Insight into the Nature of Road Collisions

Group 5 - SOEN 471 Big Data

UNIVERSITÉ Concordia UNIVERSITY

Maxime Johnson 40081684
Dominik Ludera 40062500
Alvira Konovalov 40074264
Matthew Padvaiskas 40034075

Agenda for our Presentation

01 Introduction

Research Questions, Model Selection,
Dataset Selection

02 Data Preparation

Preprocessing of data,
Feature Selection, Cleaning

03 Model Implementation

Implementation of the chosen
models, alternative models

04 Model Evaluation

Interpretation of model
results, Conclusion

1.Introduction

Research Questions

Feature Significance

Which features are most significant in determining the outcome of an accident?

Prediction

Can one predict the outcome of an accident by analyzing the attributes of an accident?

Best Model

Which machine learning technique predicts best the outcome of an accident?

Dataset Selection

Motor Vehicle Collisions in City of Toronto

- Data from **2006 - 2021**
  - Updated annually in May
- **16,861** motor vehicle collisions
- **54** features including:
  - Driver and weather conditions, time & date, location, result of collision, etc.

https://open.toronto.ca/dataset/motor-vehicle-collisions-involving-killed-or-seriously-injured-persons/

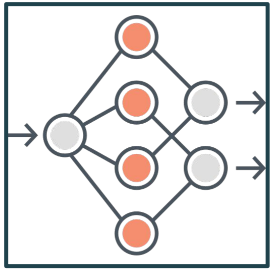
Model Selection

Random Forest

Deterministic

Split the feature space along the various features in order to optimize the gain of information

XX MODEL LICENSE RF



Neural Network

Initially Probabilistic

Each neuron watches over a specific feature space and activates once the input falls into that space

XX MODEL LICENSE NN

Models chosen as they similarly break down the problem piece by piece, but handle the data differently

2.Data Preparation

Feature Selection

We kept 25 out of 54 features.

Relevancy

Keep information relevant to our question

Uniqueness

Remove redundancy and embedded information

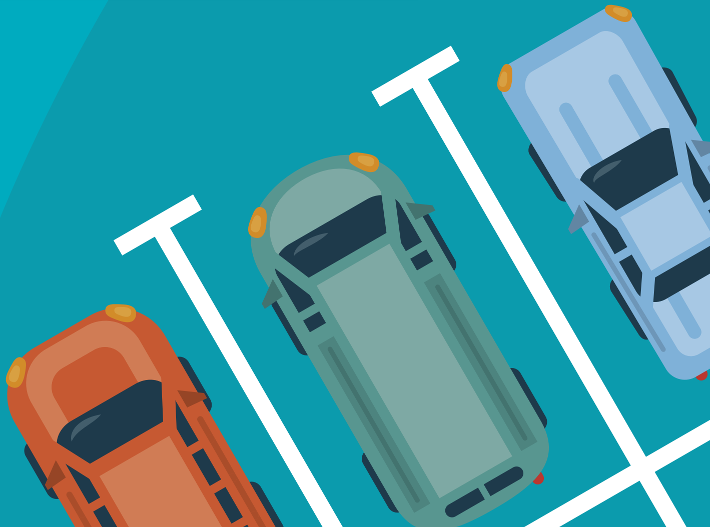
Informative

non-informative variables can add uncertainty and reduce the overall effectiveness of the model

Cleaning Data

- **Mapping binary values**
  - **Convert “Yes” and “null” by 1 and 0.**
- Reducing feature range
  - Simplify “Date” to “Month”
- Grouping similar values
  - Categorize “TRAFFCTL” into 3 classes
  - Categorize “ROAD_CLASS” into 5 classes
- Random cleaning
  - Make values uniform
- Dropping rows
- Label encoding

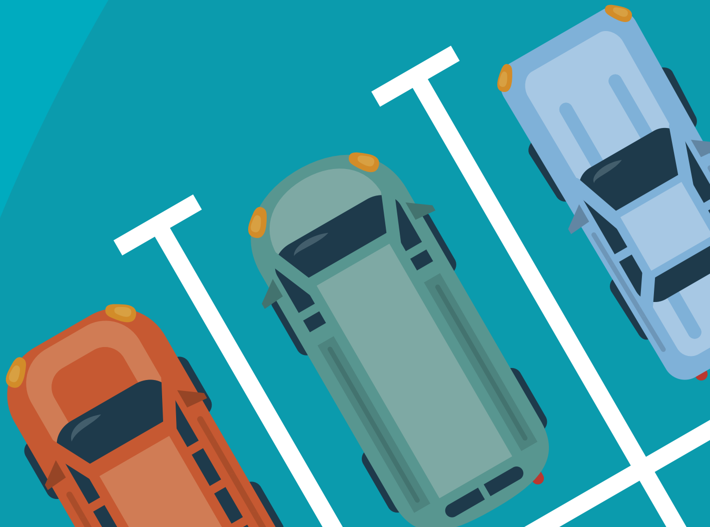
```
df.ALCOHOL.fillna(0, inplace=True)
df.ALCOHOL.replace('Yes', 1, inplace=True)

df.PEDESTRIAN.fillna(0, inplace=True)
df.PEDESTRIAN.replace('Yes',1, inplace=True)

df.SPEEDING.fillna(0, inplace=True)
df.SPEEDING.replace('Yes', 1, inplace=True)
```



Cleaning Data

- Mapping binary values
  - Convert “Yes” and “null” by 1 and 0.
- **Reducing feature range**
  - **Simplify “Date” to “Month”**
- Grouping similar values
  - Categorize “TRAFFCTL” into 3 classes
  - Categorize “ROAD_CLASS” into 5 classes
- Random cleaning
  - Make values uniform
- Dropping rows
- Label encoding

```
df['DATE'] = df['DATE'].dt.month
```



Cleaning Data

- Mapping binary values
  - Convert “Yes” and “null” by 1 and 0.
- Reducing feature range
  - Simplify “Date” to “Month”
- **Grouping similar values**
  - **Categorize “TRAFFCTL” into 3 classes**
  - **Categorize “ROAD_CLASS” into 5 classes**
- Random cleaning
  - Make values uniform
- Dropping rows
- Label encoding

```
df['TRAFFCTL'] = df['TRAFFCTL'].replace(['Traffic Signal', 'Stop Sign',
                 'Pedestrian Crossover', 'Yield Sign',
                 'Streetcar (Stop for)', 'Traffic Gate'],
                 'Passive Control')
df['TRAFFCTL'] = df['TRAFFCTL'].replace(['Police Control', 'School Guard',
                 'Traffic Controller'], 'Active Control')
```

Cleaning Data

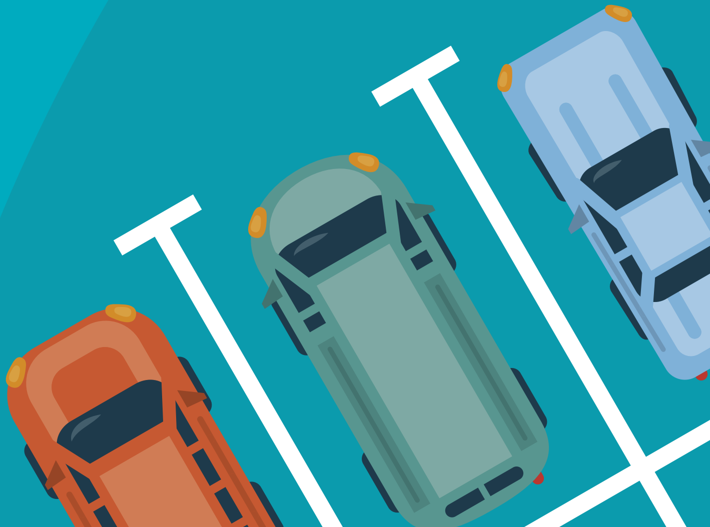
- Mapping binary values
  - Convert “Yes” and “null” by 1 and 0.
- Reducing feature range
  - Simplify “Date” to “Month”
- Grouping similar values
  - Categorize “TRAFFCTL” into 3 classes
  - Categorize “ROAD_CLASS” into 5 classes
- **Random cleaning**
  - **Make values uniform**
- Dropping rows
- Label encoding

```
df.INVAGE = df.INVAGE.replace(['unknown'], 'Unknown')
```

Cleaning Data

- Mapping binary values
  - Convert “Yes” and “null” by 1 and 0.
- Reducing feature range
  - Simplify “Date” to “Month”
- Grouping similar values
  - Categorize “TRAFFCTL” into 3 classes
  - Categorize “ROAD_CLASS” into 5 classes
- Random cleaning
  - Make values uniform
- **Dropping rows**
- Label encoding

```
df.drop(df[df.LOCCOORD.isnull()].index, inplace=True)
df.drop(df[df.LIGHT == 'Other'].index, inplace=True)
```



Cleaning Data

- Mapping binary values
  - Convert “Yes” and “null” by 1 and 0.
- Reducing feature range
  - Simplify “Date” to “Month”
- Grouping similar values
  - Categorize “TRAFFCTL” into 3 classes
  - Categorize “ROAD_CLASS” into 5 classes
- Random cleaning
  - Make values uniform
- Dropping rows
- **Label encoding**

```
df['ACCLASS'] = df['ACCLASS'].astype('category').cat.codes
df['INITDIR'] = df['INITDIR'].astype('category').cat.codes
df['LIGHT'] = df['LIGHT'].astype('category').cat.codes
df['VISIBILITY'] = df['VISIBILITY'].astype('category').cat.codes
df['RDSFCOND'] = df['RDSFCOND'].astype('category').cat.codes
df['ROAD_CLASS'] = df['ROAD_CLASS'].astype('category').cat.codes
df['TRAFFCTL'] = df['TRAFFCTL'].astype('category').cat.codes
df['INVAGE'] = df['INVAGE'].astype('category').cat.codes
df['LOCCOORD'] = df['LOCCOORD'].astype('category').cat.codes
df['MANOEUVER'] = df['MANOEUVER'].astype('category').cat.codes
```

3.Model Implementation

3 Groups of Models

Group 1 Models

Original dataset

Group 2 Models

Undersampling

Group 3 Models

Feature reduction

Implementation Approach

DATA SPLITTING

70% training data
30% testing data

PARAMETER SEARCH

Tune the hyper-parameters with **GridSearchCV**

MODEL TRAINING

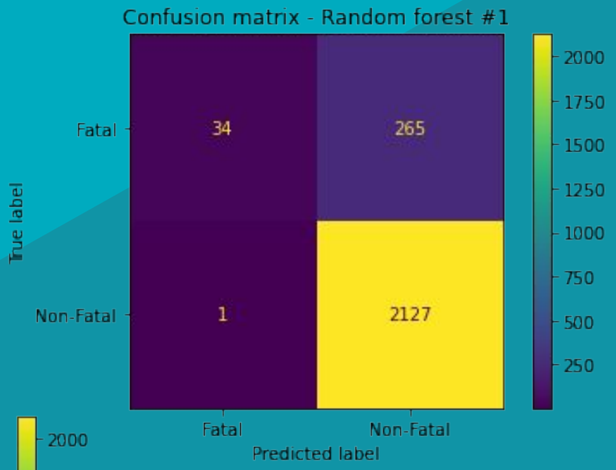
Train model using parameters of best estimator

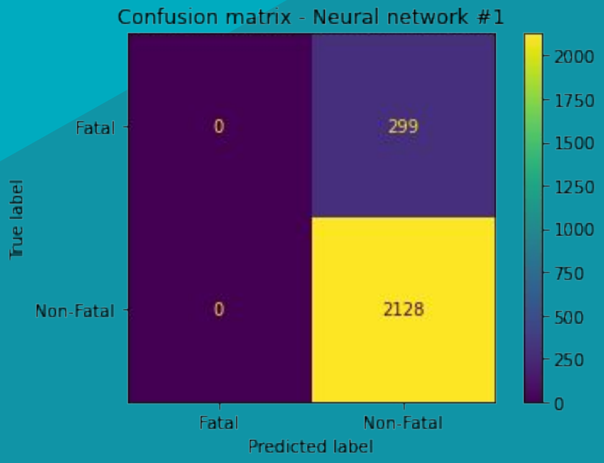
```
params = {
    'n_estimators': [100, 500, 1000],
    'criterion': ['gini', 'entropy'],
    'min_samples_split': [2, 4, 5, 10, 13],
    'min_samples_leaf': [1, 2, 5, 8, 13]
}

forest =
GridSearchCV(RandomForestClassifier(random_state=0), params)
forest.fit(X_train, y_train)
```

```
params = {
    'hidden_layer_sizes': [(50,50,50), (50,100,50), (100,)],
    'activation': ['tanh', 'relu'],
    'solver': ['sgd', 'adam'],
    'alpha': [0.0001, 0.05],
    'learning_rate': ['constant','adaptive'],
}
nn = GridSearchCV(MLPClassifier(max_iter=100), params,
n_jobs=-1, cv=3)
nn.fit(X_train, y_train)
```

Group 1 Models

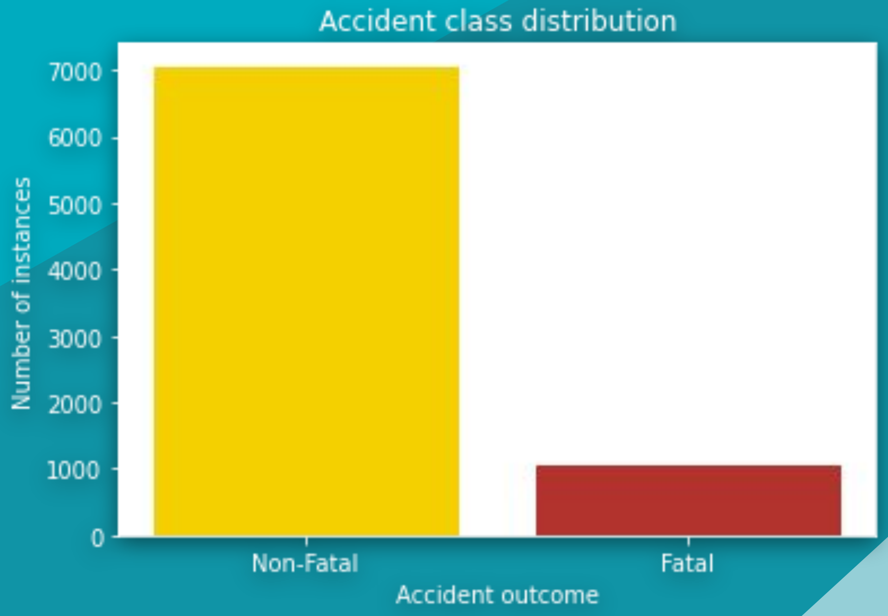
- Original dataset
- 25 features
- **The Metric Trap:**
    - **Misleading accuracy score!**
    - **Random forest #1**
        - 89% accuracy
    - **Neural network #1**
        - 88% accuracy
- Classifies all accidents as non-fatal (majority class)
- **Fails** to capture fatal accidents (minority class)





Group 1 Models

- Imbalanced class distributions
- **Class imbalance problem:**
  - Classifiers predict everything as the majority class (Non-fatal)
- **Solution - undersampling:**
  - Randomly delete instances from the majority class



Group 2 Models

- Undersampling based on Near Miss method (`imblearn` library)

```
undersample = NearMiss(version=1, n_neighbors=3)
X,y = undersample.fit_resample(X, y)
```

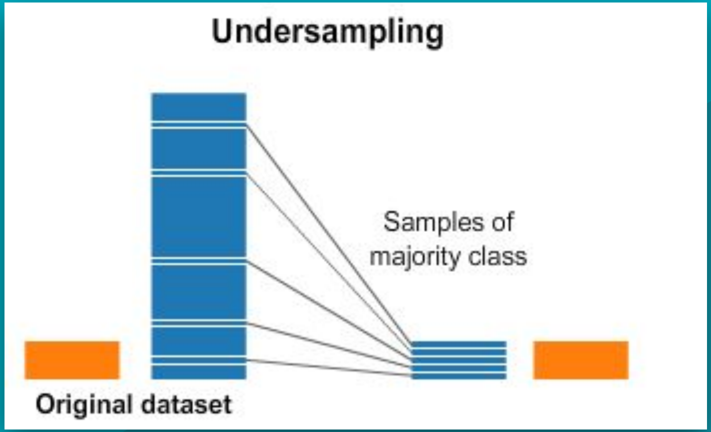
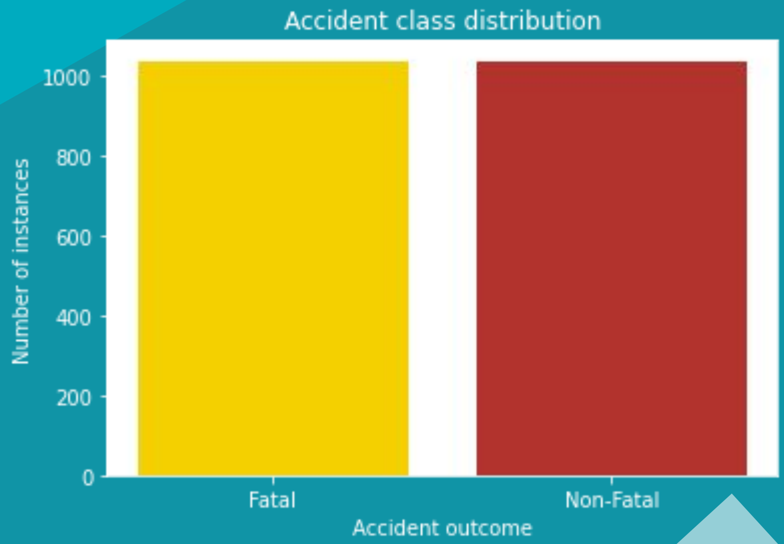


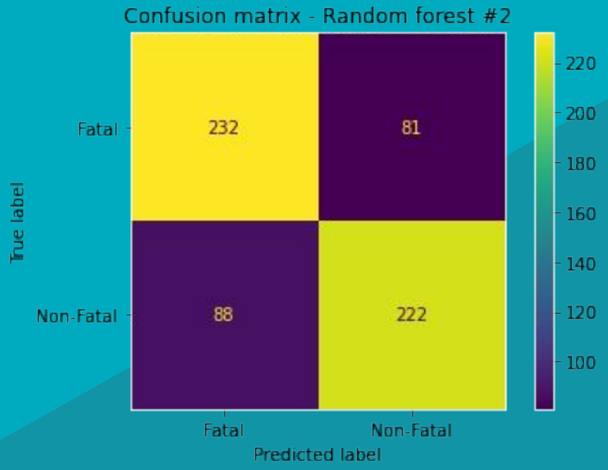
Image credit:
https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/

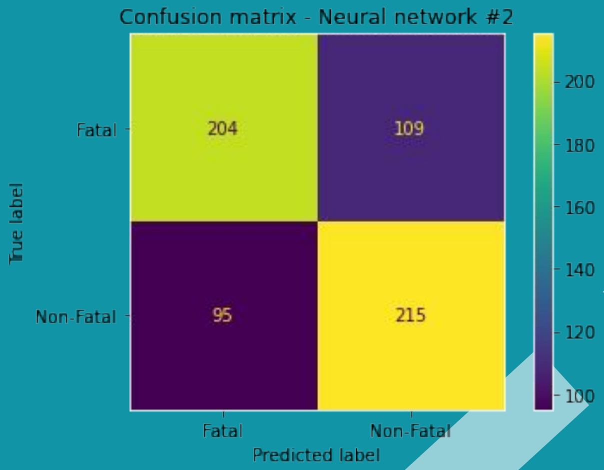


Group 2 Models

- Random forest #2 and neural network #2 using undersampled dataset
- Captures both classes (Non-fatal and fatal) equally
- Performance decreased but is more reasonable

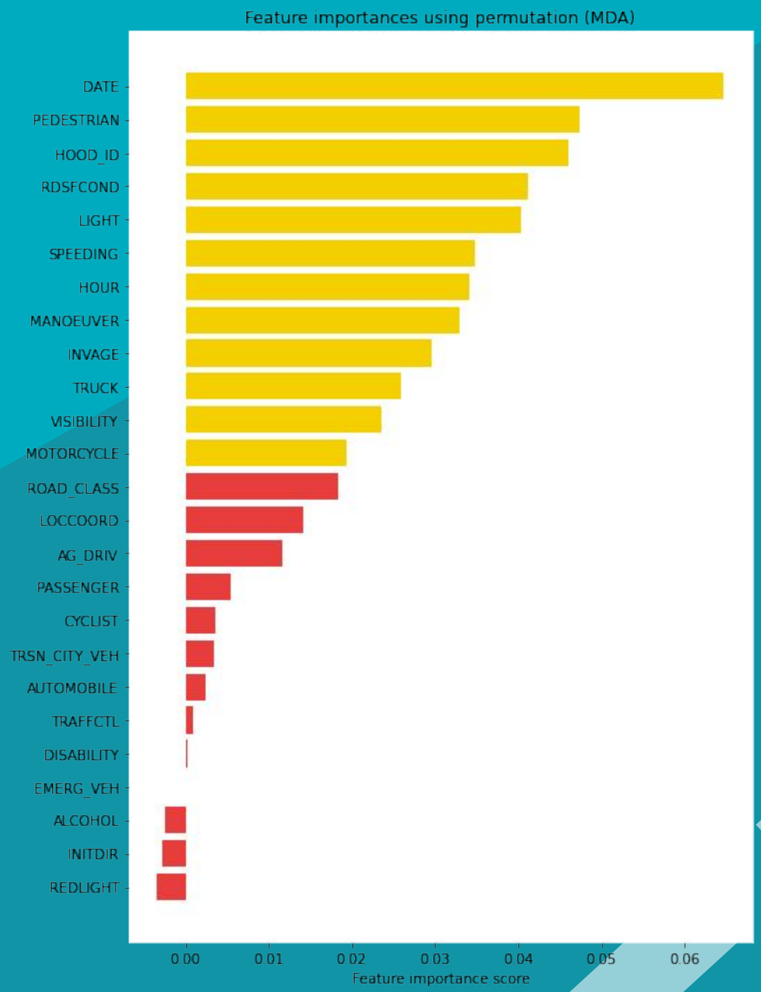
MODEL	ACCURACY	PRECISION	RECALL	F1
random forest 2	80.26	80.57	80.26	80.21
neural network 2	74.96	76.99	74.96	74.46





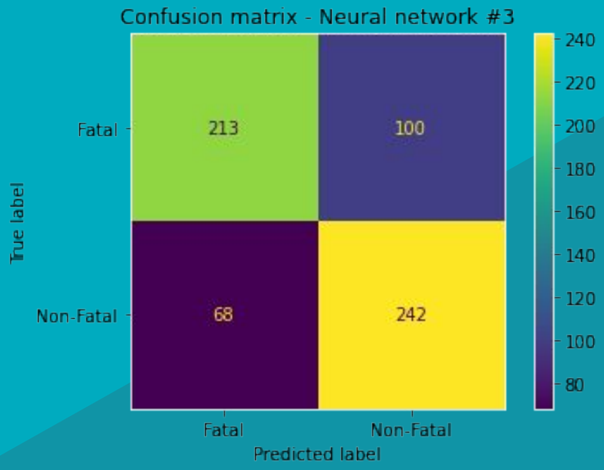
Group 3 Models

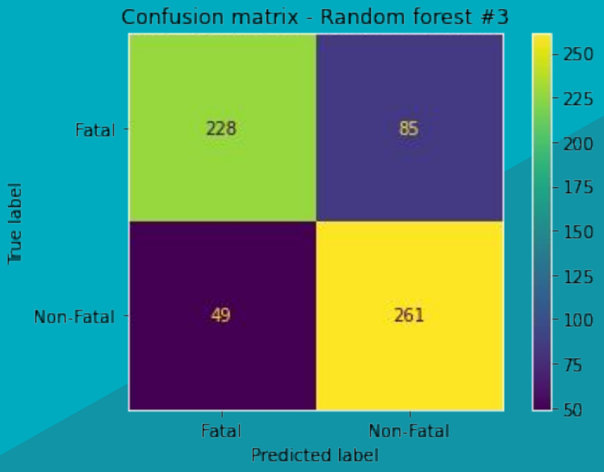
- Dataset with many features can lead to **overfitting**
- **Feature reduction:**
  - Calculate `feature_importances` random forest #2 and feature permutation
  - Keep **12 most influential features**



Group 3 Models

- Results are still satisfactory after reducing features from 54 to 12 features!





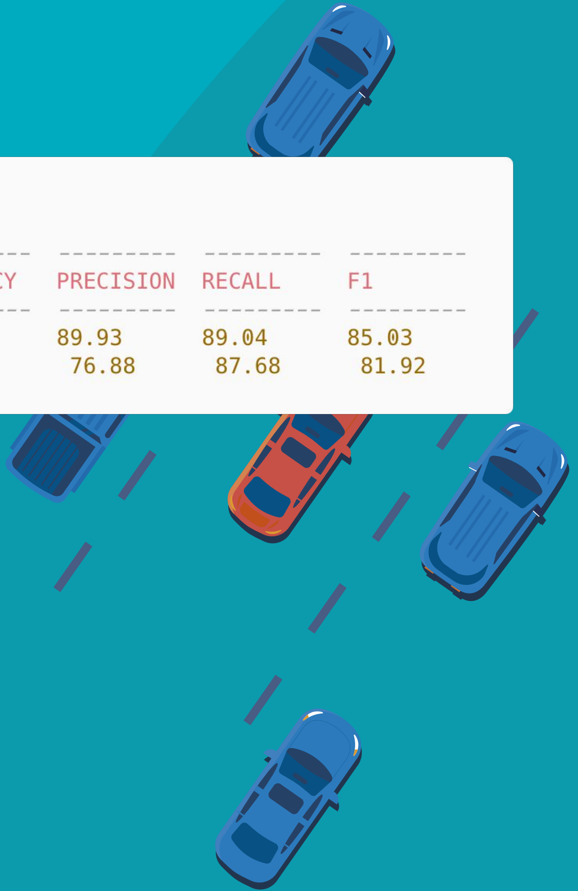
MODEL	ACCURACY	PRECISION	RECALL	F1
random forest 3	78.49	78.89	78.49	78.43
neural network 3	73.03	73.29	73.03	72.97

4.Model Evaluation & Conclusions

Model Evaluation

- **Models pre / post data resampling**
- Effects of feature reduction
  - Minor decrease in random forest performance
  - Increase in neural network performance
- Final model comparisons

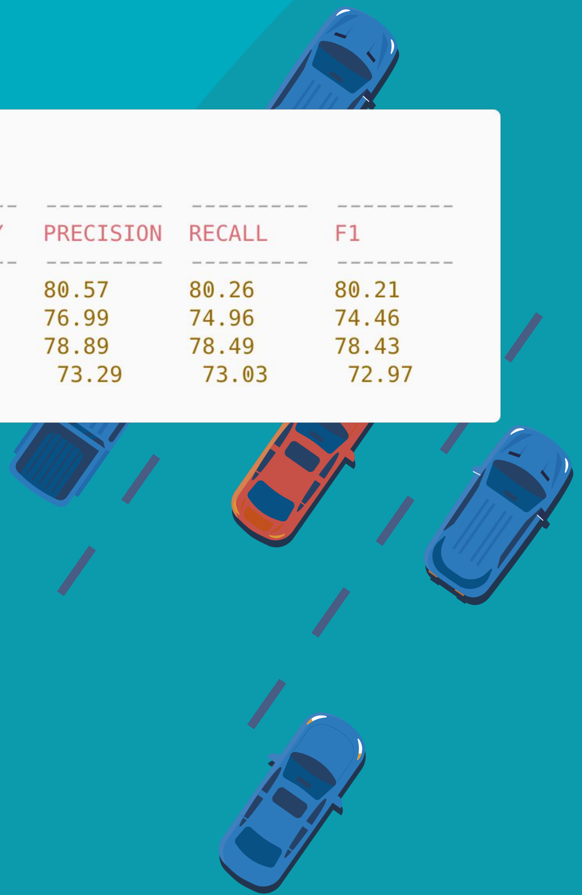
MODEL	ACCURACY	PRECISION	RECALL	F1
random forest 1	89.04	89.93	89.04	85.03
neural network 1	87.68	76.88	87.68	81.92



Model Evaluation

- Models pre / post data resampling
- **Effects of feature reduction**
    - **Minor decrease in random forest performance**
    - **Increase in neural network performance**
- Final model comparisons

MODEL	ACCURACY	PRECISION	RECALL	F1
random forest 2	80.26	80.57	80.26	80.21
neural network 2	74.96	76.99	74.96	74.46
random forest 3	78.49	78.89	78.49	78.43
neural network 3	73.03	73.29	73.03	72.97



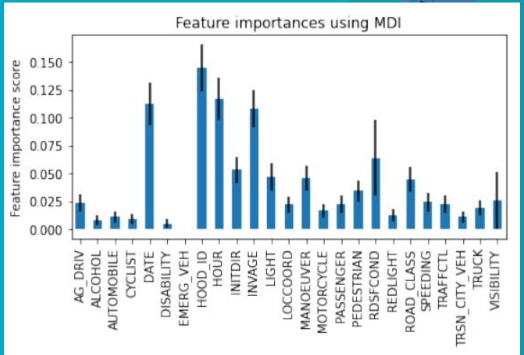
Model Evaluation

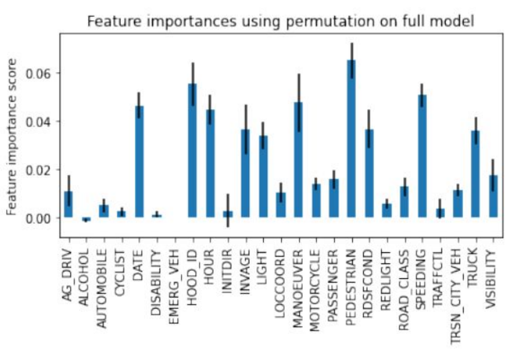
- Models pre / post data resampling
- Effects of feature reduction
  - Minor decrease in random forest performance
  - Increase in neural network performance
- **Final model comparisons**

MODEL	ACCURACY	PRECISION	RECALL	F1
random forest 2	80.26	80.57	80.26	80.21
neural network 2	74.96	76.99	74.96	74.46
random forest 3	78.49	78.89	78.49	78.43
neural network 3	73.03	73.29	73.03	72.97

Conclusions

- **Which features were most significant in determining our model?**
  - **Neighborhood and month quite important**
  - **Truck, pedestrian, or speeding involved**
  - **Low importance of alcohol or narcotics**
- Can we predict the outcome of accidents?
- Which machine learning technique provided the best outcome?





Conclusions

- Which features were most significant in determining our model?
  - Neighborhood and month quite important
  - Truck, pedestrian, or speeding involved
  - Low importance of alcohol or narcotics
- **Can we predict the outcome of accidents?**
- Which machine learning technique provided the best outcome?

Conclusions

- Which features were most significant in determining the outcome of an accident?
  - Neighborhood and month quite important
  - Truck, pedestrian, or speeding involved
  - Low importance of alcohol or narcotics
- Can we predict the outcome of accidents?
- **Which machine learning technique provided the best outcome?**

MODEL	ACCURACY	PRECISION	RECALL	F1
random forest 2	80.26	80.57	80.26	80.21
neural network 3	73.03	73.29	73.03	72.97

Thank You
Any Questions?

UNIVERSITÉ
Concordia
UNIVERSITY