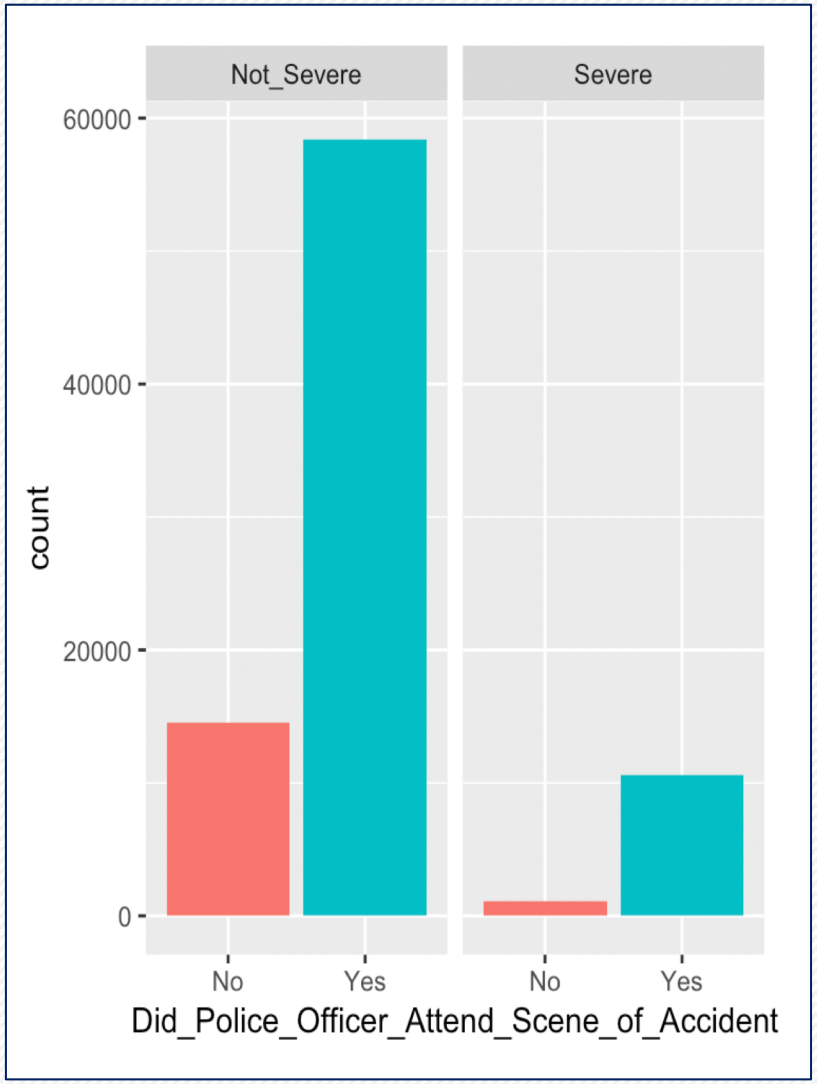
SELECTED FEATURE VISUALIZATION





- No significant differences in the distribution of severe and not severe group among different time within one day. We see more accidents and also more severe accidents during the rush hours.
- The distribution of accident within one days seem to be different during weekends More specifically, the number of severe accidents seems to be more uniformly distributed during the day as opposed to have peak numbers at certain hours

- Police force is more likely to attend the scene of severe accident

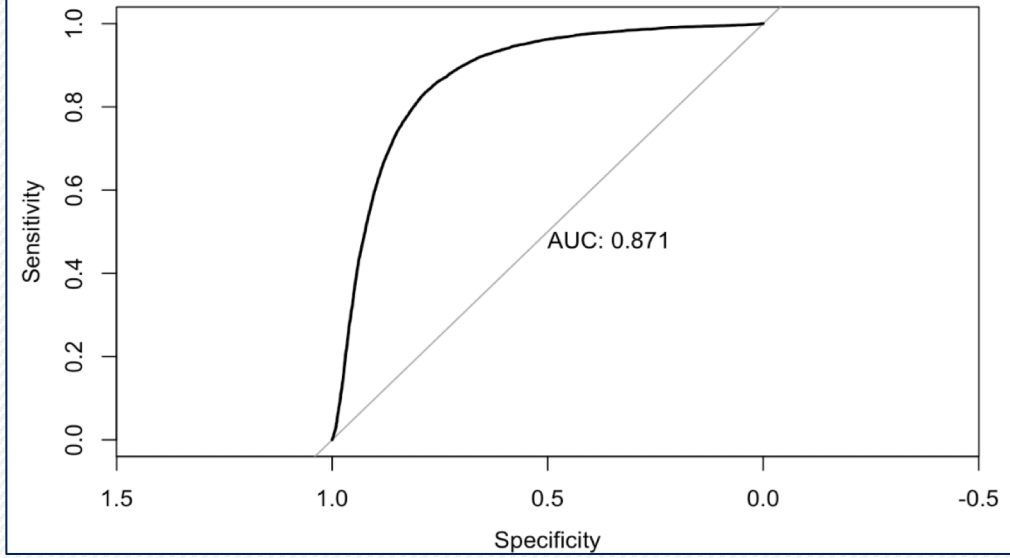


Note: Clustering was also implemented, but the visualized result is not informative (see the html)

PREDICTIVE ANALYSIS – MODEL SELECTION WITH 10-FOLD CV

Model Name	Accuracy	Notes
Random Forest	7%	
Naïve Bayes	86%	• Extreme low sensitivity – considerably limited by the prior probability
SVM	Not Implemented	• SVM uses gradient descent that works best with continuous variables and it assumes that the observations can be divided by a linear hyperplane, which is not applicable for our dataset (with all categorical variables).
Logistic Regression: high importance features	87%	• Logistic regression yields predicting accuracy higher than the default classifier • Shortest model training time • When reducing the number of the features to only the most important features, no significant decrease in predicting accuracy.
Logistic Regression: high + Medium importance features	87%	
Logistic Regression: high + Medium+ Low importance features	87%	

Table 2: Model Comparison



Note:

- Accuracy of the default classifier is 86%
- For random forest and logistic regression, the models are trained with the Upsampled data and tested on the unbalanced data reflecting the reality
  - Considerably Higher predicting accuracy can be achieved if the test set is balanced as well

CONCLUSIONS AND RECOMMENDATIONS

Predictive Features Based on Importance Level

- High Importance Features: Junction Detail, Light Conditions, Hours, and speed limit
- Medium Importance Features: Road Class and Road Type
- Low Importance Features: Policy Attendance, Urban or Rural Areas, and Day of Week, and Road surface condition

Predictive Model

- Logistic Regression Model (Accuracy of 87 % - Slightly higher than the default classifier
- Using only important features did not decrease predicting accuracy

Policy Recommendations

- ✓ Preventing Severe Accident
  - ✓ Adjust speed limit for road with many accident records and Improve light condition
- ✓ Policy Jurisdiction
  - ✓ Increase police force in Staggered T junction, private entrance / drive way, and one way roads where severe accidents are more likely to happen

Table 3: The Percentage of Severe Accident

Junction_Detail	Count	Percentage
T/staggered	6945	0.15366
Private drive or entrance	636	0.15132
More than 4 arms	273	0.13943
Other junction	391	0.13797
Crossroads	1850	0.13215
SlipRoad	242	0.11646
Mini-roundabout	162	0.09963
Roundabout	1201	0.09432

Road_Type	Count	Percentage
Single carriageway	9357	0.14738
One way	237	0.14434
Dual carriageway	1045	0.12248
Roundabout	946	0.09783
Slip road	98	0.09032
Unknown	17	0.08057

PROJECT REFELCTION

- Only have variables describing the external conditions
  - No information about the drivers and vehicles --- Equally important predictors
- All features are categorical variables
  - Limits choices of classification algorithms and effective visualization methods
- The distribution of outcome variable is unbalanced
  - The prior possibility constrained the performance of Naive Bayes model

    - Hard to achieve higher predicting accuracy than a default classifier (86%)
  - No enough information about the high-severity traffic accidents.