Exploring the Potential of Machine Learning in Insurance Claims

- Dataset: UK Traffic Accidents

Contents

INTRODUCTION

Data Set ： UK Traffic Accident 2005

- Goal: Help Insurance Company To
  - Evaluate insurance claims payment
  - Evaluate underwriting of insurance plan
  - Insurance claim fraud prediction/investigation
- Three Target Accident Severity
  - Slight
  - Serious
  - Fatal
- Multi-Classification Problem

DATA CLEANING

First...

- Replace Errant/Misspelled Values
- Drop Blank Rows with Blank Values
- Format Date in Dateline format and add Column for Month and Hour

Overview of UK Accidents Dataset

Overview

Dataset info

Number of variables	25
Number of observations	126288
Total Missing (%)	0.0%
Total size in memory	24.1 MiB
Average record size in memory	200.0 B

Variables types

Numeric	13
Categorical	10
Boolean	0
Date	1
Text (Unique)	0
Rejected	1
Unsupported	0

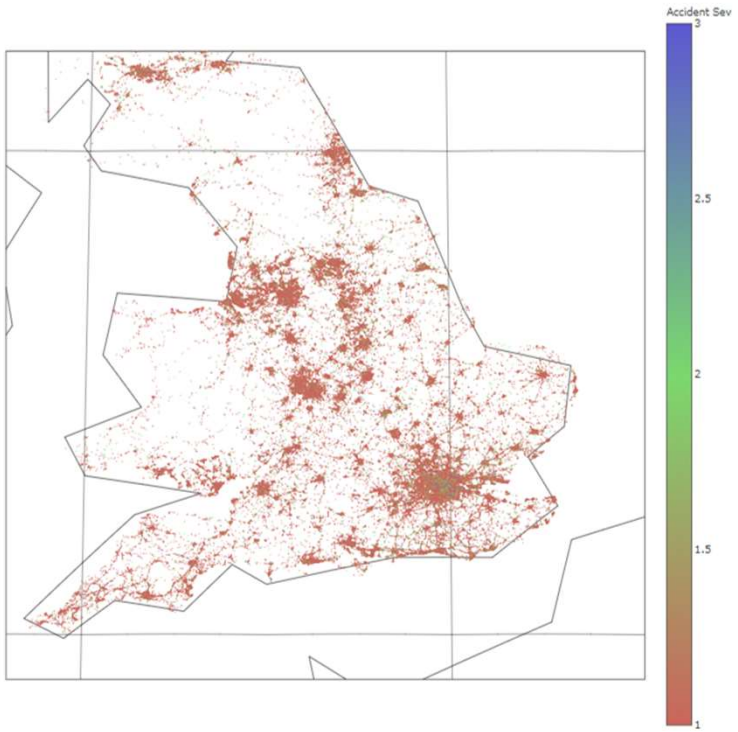
- 88% of the data belongs to slight and 11% to serious and 1% to fatal
- 71.6% of accident occur in Urban while 28.4 % occur in Rural
- A lot of the data are skewed to one variable

Metadata Information

- Accident Index: index identifier of accidents
- Longitude: longitude coordinates of accident
- Latitude: latitude coordinates of accident
- Accident Severity: the severity of accident
- Carriageway Hazard: none/ other object on road/ any animal in carriageway/ pedestrian in carriage/ previous accident/ vehicle load on road
- Date: the date that accident that occurred in the format DD/MM/YYYY
- Did Police officer attend scene of accident? (1- No, 2- Yes, 3- Yes with Ambulance)
- Junction Control: Was there junction control at the location of accident
- Light Conditions: the light condition of accident
- Number of Casualties
- Number of Vehicles
- Pedestrian crossing-human control
- Pedestrian crossing physical facilities
- Road Surface Conditions
- Road Type
- Special Conditions at site
- Speed limit
- Urban or Rural Area
- Weather Conditions

DATA VISUALIZATION

Map


Accident Sev
3
2.5
2
1.5
1


England
NORTH SEA
IRISH SEA
ENGLISH CHANNEL
BELFAST
MANCHESTER
LONDON
CARDIFF

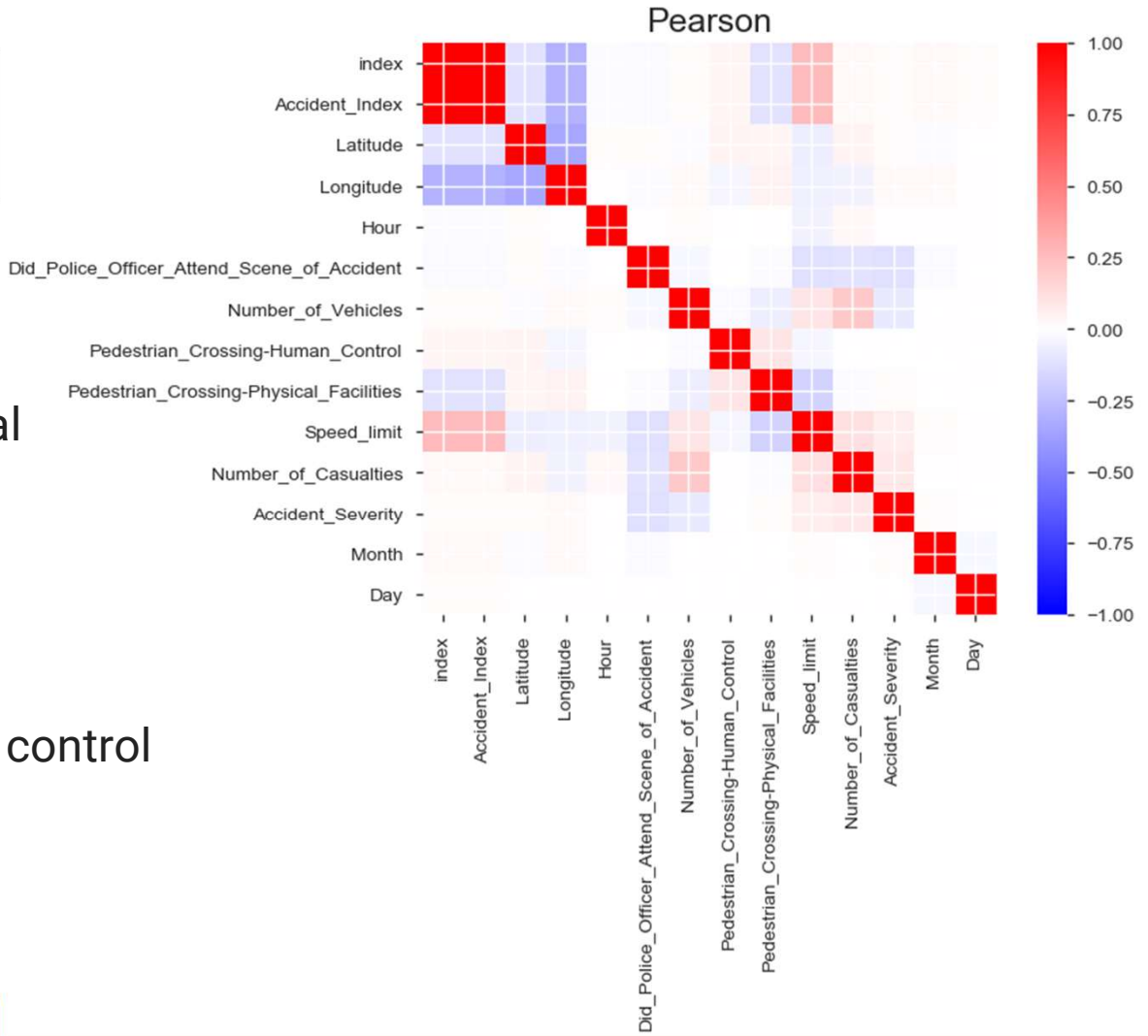
Correlation matrix

Strong Negative Correlation:

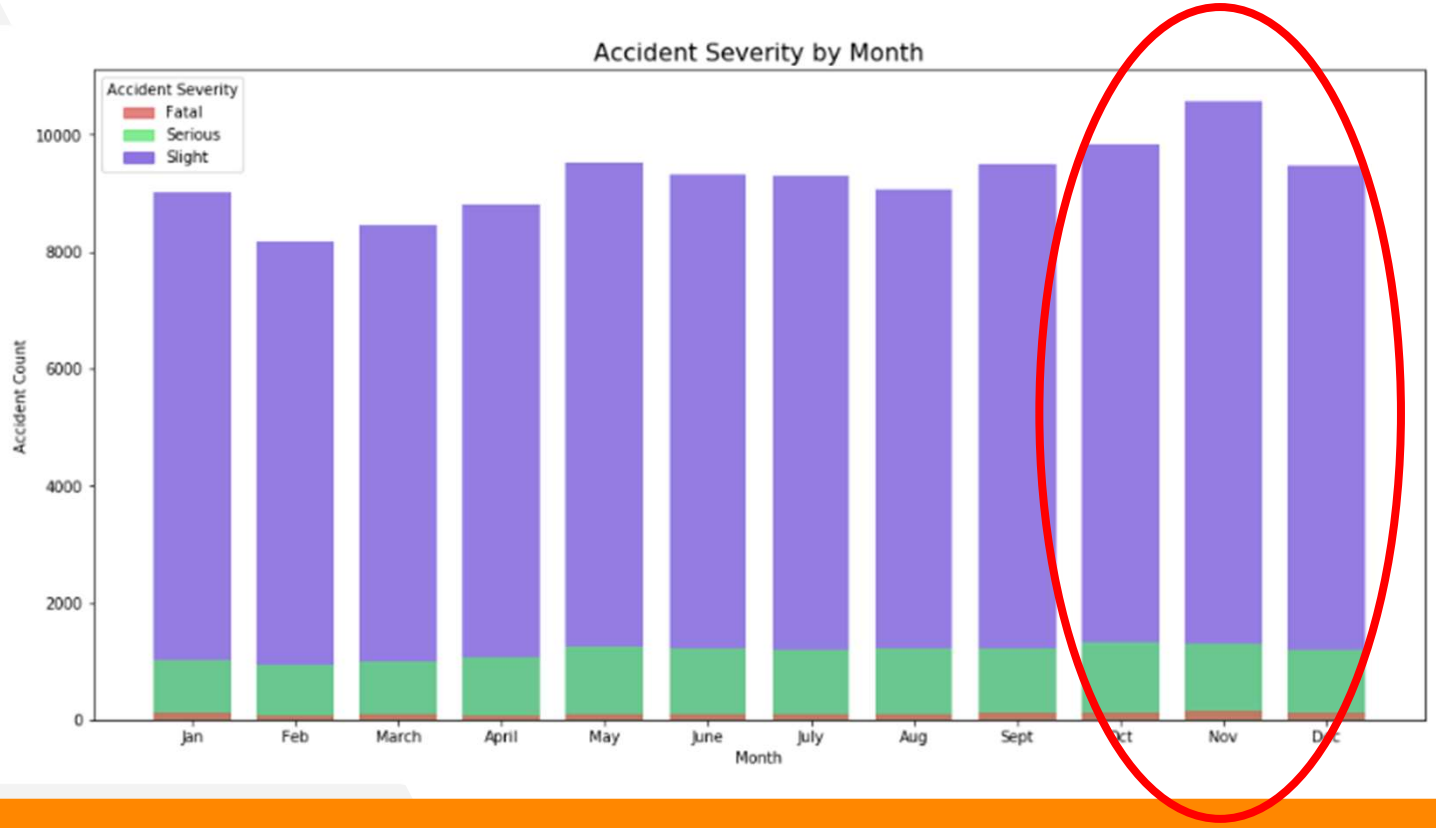
- Pedestrian crossing physical facilities

Strong Positive Correlation:

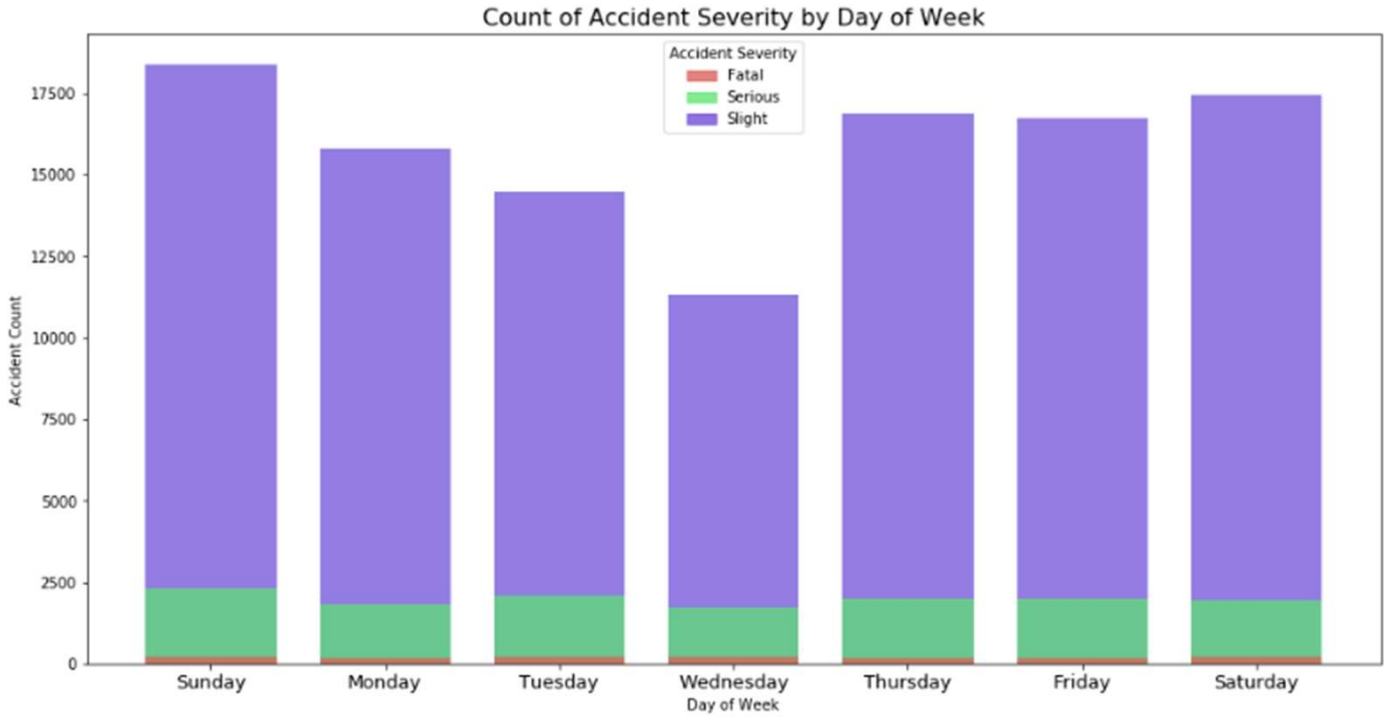
- Speed limit
- Pedestrian crossing human control



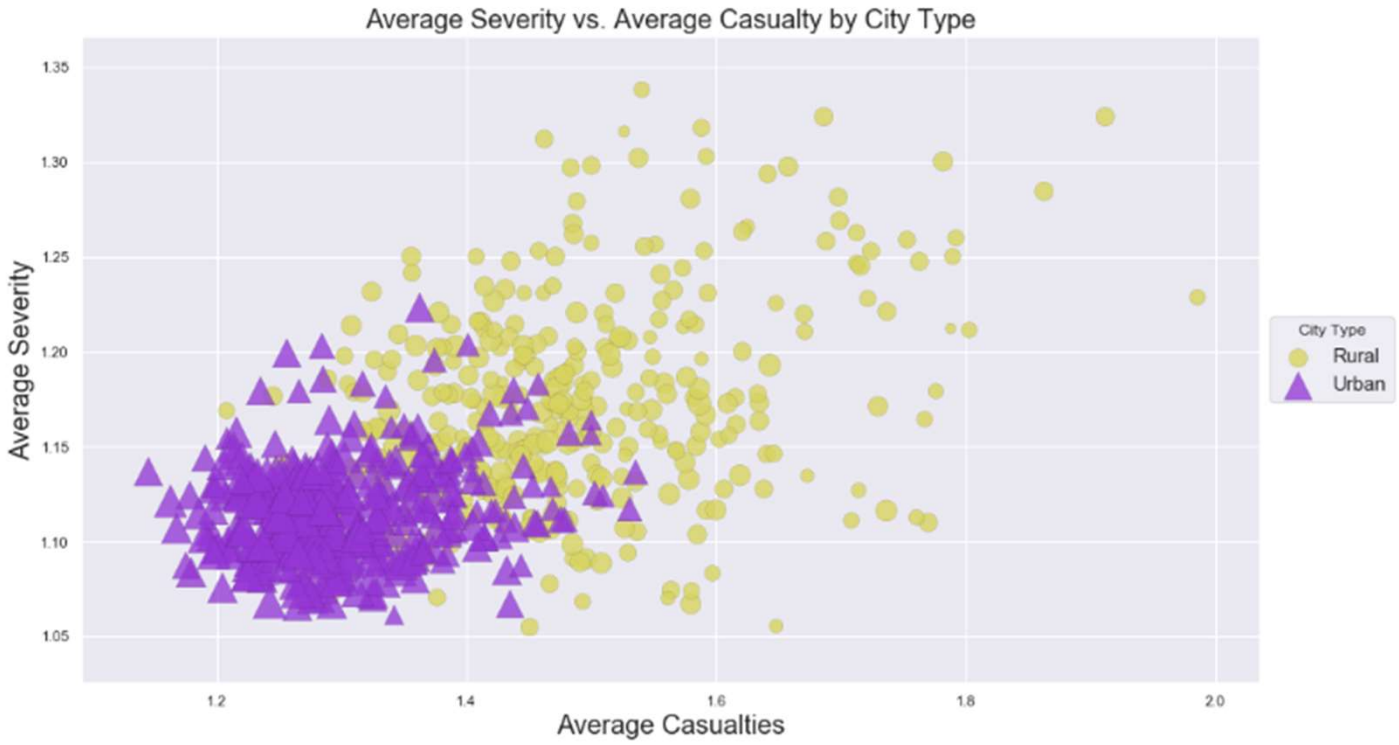
Month


Accident Severity by Month
Accident Severity
Fatal
Serious
Slight
Accident Count
10000
8000
6000
4000
2000
0
Jan
Feb
March
April
May
June
July
Aug
Sept
Nov
Month

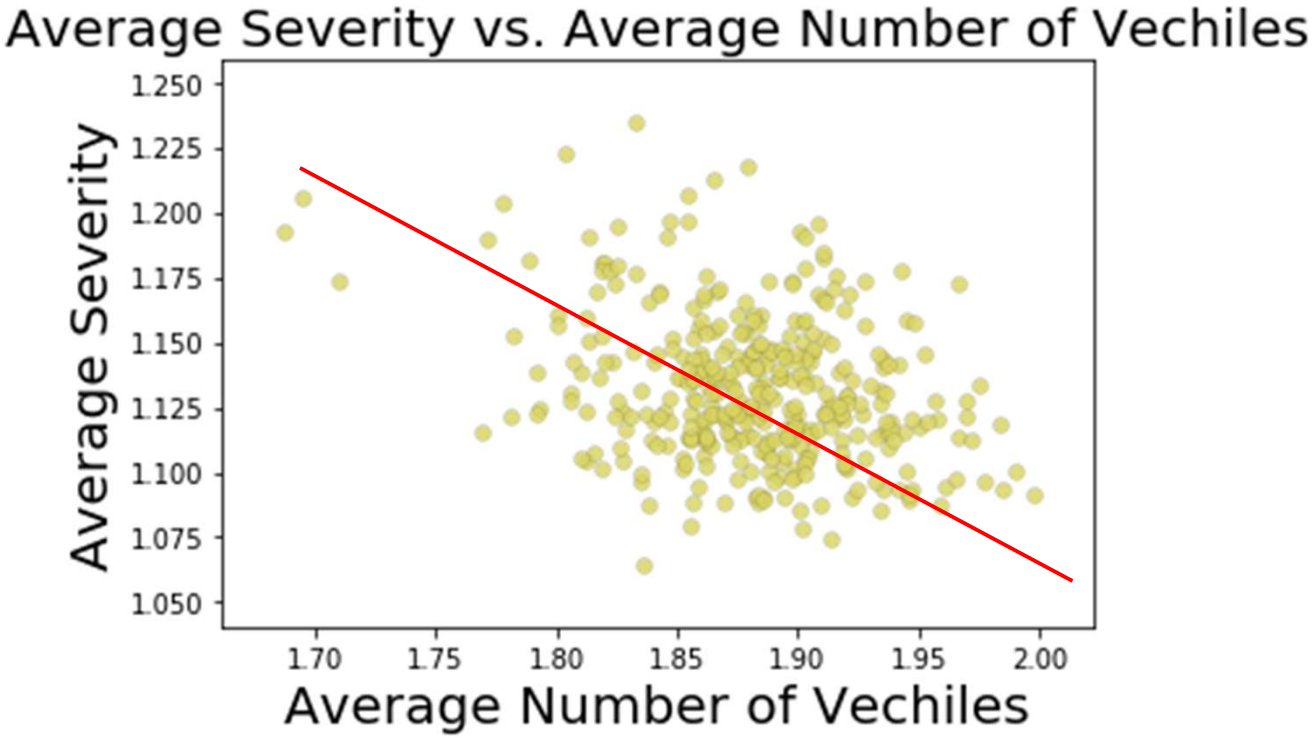
Day of Week


Count of Accident Severity by Day of Week
Accident Severity
Fatal
Serious
Slight
Accident Count
17500
15000
12500
10000
7500
5000
2500
0
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of Week

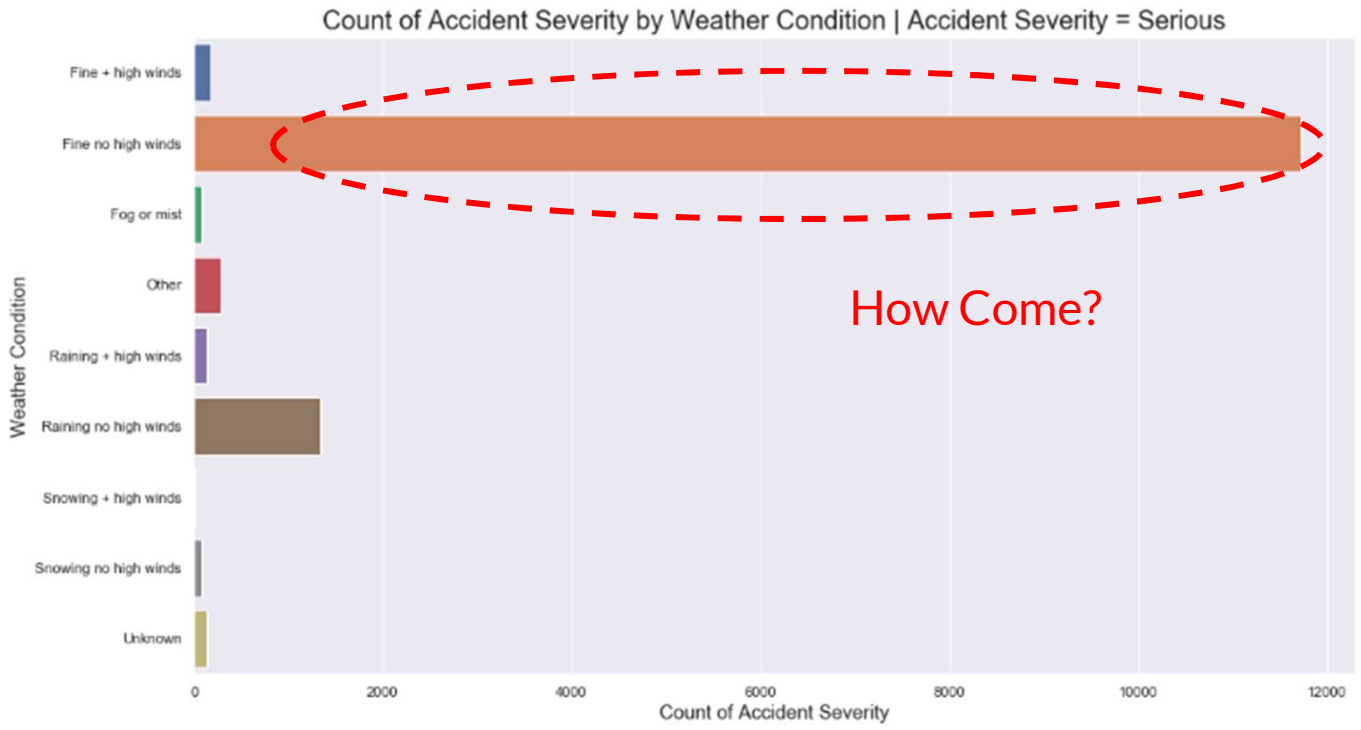
Rural Vs Urban



Number of Vech



Highly Cardinality Categorical Variable



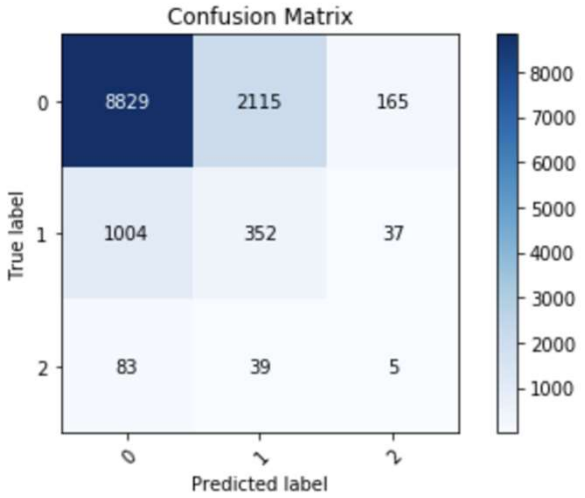
Model Building

Data Preparation

- Remove no contextual variables
  - Longitude, Latitude, Date
- Sklearn data formatting
  - Converting categorical variable to dummy variable
- Holdout Validation (90% training to 10% testing)
  - Confusion Matrix
- Highly Skewed Dataset
  - SMOTE (Synthetic Minority Over Sampling Technique)

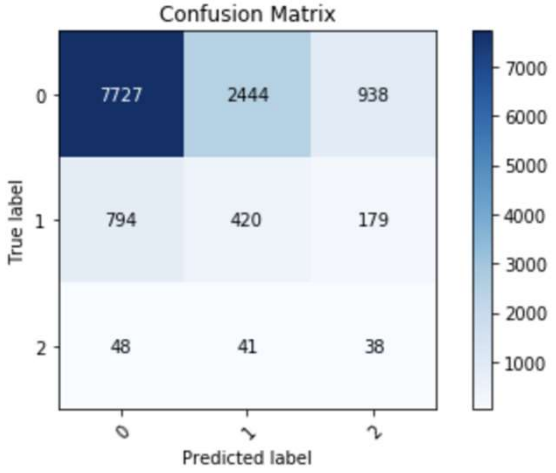
Initial building

Decision Tree 1



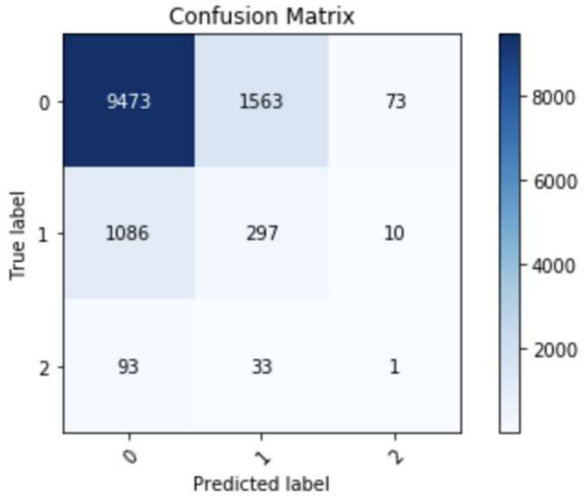
Score: 0.88

Decision Tree 2
(Standardized)



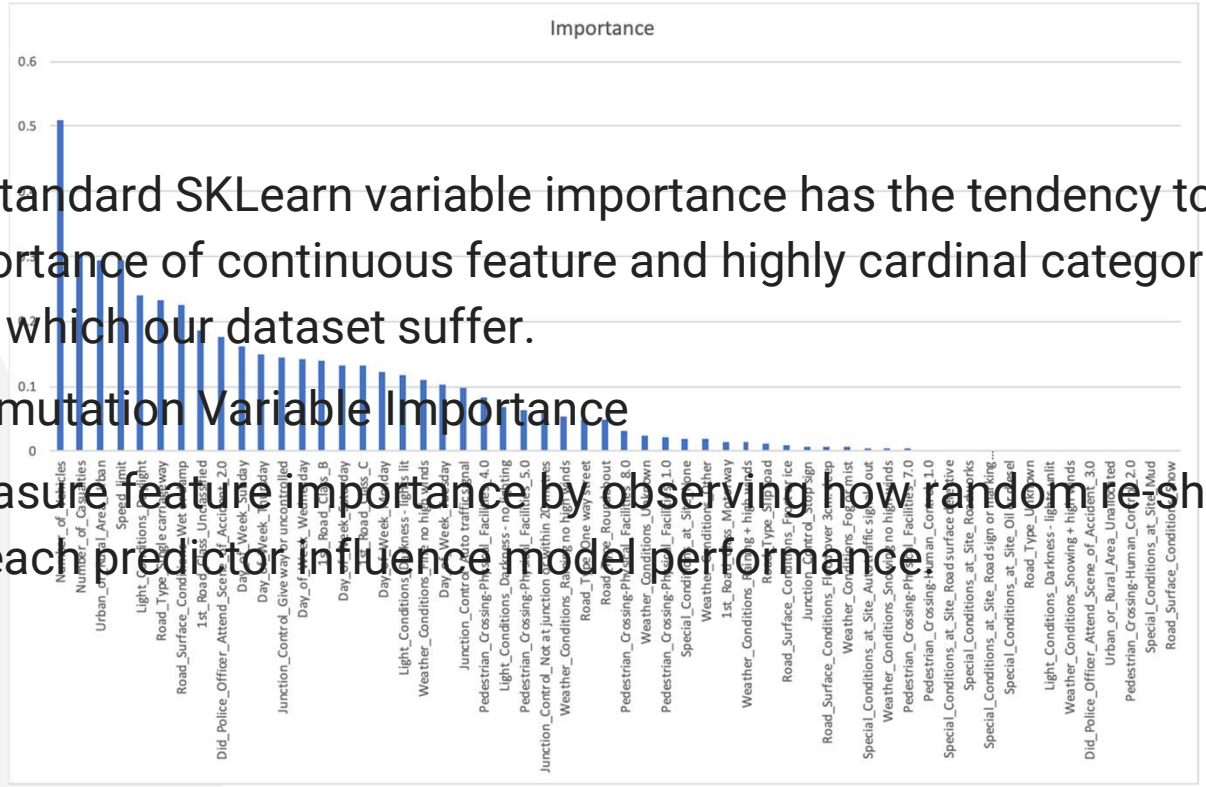
Score: 0.87

Random Forest Model



Score: 0.90

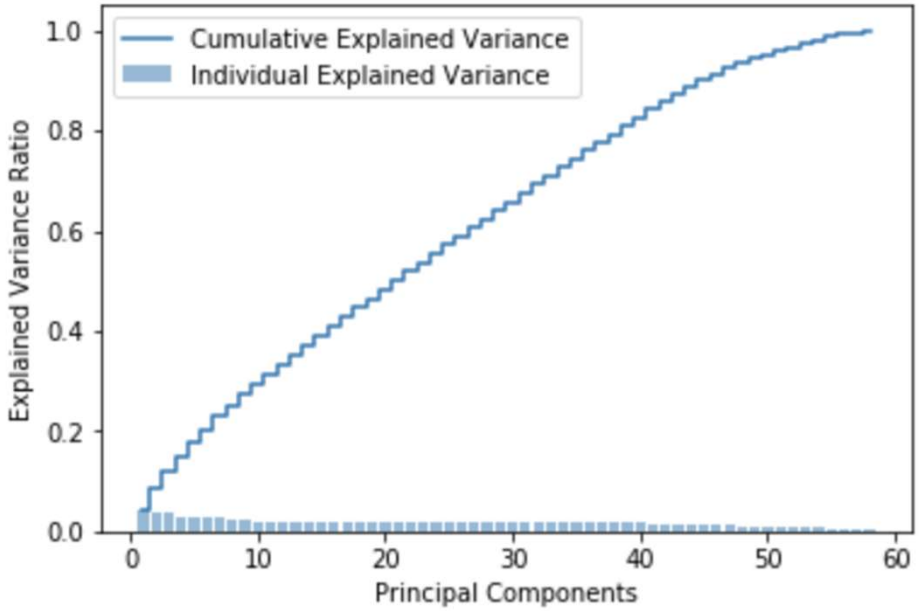
Permutation Variable Importance



As the standard SKLearn variable importance has the tendency to inflate the importance of continuous feature and highly cardinal categorical variable which our dataset suffer.

- Permutation Variable Importance
- Measure feature importance by observing how random re-shuffling of each predictor influence model performance.

PCA



- not useful to conduct dimension reduction

Machine Learning Algorithm tested

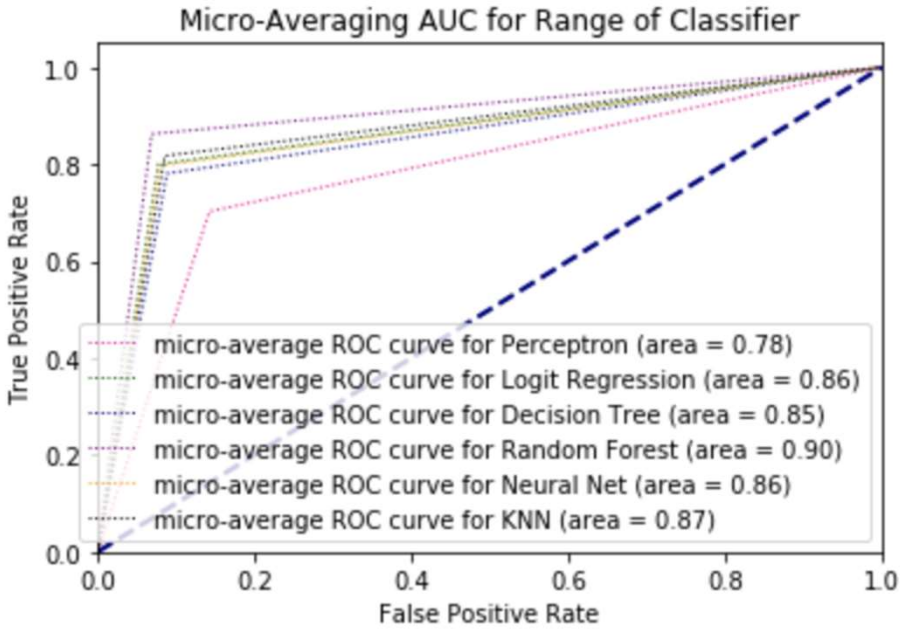
Algorithm	Scores (2.s.f)	Keep
OvA(Perceptron)	0.82	✔
OvA(Logistic Regression)	0.87	✔
Neural Net	0.88	✔
Nearest Neighbor	0.86	✔
Naive Bayes	0.10	×
Linear SVM	N/A	×
RBF SVM	N/A	×

Hyperparameter tuning using GridSearch

```
Algo: Perceptron               and 10 fold score: 0.8216
Algo: LogisticRegression       and 10 fold score: 0.8795
Algo: Decision Tree            and 10 fold score: 0.8796
Algo: Random Forest            and 10 fold score: 0.8656
Algo: Neural Net               and 10 fold score: 0.8796
Algo: Nearest Neighbors        and 10 fold score: 0.8796
```

Evaluation + Ensemble Model

ROC Curve – Micro Averaging



Learning Curve

Overfitting Model

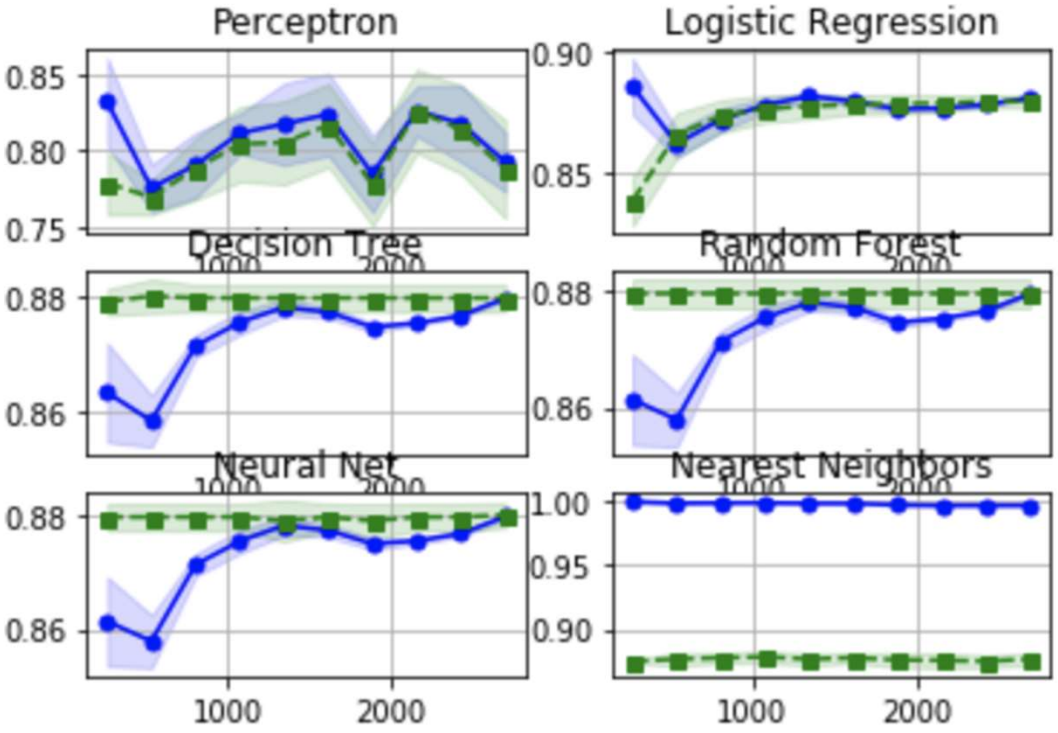
- Nearest Neighbors

Underfitting Model

- Perceptron

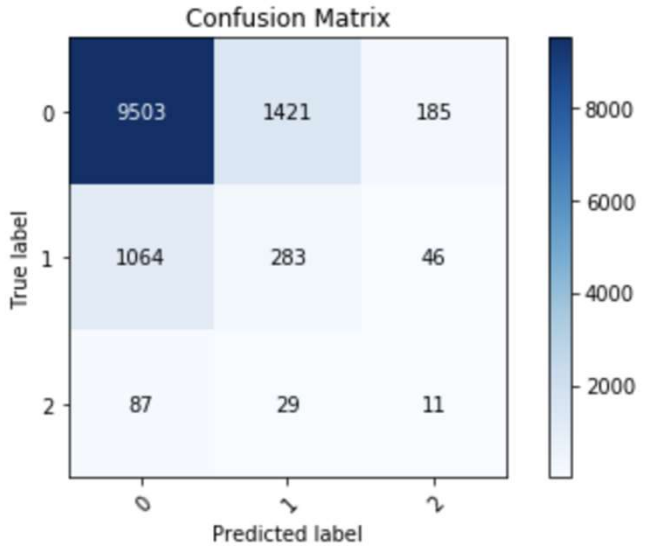
Just Right

- Logistic Regression
- Decision Tree
- Random Forest
- Neural Net



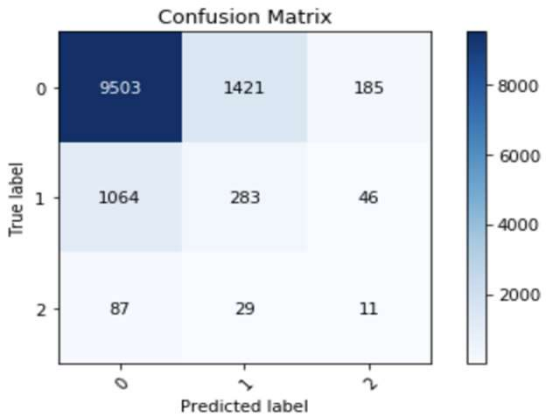
Ensemble Model

- Majority Voting Model
  - Results 88% accuracy

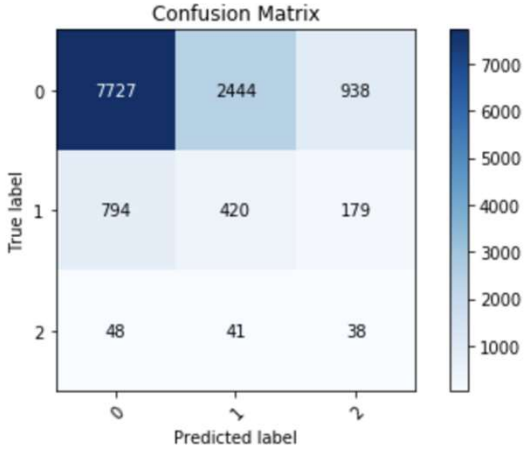


Model recommendation

- Ensemble model for Slight and Decision Tree for Serious/fatal prediction
- Will talk more about this in the improvement section



Ensemble Model



Decision Tree (Standardized)

IMPLICATION & IMPROVEMENT

Implication

Compensation Decision

- Insurance claim — Model: Predict accident severity — Help decide the amount/rate of the compensation

Fraud Claim

- Insurance claim — Model: check whether exaggerated in filing— Fraud investigation

Improvement

- To build a model with weighting of information (+ class weights to help with the skewed data problem)
- The model may have difficulty classifying serious accidents due to highly skewed dataset
- The model requires all completed data, while in reality this is often not the case

- Choose our datasets over a longer time span
- Accident prone zone and cause of fatal accidents may change over time