Data Mining the UK Road Safety Data for Actionable Insights

Georgia Tech

Introduction

Problem Statement

Traffic accidents are a pervasive issue that exists globally, impacting numerous individuals and disrupting daily life. Crash injuries are estimated to be the eighth leading cause of death globally for all age groups and the leading cause of death for children and young people 5–29 years of age. Despite its prevalence, finding sustainable solutions remains challenging.

Goal

The aim is to leverage the comprehensive and robust UK road accident and vehicle data dataset on traffic crashes and make predictions using machine learning algorithms. Our analysis will encompass descriptive investigations to gain deeper insights into the patterns of traffic crashes. The outcome of our analysis will also provide valuable guidance for policy formulation, intervention strategies and results will be advantageous for various stakeholders, including the Department of Transportation, local traffic agencies, car manufacturers, and traffic safety consultants.

Project Scope

- Our project will use relevant data, preprocess the accident and vehicle data to ensure suitability for machine learning models using various algorithms.
- Our predictions involve predicting the severity of a crash incident at a specific location based on various predictor variables.
- The aim is to provide valuable insights to the stakeholders and aid in the formulation of targeted policies, ultimately enhancing road safety and mitigating accident severity.
- Our analysis will help allocate infrastructure resources effectively for road maintenance, traffic enforcement, public awareness campaigns, infrastructure improvements and safety campaigns across all state/federal roads.
- Understanding crash patterns and causes may inspire innovation in vehicle technology, such as advanced driver assistance systems (ADAS) and autonomous driving technology, facilitating the enhancement of safety features, risk mitigation, and injury severity reduction.

Source of the Dataset

Source of the Dataset

We are using the "UK Road Safety: Traffic Accidents and Vehicles" Dataset from the Department of Transport open data website.

There are 2 files -

- **Accident_Information.csv:** 705MB sized dataset containing 34 columns of accident data from 2005 till 2017.
- **Vehicle_Information.csv:** 644MB sized dataset containing 24 columns of data about car vehicles between 2004 and 2016.

Here is the link to the dataset - https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

Dataset Structure

From an initial analysis, we were able to determine that the below factors in the data can help us build a model to determine the contribution of each factor to the probability of the severity of an accident event.

- **Environmental Factors:** Weather, light, and road surface conditions that contribute to an increase in traffic accidents
- **Infrastructure:** Specific junctions, carriageway details, police oversight, speed limit, pedestrian crossing that are more prone to accidents.
- **Geographical Factors:** Urban Vs Rural, In Scotland - where do the accidents happen mostly and reason?
- **Personal Factors:** Age of driver, Gender
- **Vehicle Factors:** Age of Vehicle, Engine Capacity, Make, Model, Vehicle Type, Model Year
- **Driving Factors:** Vehicle Maneuver

Research Questions

- What infrastructure factors are significant for a particular road segment so that DOT/local agencies could take significant steps to reduce the severity of the crash incidents?
- What driver/car information is significant so that the car manufacturer can take significant steps to reduce crash severity?
- Which infrastructure elements, such as junction design, carriageway details, and police oversight, are associated with higher accident rates?
- How do specific environmental factors such as weather, light conditions, and road surfaces contribute to the occurrence of traffic accidents?
- Could we recommend DOT or other local agencies some highway safety improvements such as highway lighting, 2-way Left turn Lane, signal control, speed limit reduction, etc. based on the exploratory data analysis?
- Could our classification model be used as a performance analytics to check the safety of our existing roadway network?
- Could a classification model based on environmental, infrastructure, geographical, personal, vehicle and driving factors help a state Department of Transportation, or a local agency allocate funds judiciously on a road safety project based on the actual need backed by data?
- To what extent can future accident predictions be precise when employing data mining methodologies with historical accident data as a basis?

Exploratory Data Analysis (EDA)

Information of the Merged Dataset

As part of EDA below are the steps we have performed:

- Accident and the Vehicle data set were merged using the Accident_Index column.
- Next we identified 21 columns based on the factors listed, that we will use for further analysis.
- Our next step was to identify and then remove the Null values from the resulting dataset. The row count is now 3,93,885 and memory usage: 66.1+ MB.
- After the removal of Null values, all the String variables are now converted into Categorical variables. This will help us in building models.

```
df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1556385 entries, 2 to 2058407
Data columns (total 21 columns):
 #   Column                   Non-Null Count    Dtype
---  ------                   --------------    -----
 0   Accident_Index           1556385 non-null  object
 1   Accident_Severity        1556385 non-null  object
 2   Date                     1556385 non-null  object
 3   Day_of_Week              1556385 non-null  object
 4   Junction_Control         1556385 non-null  object
 5   Junction_Detail          1556385 non-null  object
 6   Road_Surface_Conditions  1556385 non-null  object
 7   Road_Type                1556385 non-null  object
 8   Speed_limit              1556385 non-null  float64
 9   Time                     1556385 non-null  object
 10  Urban_or_Rural_Area      1556385 non-null  object
 11  Weather_Conditions       1556385 non-null  object
 12  Year_x                   1556385 non-null  int64
 13  Age_Band_of_Driver       1556385 non-null  object
 14  Age_of_Vehicle           1556385 non-null  float64
 15  Engine_Capacity_.CC.     1556385 non-null  float64
 16  make                     1556385 non-null  object
 17  model                    1556385 non-null  object
 18  Propulsion_Code          1556385 non-null  object
 19  Sex_of_Driver            1556385 non-null  object
 20  Vehicle_Manoeuvre        1556385 non-null  object
dtypes: float64(3), int64(1), object(17)
memory usage: 261.2+ MB
```

Georgia Tech

Information of the Merged Dataset (contd.)

Some of the subfields in our independent categorical variables are shared below. State Engineers or transportation professionals who are interested to study the safety of a road segment based on the model we developed should be cognizant of the various sub-fields inside each categorical variable we chose.

Vehicle_Manoeuvre	
Going ahead other	175217
Turning right	41702
Waiting to go - held up	32355
Slowing or stopping	31119
Going ahead right-hand bend	16702
Parked	15770
Going ahead left-hand bend	14488
Moving off	13981
Turning left	12700
Waiting to turn right	7735
Overtaking moving vehicle - offside	7387
Reversing	5498
Overtaking static vehicle - offside	5122
U-turn	3509
Changing lane to right	3073
Changing lane to left	2881
Waiting to turn left	2680

Age_Band_of_Driver	
26 - 35	83078
36 - 45	81965
46 - 55	56570
21 - 25	44579
16 - 20	36214
56 - 65	34525
Data missing or out of range	31282
66 - 75	15997
Over 75	9539
11 - 15	132
6 - 10	4

Road_Surface_Conditions	
Dry	265638
Wet or damp	113483
Frost or ice	10269
Snow	3548
Data missing or out of range	490
Flood over 3cm. deep	457

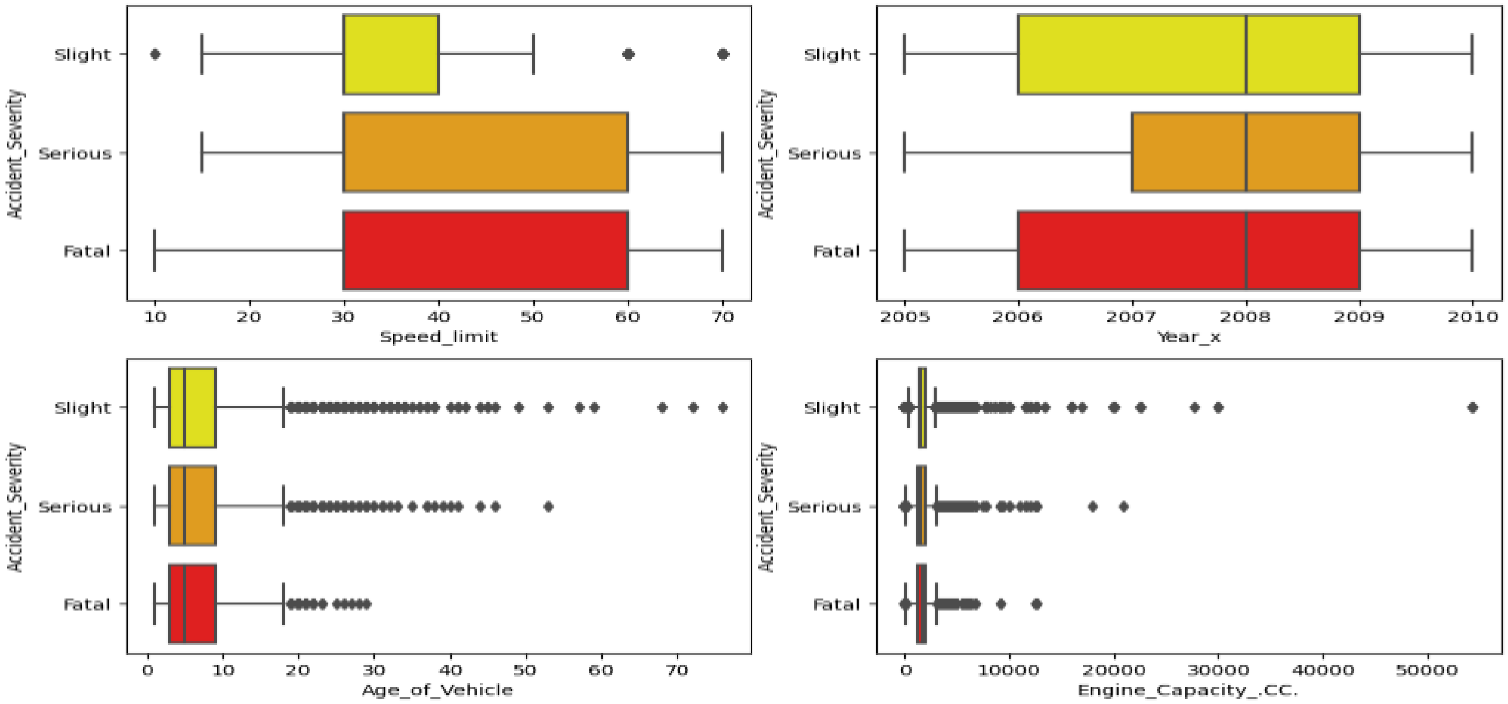
Road_Type	
Single carriageway	285607
Dual carriageway	67871
Roundabout	26864
One way street	6960
Slip road	4435
Unknown	2148

Weather_Conditions	
Fine no high winds	311576
Raining no high winds	49060
Other	10805
Unknown	6308
Raining + high winds	5055
Fine + high winds	4216
Snowing no high winds	4093
Fog or mist	2133
Snowing + high winds	543
Data missing or out of range	96

Distribution of Key factors against Accident Severity

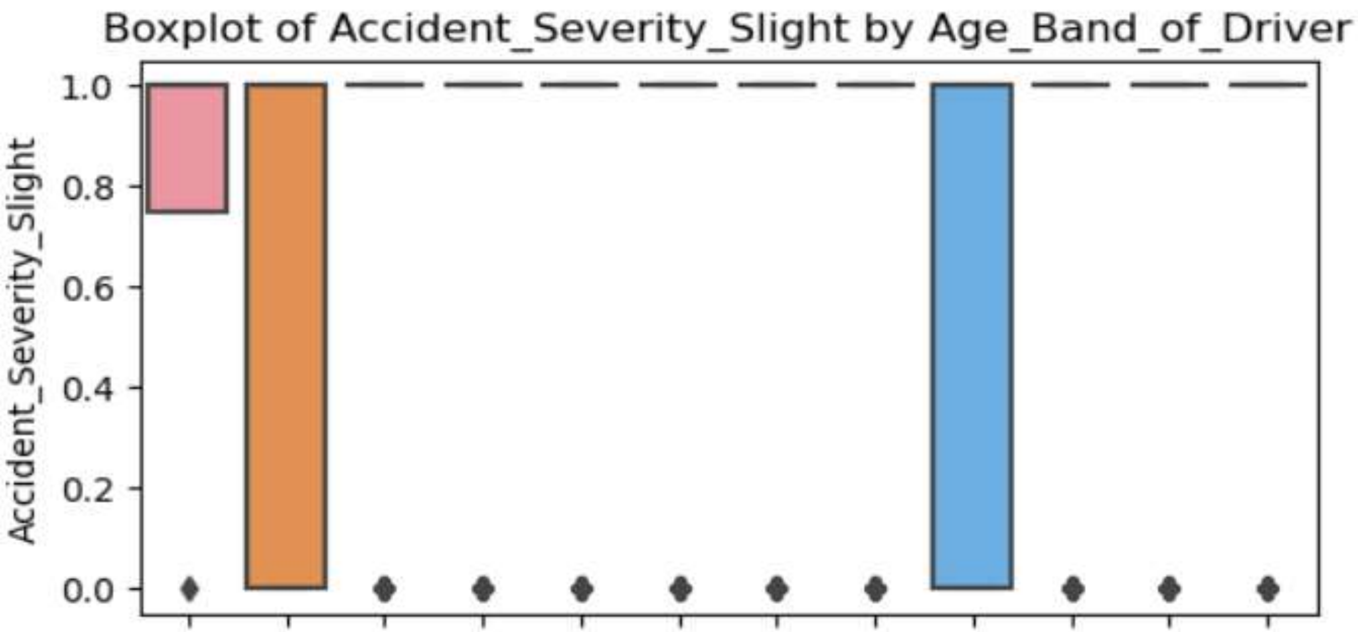
We are now using the Box plots to find the outliers. Below are images of box plots built for Accident Severity against Speed Limit, Age of Vehicle, Year and Engine capacity.

- ❑ We can see that the severity of accident increased with an increase of the speed limit
- ❑ The severity of the accident tends to increase for older vehicles. Most likely because of the advances in the car safety features over the years.



Georgia Tech

Distribution of Key factors against Accident Severity (Cont.)



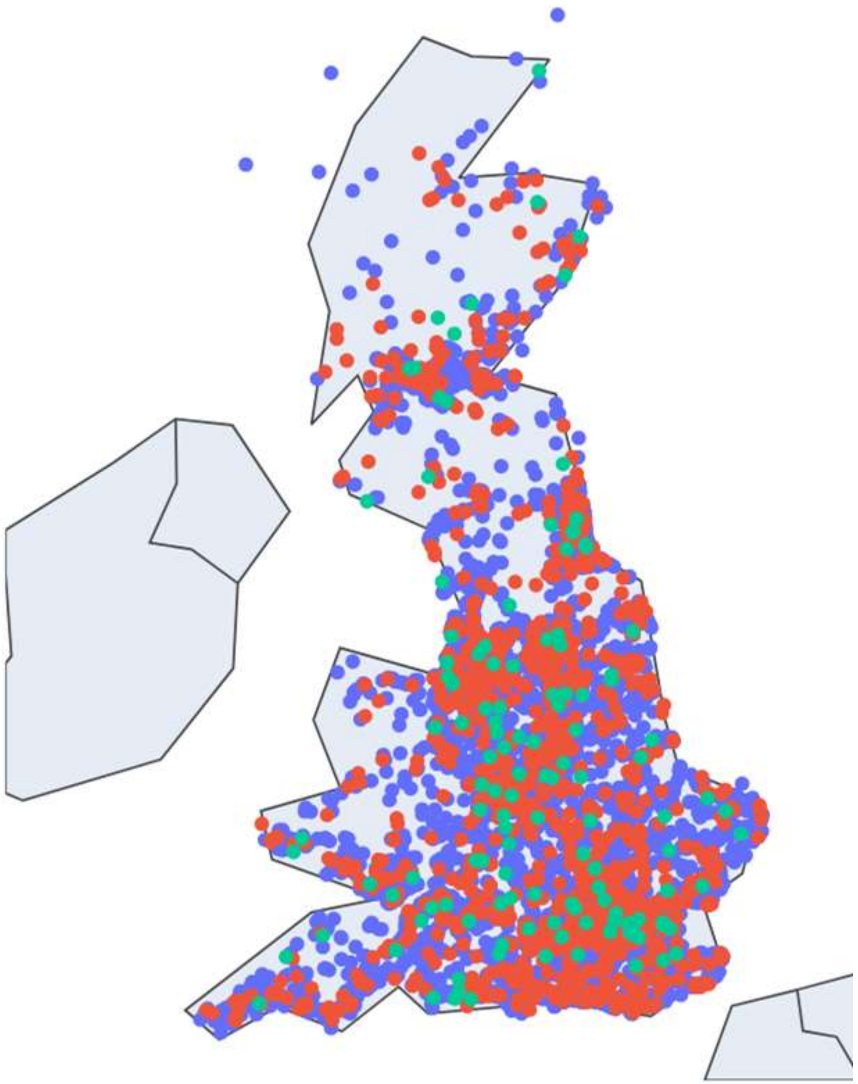
Apparently the accidents are distributed more for younger age groups and older age groups

Correlation Matrix

	Junction_Control	Junction_Detail	Road_Surface_Conditions	Road_Type	Urban_or_Rural_Area	Weather_Conditions	Age_Band_of_Driver	Sex_of_Driver	Vehicle_Manoeuvre	Accident_Severity_Slight	Speed_limit	Age_of_Vehicle	Engine_Capacity_.CC.
Junction_Control	1.00	0.38	0.00	0.15	-0.02	-0.01	0.01	-0.02	0.05	0.00	-0.04	0.00	-0.03
Junction_Detail	0.38	1.00	-0.03	0.10	0.07	-0.01	0.01	-0.01	0.14	0.02	-0.10	0.01	-0.02
Road_Surface_Conditions	0.00	-0.03	1.00	0.00	-0.08	0.51	-0.06	-0.02	-0.04	0.00	0.08	0.01	-0.01
Road_Type	0.15	0.10	0.00	1.00	0.10	-0.01	0.03	0.00	0.05	-0.03	-0.36	0.05	-0.06
Urban_or_Rural_Area	-0.02	0.07	-0.08	0.10	1.00	-0.01	0.04	0.03	0.12	0.09	-0.67	0.01	-0.04
Weather_Conditions	-0.01	-0.01	0.51	-0.01	-0.01	1.00	-0.02	-0.01	-0.02	0.03	0.02	0.01	-0.01
Age_Band_of_Driver	0.01	0.01	-0.06	0.03	0.04	-0.02	1.00	0.16	0.04	-0.01	-0.06	-0.01	0.12
Sex_of_Driver	-0.02	-0.01	-0.02	0.00	0.03	-0.01	0.16	1.00	-0.05	-0.05	-0.02	0.03	0.10
Vehicle_Manoeuvre	0.05	0.14	-0.04	0.05	0.12	-0.02	0.04	-0.05	1.00	0.08	-0.14	-0.03	0.02
Accident_Severity_Slight	0.00	0.02	0.00	-0.03	0.09	0.03	-0.01	-0.05	0.08	1.00	-0.08	-0.01	0.02
Speed_limit	-0.04	-0.10	0.08	-0.36	-0.67	0.02	-0.06	-0.02	-0.14	-0.08	1.00	-0.03	0.06
Age_of_Vehicle	0.00	0.01	0.01	0.05	0.01	0.01	-0.01	0.03	-0.03	-0.01	-0.03	1.00	-0.01
Engine_Capacity_.CC.	-0.03	-0.02	-0.01	-0.06	-0.04	-0.01	0.12	0.10	0.02	0.02	0.06	-0.01	1.00

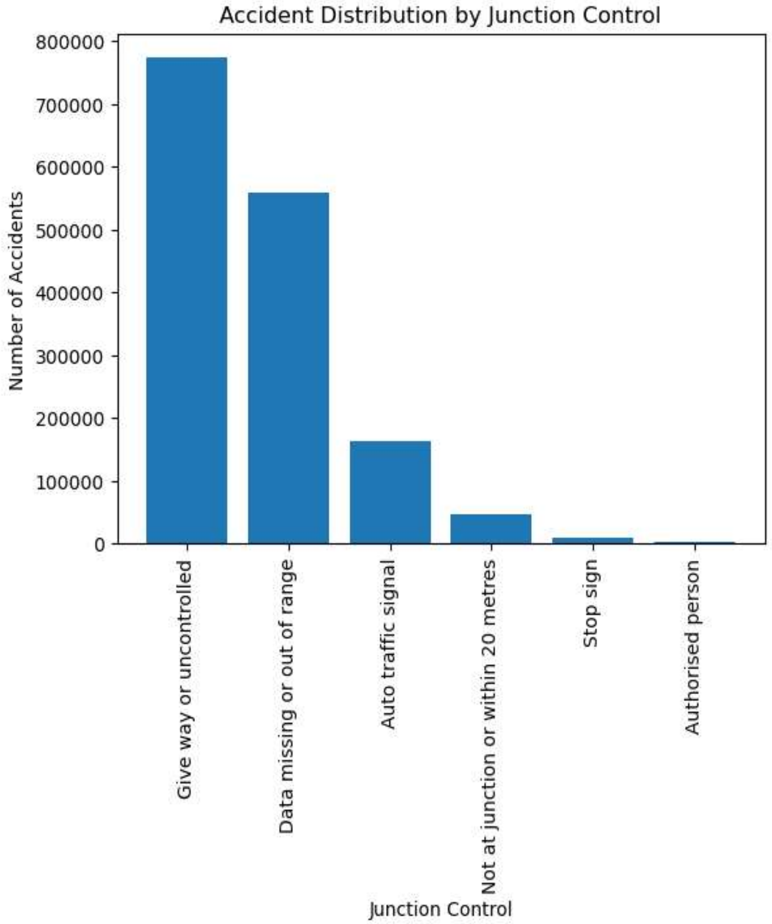
- ❑ There seems to be a higher correlation between Weather and Road surface conditions, Junction Detail and Control which makes logical sense.
- ❑ There seems to be a negative correlation between The Urban/Rural area and the Speed limit and the road type which also is valid.
- ❑ Most other factors have little correlation making the dataset good for the model building.

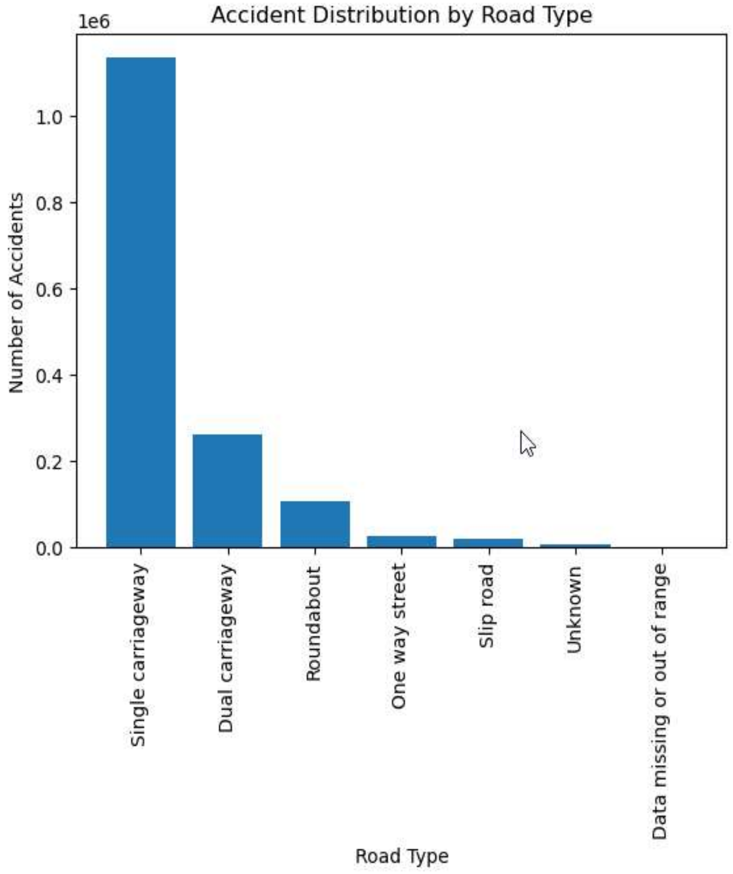
Geographical Accident Distribution

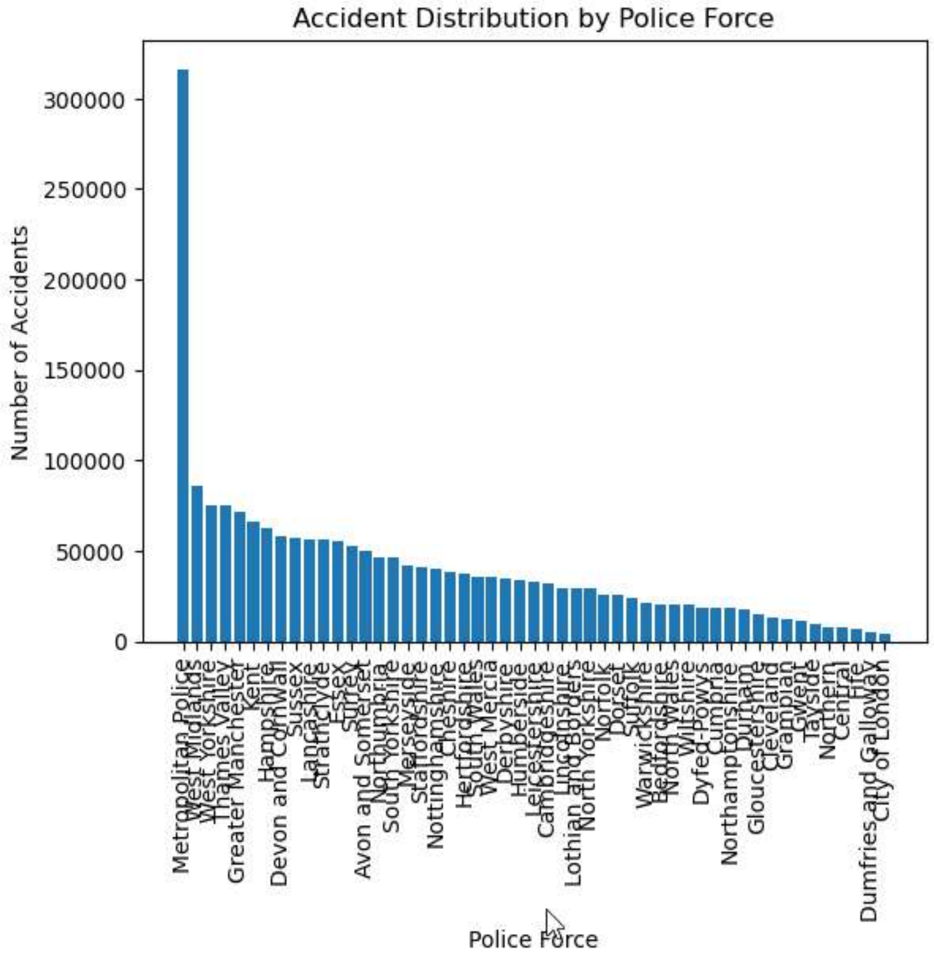


- Blue indicates slight severity, Red indicates serious and Green dots indicate fatal accidents.
- It shows that the most accidents are concentrated in high population density areas highlighting the importance of distribution of infrastructure funds based on population density

Accident Analysis by Infrastructure Factors



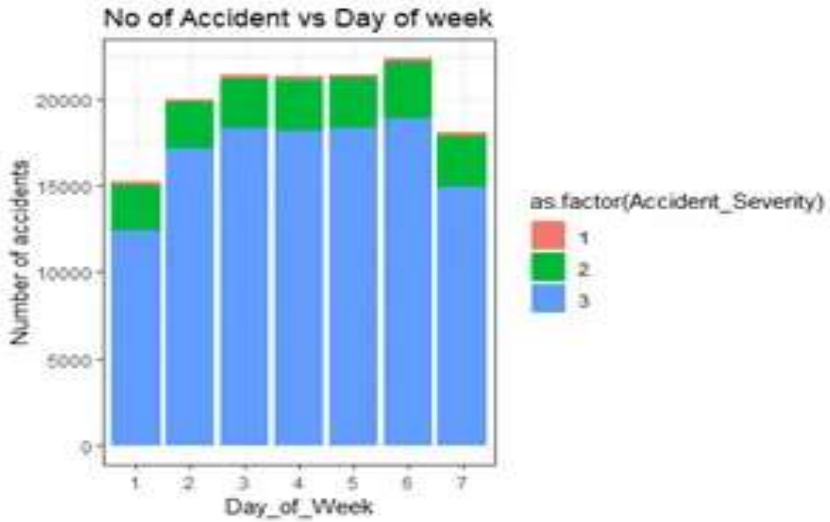


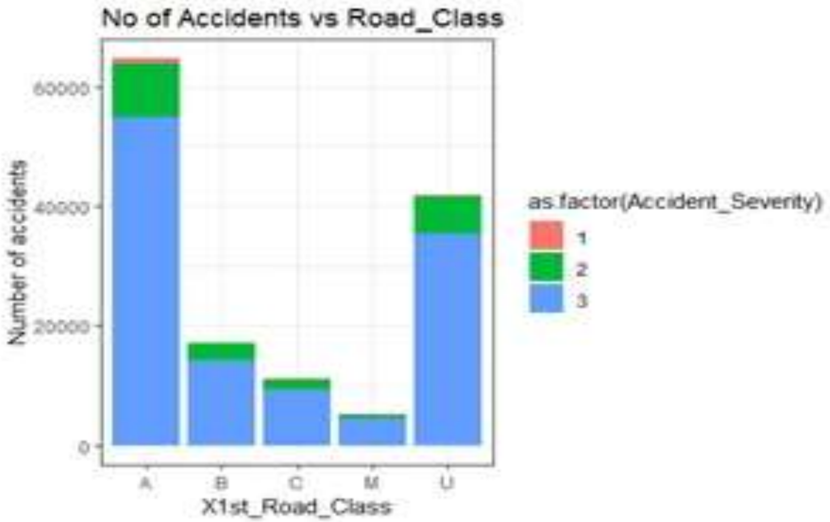


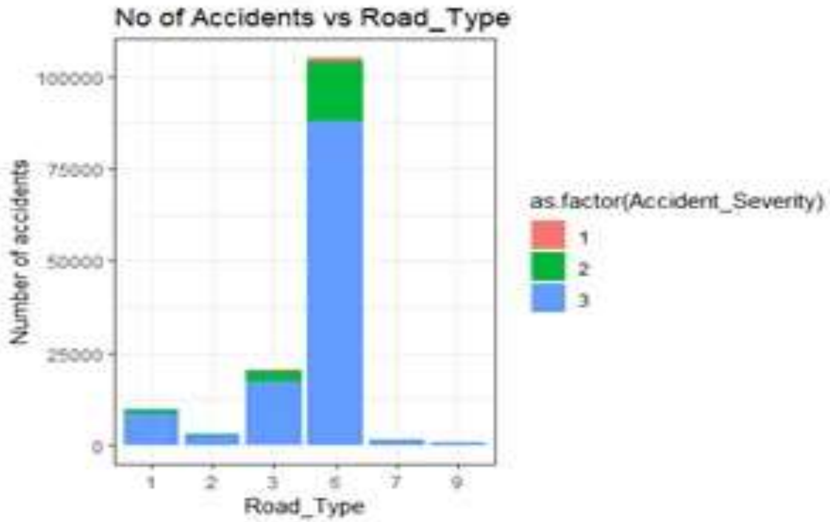
- It looks like most accidents are at uncontrolled junctions indicating the need for the infrastructural needs to decrease the accident numbers.
- Also the higher accidents on the single carriageway indicates the same.
- When it comes to Police Force, it seems most accidents were happening in the metropolitan locations

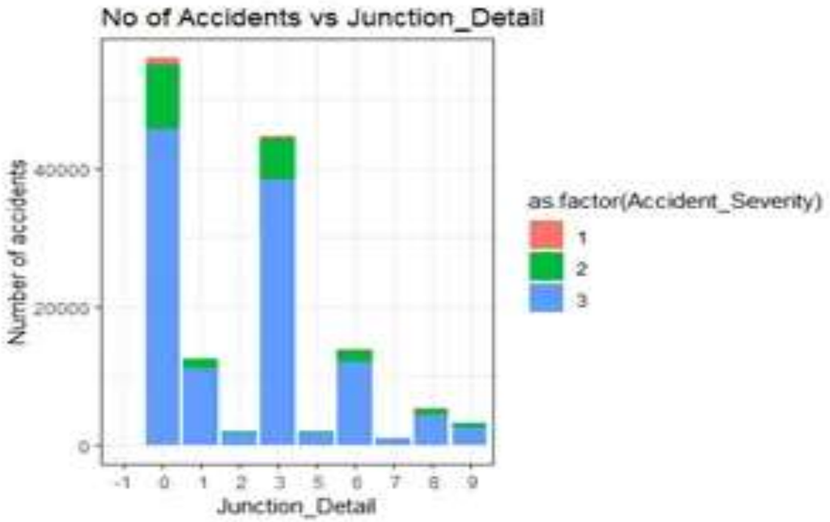
Georgia Tech

Other Insights









- ❑ Additional inferences were drawn from similar research (3) on this dataset.
- ❑ It seems more accidents tend to happen over the weekdays compared to weekends.
- ❑ Maximum accidents happened on Fridays, on Road class A, Road type 6, and no junction.
- ❑ It was inferred that the weather conditions do not play a major role as most accidents seem to have happened on clear days.

Georgia Tech

Models/Methods

Linear Discriminant Analysis(LDA)

LDA Model:

This is one of the statistical methods used for classification techniques. By using observations that have the equal covariance matrices and that are distributed normally, we will classify the new observations into already defined categories in this supervised learning technique.

Approach:

- After EDA, the dataset we obtained was partitioned and we allocated 20% of data for testing and remaining 80% as training data for performing the analysis.
- LinearDiscriminantAnalysis() method from the linear discriminant analysis python package has been used for this.
- Cross validation was performed with n=100 to measure the robustness of the model.

Results:

LDA mean testing error 0.126 and with cross validation it's 0.127. This model's recall value is 1.0, F1 score 0.932 with accuracy & precision values of 0.873

Quadratic Discriminant Analysis(QDA)

QDA Model:

This method is similar to LDA in terms of supervised learning techniques for classification but the only difference of QDA is it will not assume all observations will have same covariance matrices and hence will be used for different covariance structures scenarios

Approach:

- ❑ After EDA, the dataset we obtained was partitioned and we allocated 20% of data for testing and remaining 80% as training data for performing the analysis.
- ❑ QuadraticDiscriminantAnalysis() method from the Quadratic discriminant analysis python package has been used for this.
- ❑ Cross validation was performed with n=100 to measure the robustness of the model.

Results:

QDA mean testing error 0.129 and with cross validation it’s 0.13. This model’s recall value is 0.995, F1 score 0.93 with 0.87 accuracy & precision values of 0.874.

Naive Bayes Model

Naive Bayes Model:

One of the most popular classification methods which uses Bayes theorem for classifying the documents. This method treats all variables as independent i.e., one variable cannot impact the other variable classification which is different behavior from the above listed lda and qda methods

Approach:

- After EDA, the dataset we obtained was partitioned and we allocated 20% of data for testing and remaining 80% as training data for performing the analysis.
- GaussianNB () method from the GaussianNB python package has been used for this
- Cross validation was performed with n=100 to measure the robustness of the model.

Results:

Naive Bayes Model’s mean testing error 0.129 and with cross validation it’s 0.13. This model’s recall value is 0.995, F1 score 0.93 with 0.87 accuracy and precision values of 0.874.

Georgia Tech

Logistic Regression Model

Logistic Regression Model:

This classification technique is primarily used in predicting the input belongs to a certain category or not to prevent overfitting of data. Usually Generalized linear model (glm) or Multinomial Logistic Regression (multinorm) functions will be used for logistic regression

Approach:

- After EDA, the dataset we obtained was partitioned and we allocated 20% of data for testing and remaining 80% as training data for performing the analysis.
- LogisticRegression() method from the LogisticRegression python package has been used for this.
- Cross validation was performed with n=100 to measure the robustness of the model.

Results:

Logistic Regression Model's mean testing error 0.126 and with cross validation it's 0.127. This model's recall value is 1.0, F1 score 0.932 with 0.873 accuracy and precision.

KNN Classification Model

KNN Classification Model:

A common method that can be used for classification as well as regression and K value has to be carefully selected to have correct predictions. Smaller value of ‘K’ results in overfitting and a larger value of ‘K’ results in underfitting of data, so multiple trial and error for K values has to be performed to find a proper K value to have better mean and variance.

Approach:

- After EDA, the dataset we obtained was partitioned and we allocated 20% of data for testing and remaining 80% as training data for performing the analysis.
- KNeighborsClassifier() method from the KNeighborsClassifier python package has been used for this. We considered k values from 1 to 20 out of which 13 performed the best.
- Cross validation was performed with n=100 to measure the robustness of the model.

Results:

KNN for n=13 mean testing error 0.129 and with cross validation it’s 0.128. This model’s recall value is 0.989, F1 score 0.929 with 0.869 accuracy and precision of 0.876.

Random Forest Classification Model

Random Forest Classification Model:

Random Forest is a machine learning algorithm that can be used for classification as well as regression cases. This machine learning method operates by constructing the multitude of decision trees during training and outputting the classification or regression of individual trees. Major advantages using this method are reduce overfitting and more robust approach

Approach:

- ❑ After EDA, the dataset we obtained was partitioned and we allocated 20% of data for testing and remaining 80% as training data for performing the analysis.
- ❑ Random Forest Classifier () method from the Random Forest Classifier python package has been used for this.
- ❑ Hypertuning is done to highlight the best performance of the model.

Results:

- ❑ Random Forest mean testing error without hypertuning is 0.145 and with hypertuning it's 0.127
- ❑ The accuracy and precision values are 0.855 and 0.877 respectively, recall is 0.97 and F1 is 0.921 without hypertuning.
- ❑ Random Forest Best Parameters: {'n_estimators': 50, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_depth': 10}
- ❑ The accuracy and precision values are 0.873 and 0.874 respectively, recall is 1 and F1 is 0.932 with hypertuning.

Boosting Model

Boosting Model:

Boosting is a powerful machine learning algorithm used for accuracy improvement in case of classification and regression. It works on principle by combining the multiple decision trees predictions to create a stronger one where it iteratively train new models, with each subsequent model focusing on the instances that the previous model had issues in classifying correctly

Approach:

- ❑ After EDA, the dataset we obtained was partitioned and we allocated 20% of data for testing and remaining 80% as training data for performing the analysis.
- ❑ Gradient Boosting Classifier () method from the Gradient Boosting Classifier python package has been used for this
- ❑ Hypertuning is done to highlight the best performance of the model.

Results:

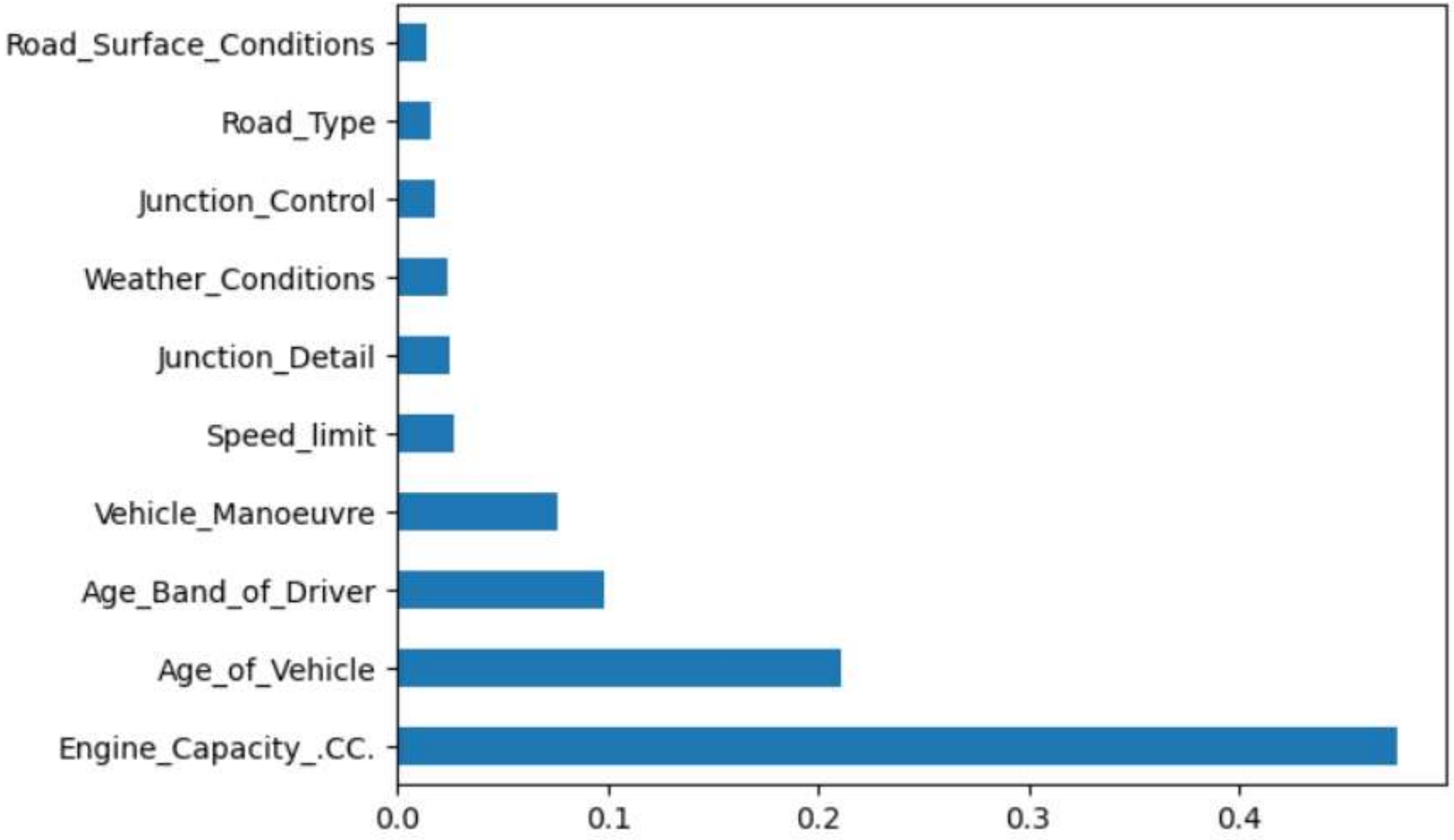
- ❑ Boosting mean testing error without hypertuning is 0.126 and with hypertuning is 0.126.
- ❑ The accuracy and precision values are 0.874 and 0.875 respectively, recall is 0.998 and F1 is 0.932 without hypertuning.
- ❑ Boosting Best Parameters: {'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 5, 'learning_rate': 0.05}
- ❑ The accuracy and precision values are 0.874 and 0.875 respectively, recall is 0.998 and F1 is 0.932 with hypertuning.

Georgia Tech

Feature Importance

Method:

- ❑ It is of value to know which independent variables are of most importance to the severity of accident.
- ❑ We used the ExtraTreesClassifier from sklearn to identify the same
- ❑ We picked the top 10 and it seems the capacity of engine is the most significant.
- ❑ This indicates the importance of the intrinsic qualities of the vehicles compared to everything else.



Results

Comparison of the Model Results and Findings

Models Result without Cross Validation

Models	LDA	QDA	Naive Bayes	Logistic Regression	KNN Classification n=13
Mean Testing Error	0.127	0.13	0.13	0.127	0.129
Accuracy	0.873	0.87	0.87	0.873	0.871
Recall	1.0	0.995	0.995	1.0	0.994
F1 score	0.932	0.93	0.93	0.932	0.931
Precision	0.873	0.874	0.874	0.873	0.875

Models Result with Cross Validation

Models	LDA	QDA	Naive Bayes	Logistic Regression	KNN Classification
Mean Testing Error with CV	0.126	0.129	0.129	0.126	0.128

Georgia Tech

Comparison of the Model Results and Findings

Models Result without Hypertuning

Models	Random Forest	Boosting
Mean Testing Error	0.145	0.126
Accuracy	0.855	0.874
Recall	0.97	0.998
F1 score	0.921	0.932
Precision	0.877	0.875

Models Result with Hypertuning:

Models	Random Forest	Boosting
Mean Testing Error	0.127	0.126
Accuracy	0.873	0.874
Recall	1.0	0.998
F1 score	0.932	0.932
Precision	0.874	0.875

Confusion Matrices

Baseline Models

```
Confusion Matrix:
[[     0  9984]
 [     0 68793]]
```

LDA

```
Confusion Matrix:
[[     0  9984]
 [     0 68793]]
```

Logistic Regression

```
Confusion Matrix:
[[   92  9892]
 [  339 68454]]
```

QDA

```
Confusion Matrix:
[[  244  9740]
 [  386 68407]]
```

KNN (for n=13)

```
Confusion Matrix:
[[   72  9912]
 [  319 68474]]
```

Naive Bayes

Ensemble Models After Hypertuning

```
Confusion Matrix:
[[   32  9952]
 [   18 68775]]
```

Random Forest Classifier

```
Confusion Matrix:
[[  166  9818]
 [  142 68651]]
```

Gradient Boosting Classifier

LESS TYPE 1 AND TYPE 2 ERRORS IN ENSEMBLE METHODS COMPARED TO BASELINE METHODS.

Georgia Tech

Result Analysis

- Logistic regression & LDA has the lowest mean testing error of 0.127 among all the baseline models
- **Boosting has the slightly lowest mean testing error with a value of 0.126 compared to baseline models.**
- Random Forest has the highest mean testing error of 0.145 without hypertuning.
- With Cross Validation - Logistic Regression , LDA has the mean testing error of 0.126 with an accuracy & precision of 0.873 i.e., we are able to predict the impact of severity 87.3% accurately
- Hypertuning has been performed for Random Forest & Boosting methods and could see a considerable change in mean testing error from 0.145 to 0.127 in case of Random Forest whereas Boosting holds same value of 0.126 with accuracy & precision values around 87%
- Confusion matrix of ensemble methods looks good with less type 1 & type 2 errors compared to baseline methods.

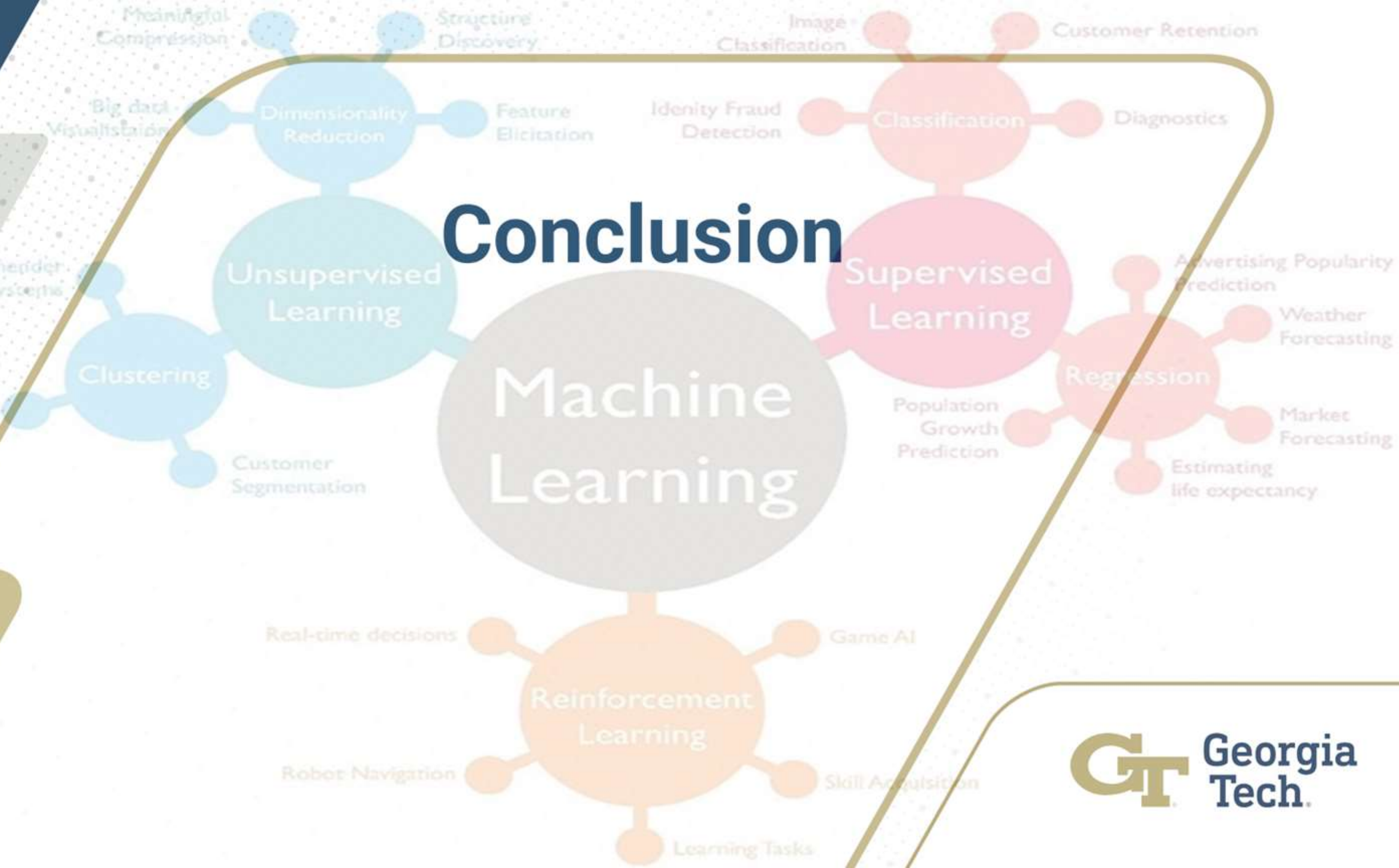
Georgia Tech

Result Analysis (Cont.)

- Based on the EDA, model results and prior research (4), below inferences were made -
  - Younger drivers tend to be more likely in an accident compared to the experienced drivers.
  - Poor visibility and the weather conditions contributing to that raise the risk of an accident as they can contribute to poor decisions.
  - Most accidents involve 1 or 2 vehicles. But when the number of vehicles involved increases, it raises the severity & fatality of the accident indicating the impact of the chain reaction

Georgia Tech

Conclusion



Georgia Tech

Conclusions

Based on the EDA, model results and prior research (4), below inferences were made -

- ❑ Infrastructure factors such as junction, road detail are significant for a particular road segment are important to determine the risk of a road for accidents. DOT/local agencies could analyse this analysis to identify the roads that need additional resources.
- ❑ While the age and experience of the driver plays a major role in determining the propensity of an accident, the engine capacity is an important factor that can reduce the accident severity.
- ❑ Maximum accidents happened on Fridays, on Road class A, Road type 6, and no junction.
- ❑ While the condition of the road is important in determining the accident severity, that itself is not significant alone when isolated.
- ❑ Since engine capacity has the highest impact on the classification models, the car manufacturer companies or the State Department of Transportation could make a policy which discourages drivers to drive cars with engine over 150,000 miles on it.

Georgia Tech

Conclusions (Cont.)

- Younger drivers tend to be more likely in an accident compared to the experienced drivers.
- Poor visibility and the weather conditions contributing to that raise the risk of an accident as they can contribute to poor decisions.
- Most accidents involve 1 or 2 vehicles. But when the number of vehicles involved increases, it raises the severity & fatality of the accident indicating the impact of the chain reaction.
- The classification models built in this project delivered a superior accuracy in predicting the severity of the crash indicating the importance to continue the research in this space,
- The classification models based on environmental, infrastructure, geographical, personal, vehicle and driving factors help a state Department of Transportation, or a local agency allocate funds judiciously on a road safety project based on the actual need backed by data
- It's possible to build and update the model that can indicate the riskiness of a geographical location combined with other factors like the driver attributes and vehicle attributes to forecast an accident and it's severity

Georgia Tech

Future Work

Although the current model demonstrated promising performances, there remains ample opportunities for refinement and optimizations -

- ❑ In the future we can experiment with different algorithms such as neural networks/deep learning to capture various patterns.
- ❑ Incorporating additional data sources beyond UK accident records can enrich the analysis for other countries
- ❑ Exploring advanced visualization techniques can help communicate findings more effectively and uncover hidden patterns in the data.
- ❑ Our project could be involved in developing recommendations for public policy interventions aimed at reducing road accidents.
- ❑ Integrating external factors and events into future analysis is imperative for comprehensive understanding of road accidents.
- ❑ We dealt with accident data in this project. But if some data collection project is done to count all the vehicles that pass through the roads, we can also estimate the likelihood of an accident.

Work Distribution Summary

#	Name	Responsibilities
1	Praneetha Kommineni	Data Research, Model Building, Feature Extraction, Cross Validation, Data Mining and Hypertuning
2	Sai Pooja Panda	Prior Research, Documentation, Inferences of Results, EDA, Evaluation of final results
3	SaiChandan Duggirala	Project Management, EDA, Inferences of Results, Scope of Work, Data Analysis, Documentation
4	Sayanta Barman	Machine Learning Model Evaluation, Model Building, Feature Extraction, Cross Validation, Data Mining and Hypertuning
5	Tabassum Shahid	Project Lead, Confusion Matrices, Meeting scheduling, Coordination, Documentation, Action Items and Future Scope

Georgia Tech

Reference

1. https://www.data.gov.uk/dataset/208c0e7b-353f-4e2d-8b7a-1a7118467acc/gb-road-traffic-counts
2. https://www.cdc.gov/injury/features/global-road-safety/index.html#:~:text=Crash%20injuries%20are%20estimated%20to,crashes%20than%20from%20HIV%2FAIDS.
3. https://medium.com/analytics-vidhya/analysis-of-uk-accident-dataset-1d4abf773e68
4. https://arxiv.org/ftp/arxiv/papers/2309/2309.13483.pdf

Georgia Tech

Georgia Tech