PREDICTING COLLISION SEVERITY IN SEATTLE

IBM DATA SCIENCE PROFESSIONAL CAPSTONE PROJECT

PREPARED BY MARK SNUFFIN

BACKGROUND

- This presentation is for the Capstone Project to complete the IBM Data Sciences Professional Program
- The storyline is fictional study sanctioned by SDOT and performed by internal IT staff to determine the feasibility of using Machine Learning and Data Sciences to build an application that informs the public of potential high-risk driving situations
- To gain traction within SDOT the study focused on using machine learning predict if it collision results in an injury

Seattle Times

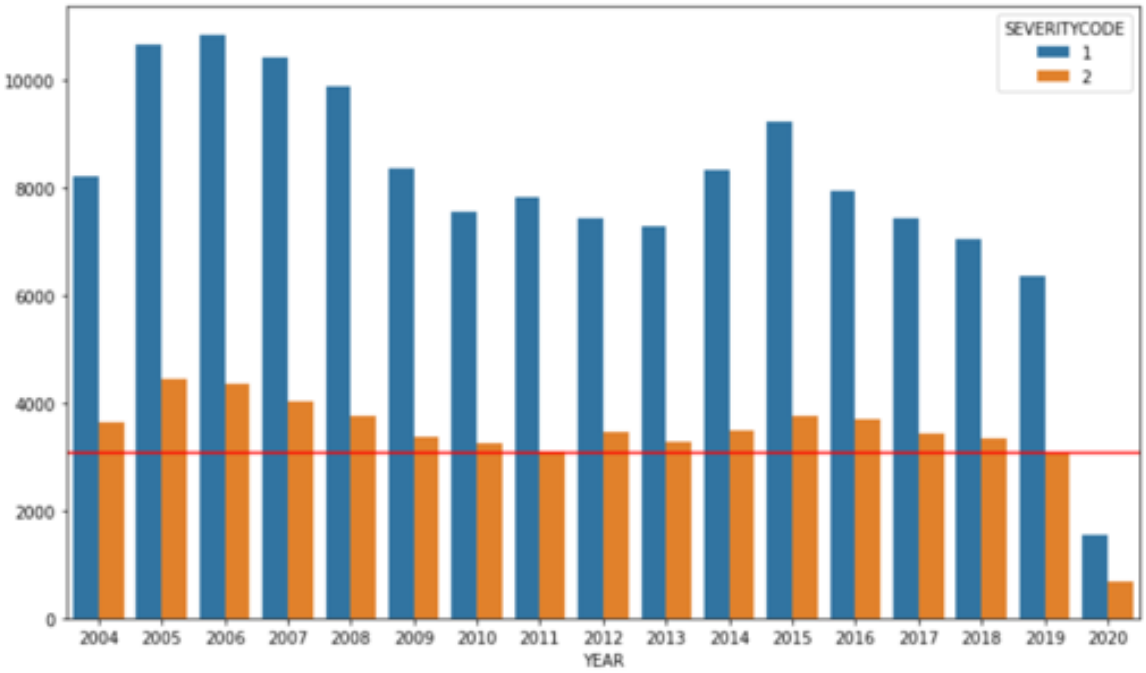
WE ALWAYS NEED TO DO BETTER

- While collisions overall in Seattle are down, those with injury have remained steady for over a decade
- Can SDOT utilize Data Sciences and Machine Learning to provide tools to the public and emergency services to increase safety and reduce risk?

INTRODUCTION

- Since 2004 collisions without an injury have been on the decline
- However collisions with an injury have remained steady
- Other factors such as driver distraction are an ongoing issue as mobile technologies introducing new challenges to public safety
- SDOT has collected data from 2004 where it is generally used for basic reporting and statics
- Emerging technologies are enabling new ways to approach this problem utilizing the data SDOT already has

Seattle Collisions from 2014 to May 2020



Orange bars represent a collision with at least one injury.
Blue bars represent collision with no injury, property only.
Note that 2020 has partial data up to May

INTRODUCTION

- For the 2021 budget SDOT is has tasked the Internal Technology Department with determining ways to leverage emerging technologies to improve public safety on the roads
- Rather then provide static information the SDOT administration is seeking tools and technologies that would provide the public and the public sector agencies with better tools maximize safety and awareness, providing actionable intelligence is a key objective
- SDOT Technology Division proposed the use of Data Sciences and Machine Learning to meet this objective
- These technologies can be used to
  - Influence Public Service Announcements, e.g., Don't Text and Drive, Drunk Driving, Road Rage, based on trends
  - Help first responder in quickly ascertaining high risk collisions to react with the appropriate level of resources
  - Assist drivers with determining how risky a trip could be based on the environmental factors and past history
- Before allocating funds in the budget management requested a study that would leverage the existing collision data to determine the efficacy of this approach and whether the technology division has the appropriate resources to deliver this capability

INTRODUCTION - THE PROBLEM

- SDOT Technology Division to perform a study to predict Injury Collisions
- Study is to determine if Machine Learning can be used to accurately predict a collision will involve an injury
- Accurately predicting collision severity automatically would demonstrate the departments ability to leverage this technology
- A caveat is that no additional data collection is allowed, all predictions must use the existing collision data maintained by SDOT
- The technology division needs to present this findings to the administration and if the results are positive funding will be allocated to utilize Machine Learning and Data Sciences

DATA USED FOR THE STUDY

- Data was obtained from the Seattle Department of Transport (SDOT) Management Division from 2004 to present
- There are 194,673 collisions recorded with 38 attributes corresponding to each collision. This data is entered by the police, traffic services and the state
- Contains information including location, temporal, environmental, human factors and descriptors that define the collision
- The dataset was analyzed to extract the best features to predict Injury Collisions
- Data was cleaned to remove abnormalities and missing data was replaced with estimated values where possible
- From this dataset 14 features were extracted along with the severity code

DATA PREPARATION – FEATURES SELECTED

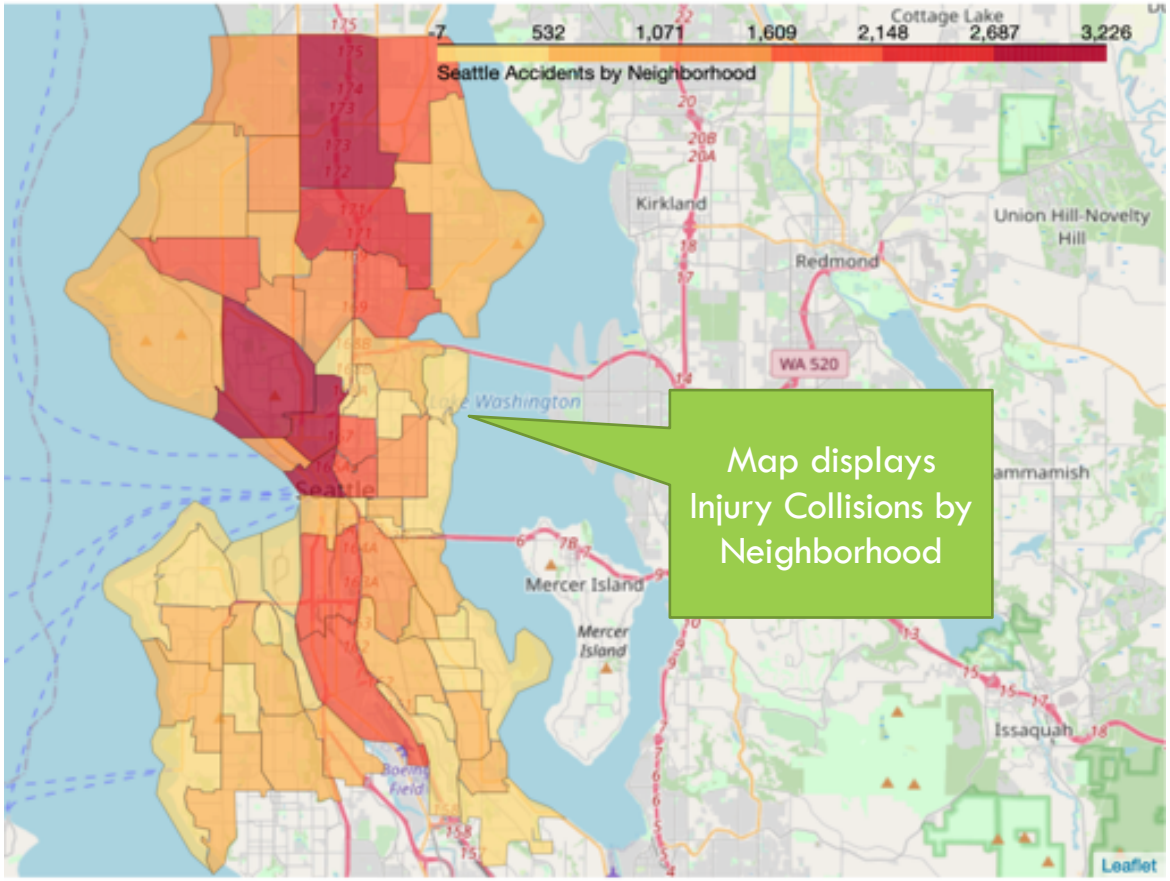
Feature Descriptions	
SEVERITYCODE The target variable, can be Property or Injury	INATTENTIONIND Was the driver was distracted?
ADDRTYPE Generalized location of the collision, e.g., intersection, block or alley	UNDERINFL Was the driver was intoxicated?
COLLISIONTYPE Technical description of the how the vehicle(s) collided	WEATHERCOND, ROADCOND, LIGHTCOND Environmental conditions
PERSONCOUNT Number of people involved	PEDROWNOTGRNT Was a pedestrian granted the right of way?
VEHCOUNT Number of vehicles involved	SPEEDING Was driver exceeding the posted speed limit?
JUNCTIONTYPE If the collision occurred at a junction what was the type	HITPARKEDCAR Was a parked vehicle involved?
SDOT_COLCODE A code assigned by the SDOT describing the collision	

EXPLORATORY DATA ANALYSIS (EDA)
TARGET VARIABLE – COLLISION SEVERITY

- Defined as SEVERITYCODE
- Chart shows collision severity code over the lifetime of the dataset
- Has two categorical codes
  - Category 1 - Property Damage Only Collision
  - Category 2 - Injury Collision
- Dataset has 136,485 Property Collisions
- And 58,188 Injury Collisions
- Study goal is to accurately predict Injury collisions

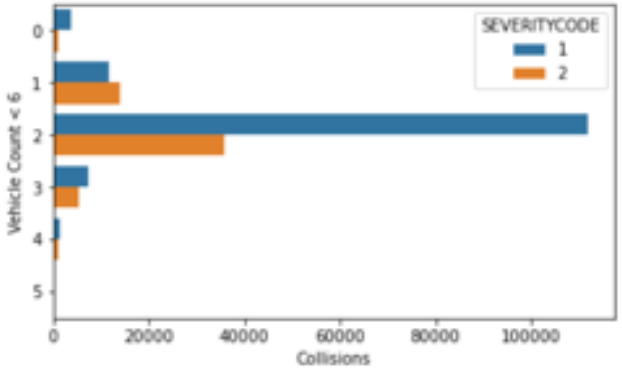
EDA – LOCATION FEATURES

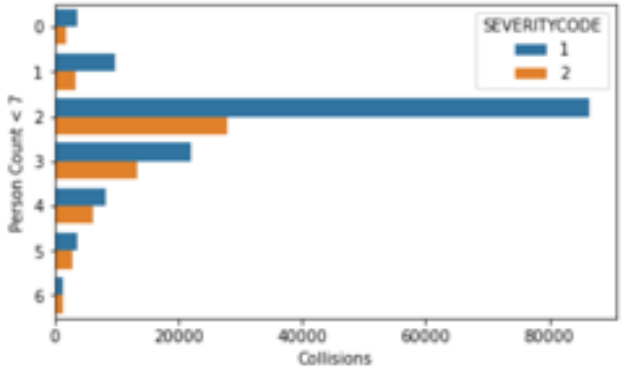
- Displaying collisions on the map clearly indicates hotspots exist, especially those within major interstates or downtown
- While relevant these did not show a strong correlation with predicting Injury Collisions
- Generalized categorical locations such as intersections and blocks showed a higher correlation
- Street(s) where a collision occurred were captured as text and not useful. These could have a lot of value but require parsing

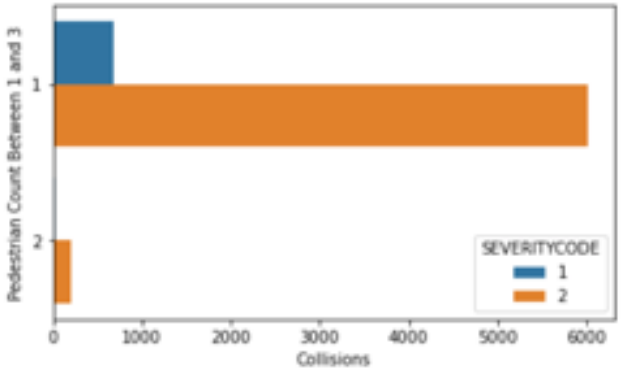


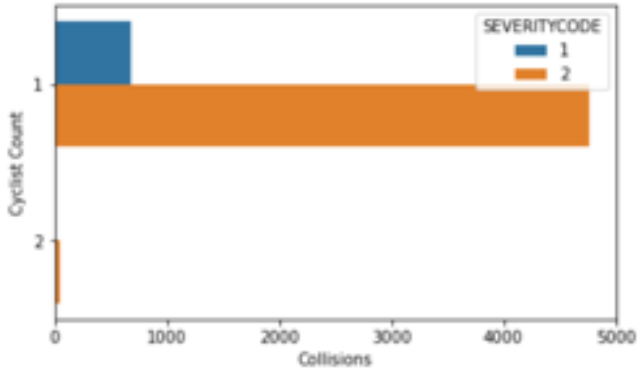
EDA - COUNTS

- For vehicle counts the majority are two vehicles colliding
- Most collisions occur with two to three people involved. Three and four people involved in an accident has a much stronger correlation to severity
- Just about any collision related to with a pedestrian or cyclist results in an injury (there is not designation for a motorcycle)
- All the counts could have been binned but were not. This could be a good improvement in the future



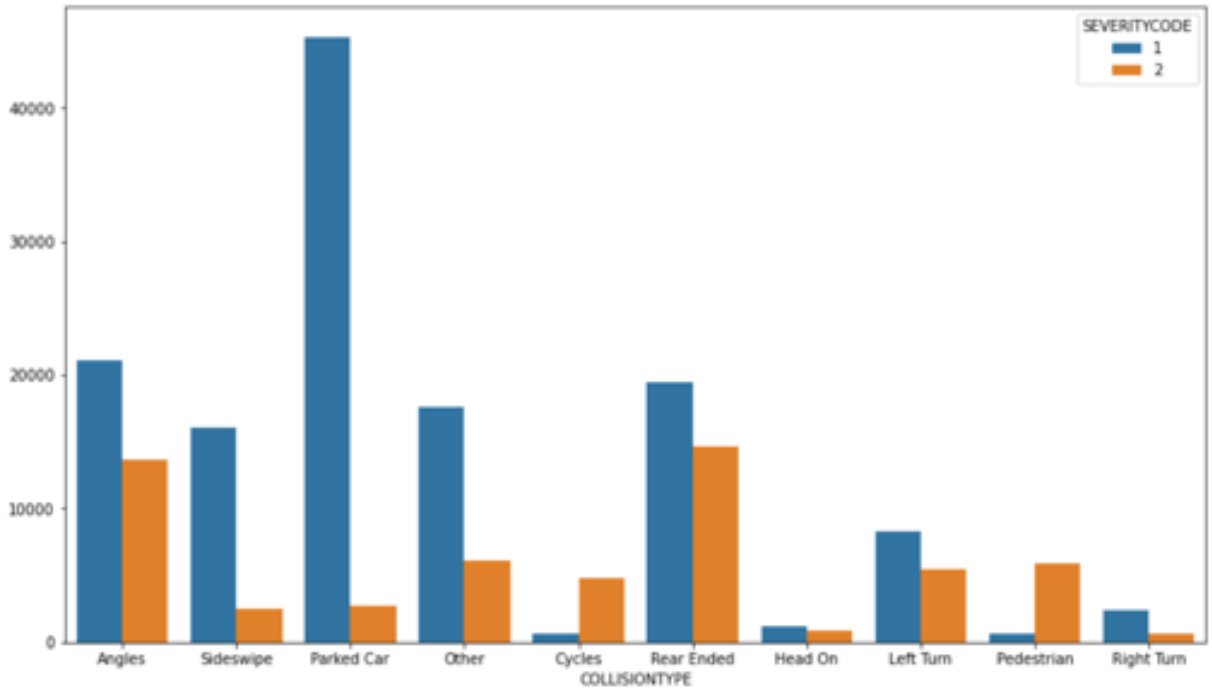






EDA – COLLISION FEATURES

- Collision type has a strong correlation with severity (2nd best overall)
- Hitting a parked car is by far the most frequent but also a good indicator of when there is not an injury
- Rear Ended, when the vehicle hits another vehicle from behind has the greatest chance of severity being an Injury
- There is overlap with SDOT Collision Codes. Both where kept in for modelling though
- Any time a pedestrian is involved there is a high chance of injury
- Left turns are more dangerous than right turns.



EDA – COLLISION FEATURES

- SDOT Collision Code offers a good indicator of severity
- Each sample has two collision codes that overlap. One is assigned by the state and the one shown here by the city. There is overlap here but city codes look better and more description – reviewing those was out of scope
- Overall this code was one of the features with the highest correlation coefficient
- This code was defined for every collision

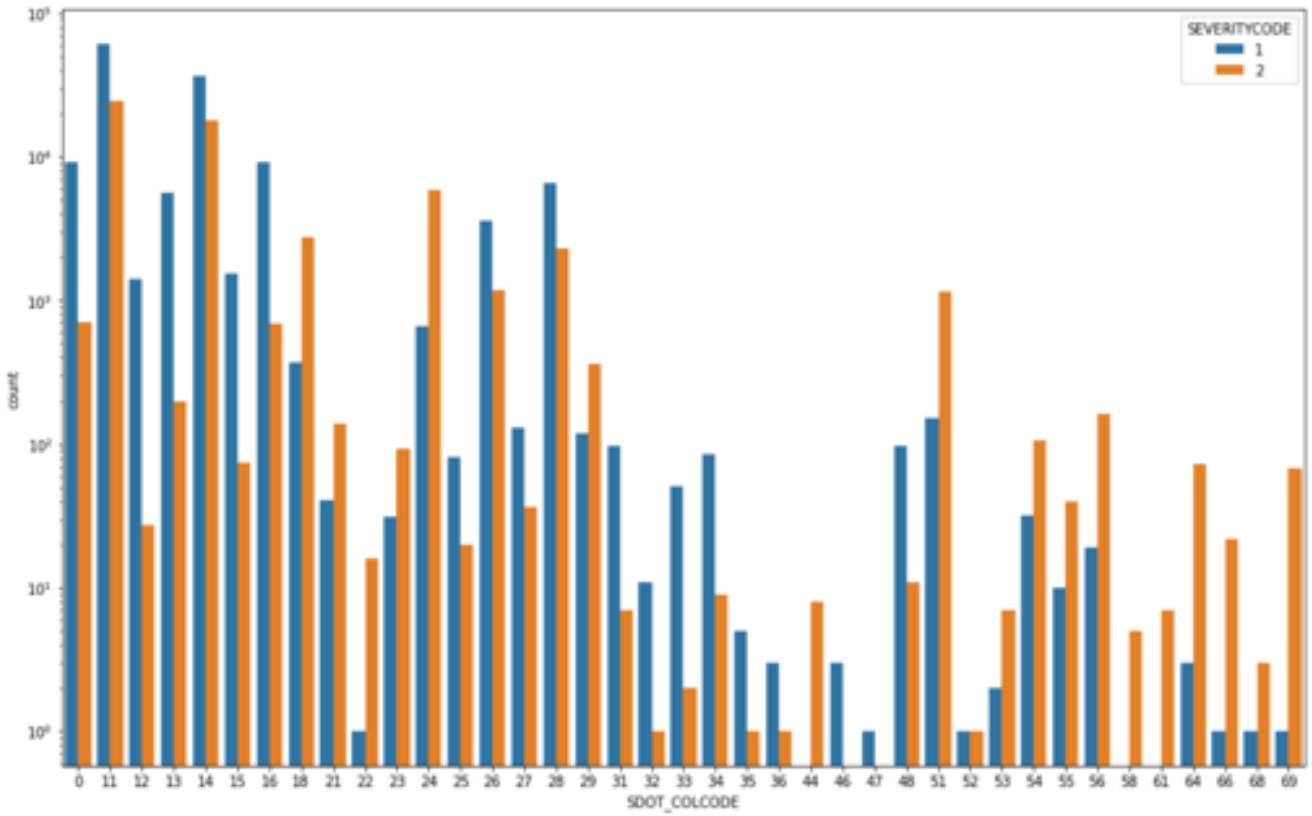
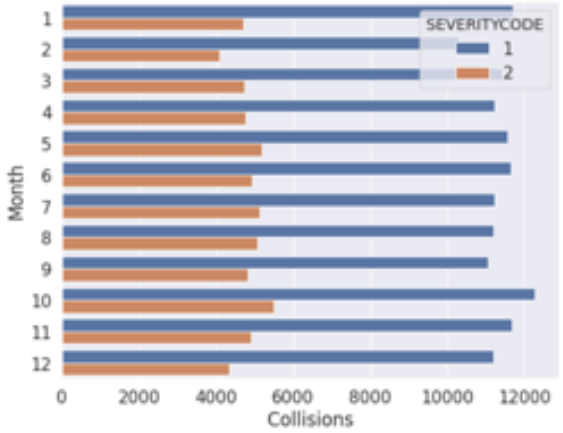
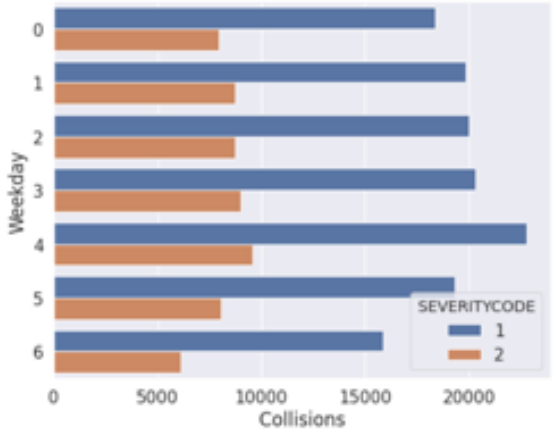


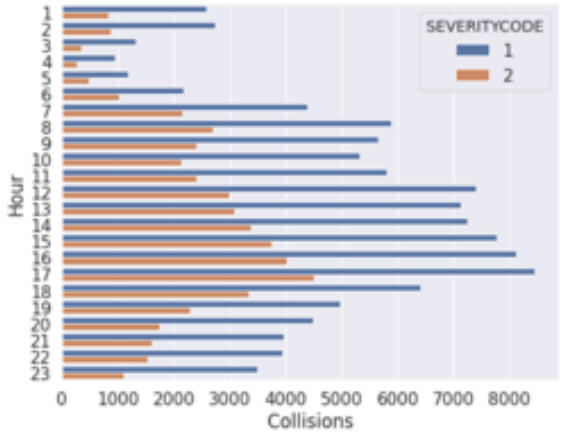
Chart showing log scale

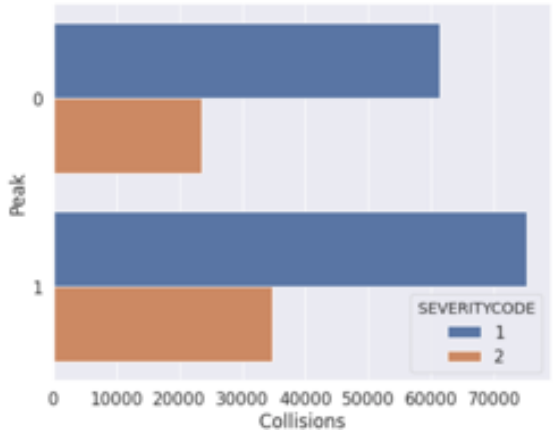
EDA – TEMPORAL FEATURES

- Date was excluded from the feature set
- Various components of the date the collision took place were analyzed
- The day did not offer any strong relation to Injury severity however most collisions occurred at the end of the week
- Hour of day had a stronger relation but not strong enough
- Month was pretty much equal across the board
- Peak was an engineered feature that has promise however it was not used



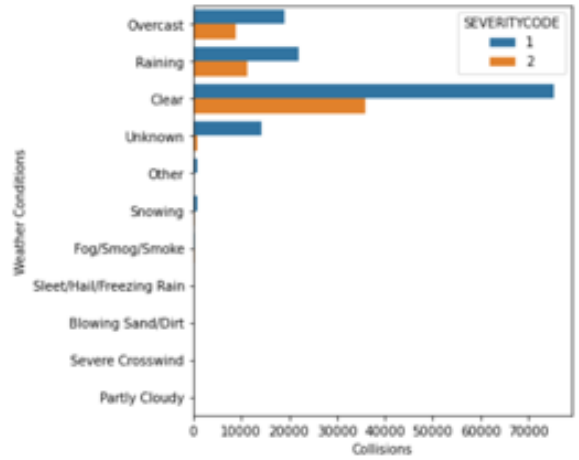


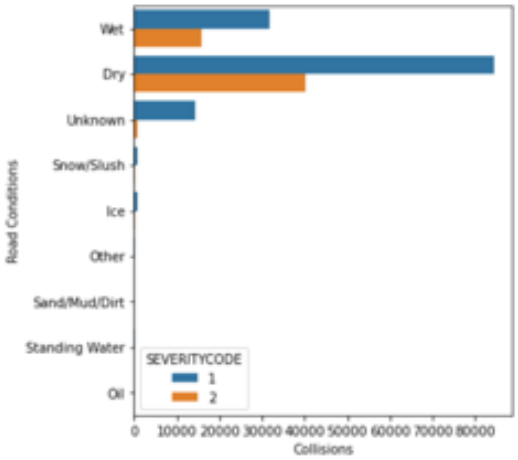


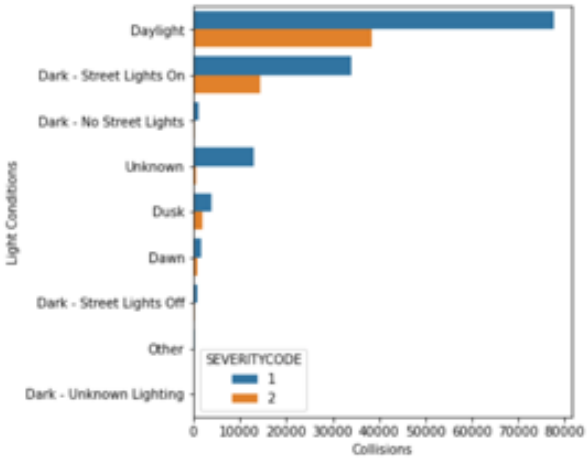


EDA – ENVIRONMENTAL

- Weather conditions indicate that most of collisions occur when the weather is clear
- When road conditions are dry there are the most collisions
- Daylight is when the most occur
- It rains a lot in Seattle so it's not surprising to see rain and overcast resulting in most collisions when not clear. Combining these was considered but left these separate
- Wet roads and weather are redundant but were left in, this could be an optimization
- Unknown lighting conditions was left in as-is

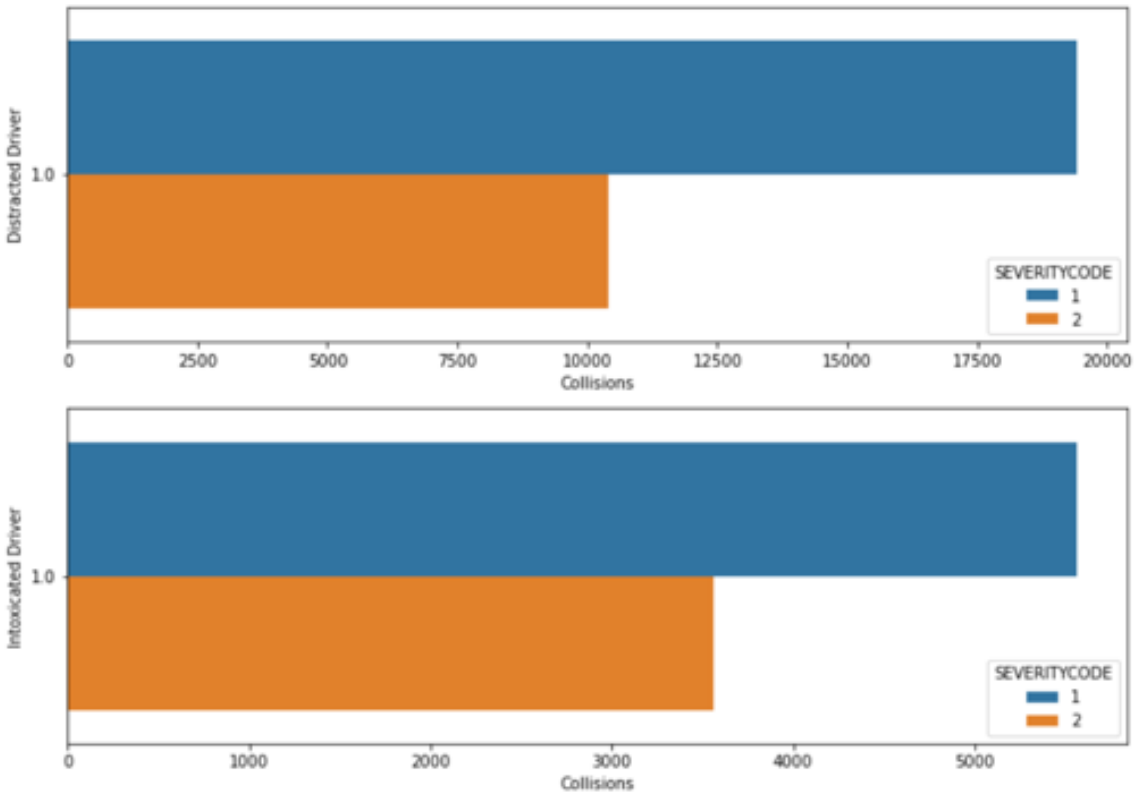






EDA – HUMAN FACTORS

- Distracted drivers are an issue
- Especially considering this number has increased over time – there are more distractions now that everyone has phones and other devices in the car
- Intoxicated drivers have a relatively low sample but when intoxicated there is stronger chance of an Injury



DATASET BALANCING

- Started with Total collisions = 194,673
  - Property Collisions = 136,485
  - Injury Collisions = 58,188
- Unbalanced data set will bias the models to favor the majority class
- Random Under Sampling was used for balancing the targets, Property == Injury
- Reduced Property Collisions down to 58,188
- Final dataset was 116,376 collisions in the dataset
- Other resampling techniques, such as SMOTE, were not explored but could be a better choice. Oversampling the minor class in this case could have it benefits

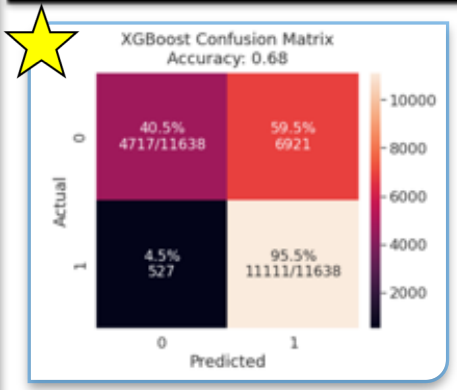
MODELLING

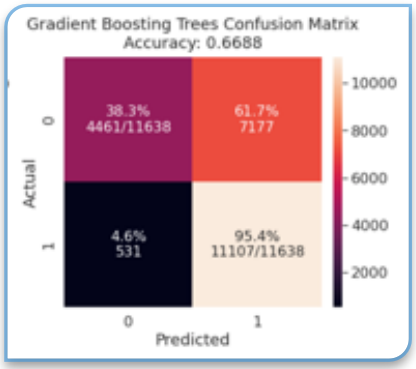
- Machine Learning Models used -
  - XGBoost
  - Gradient Boosting Trees
  - Decision Tree
  - Random Forest
  - Logistic Regression
- All provide good support binary classification

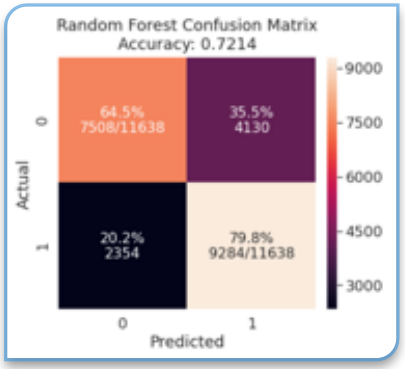
RESULTS DASHBOARD

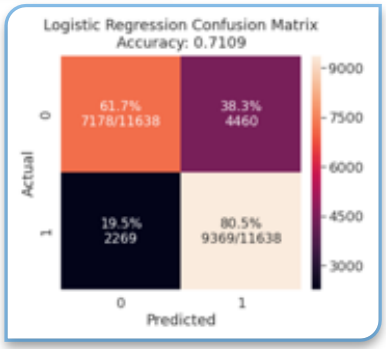
- XGBoost is best based on number of True Positives. GBT is 2nd
- XGBoost also has best Jaccard Score
- Random Forest was worst TP but overall best accuracy
- LR and DT honorable mention. Shows LRs flexibility for binary classification
- Overall Jaccard Score is very close for all models within .006 STD

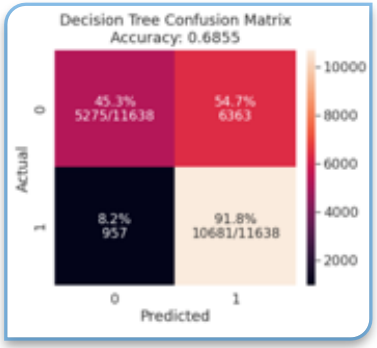
	Model	Accuracy	Jaccard	F1-Score	Log Loss	True -	False +	False -	True +
0	Gradient Boosting Trees	0.668843	0.590327	0.639449	0.597680	4461	7177	531	11107
1	XGBoost	0.680014	0.598685	0.653896	0.601270	4717	6921	527	11111
2	Random Forest	0.721430	0.588787	0.719798	0.528355	7508	4130	2354	9284
3	Decision Tree	0.685513	0.593356	0.667581	0.551461	5275	6363	957	10681
4	Logistic Regression	0.710904	0.581998	0.708319	0.534020	7178	4460	2269	9369





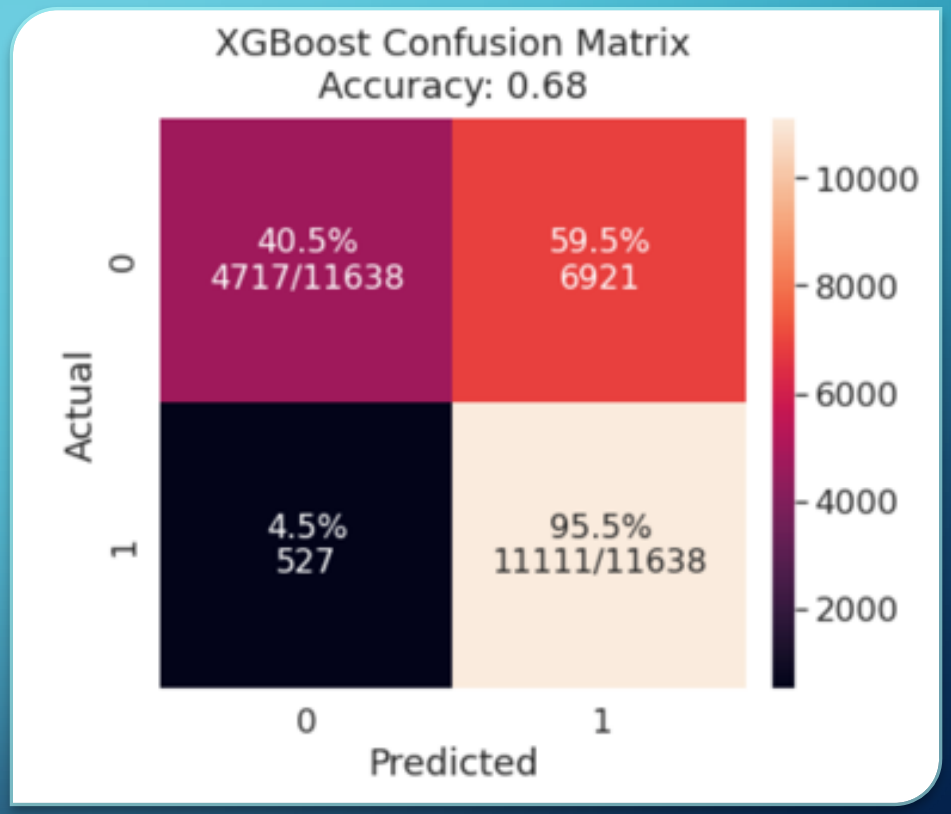






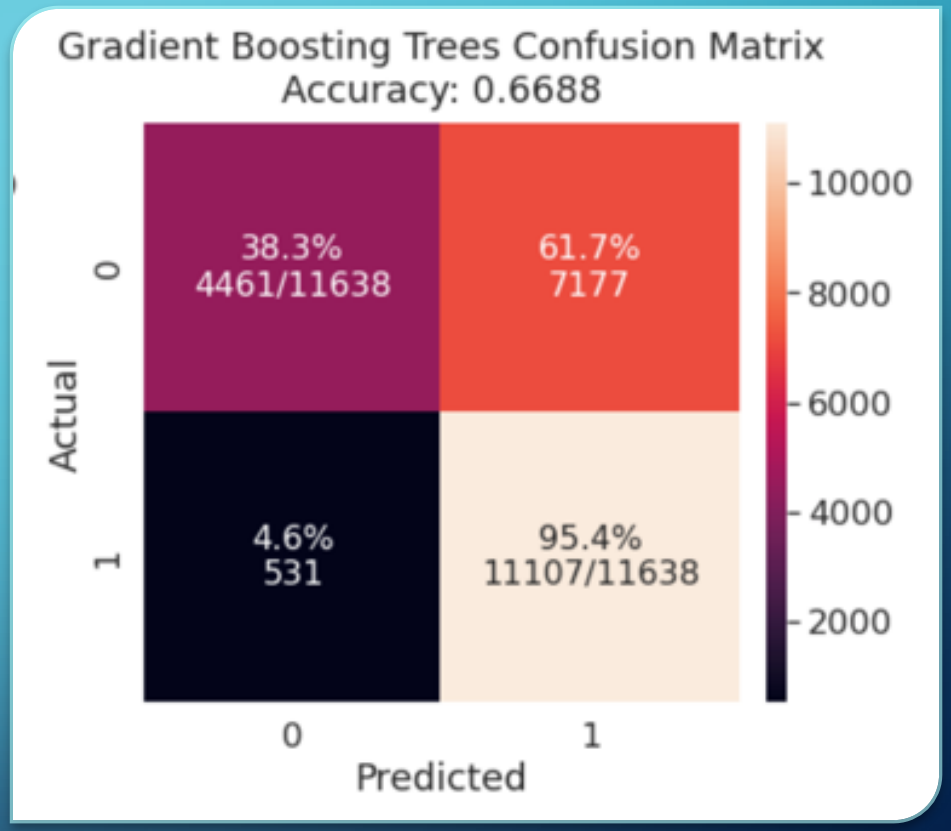
1ST PLACE XGBOOST

- General Accuracy - 0.680014
- Jaccard - 0.598685
- F1 - 0.653896
- Log Loss - 0.601270
- Top 5 Features
  - Collision Type = Parked Car
  - Collision Type = Angles
  - Collision Type = Pedestrian
  - Collection Type = Sideswipe
  - Road Condition = Unknown



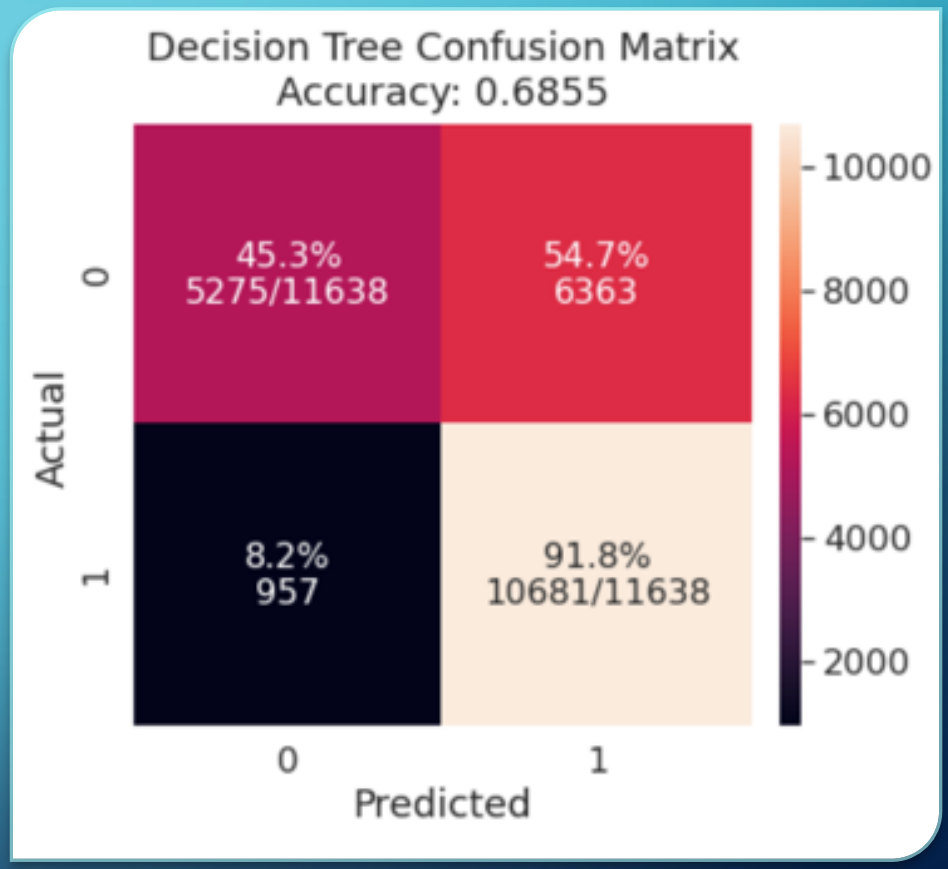
2ND PLACE GRADIENT BOOSTING TREES

- General Accuracy - 0.668843
- Jaccard - 0.590327
- F1 - 0.639449
- Log Loss - 0.597680
- Top 5 Features
  - Collision = Parked Car
  - Light Condition = Unknown
  - Junction Type = Intersection
  - Person Count
  - Collection Type = Angles



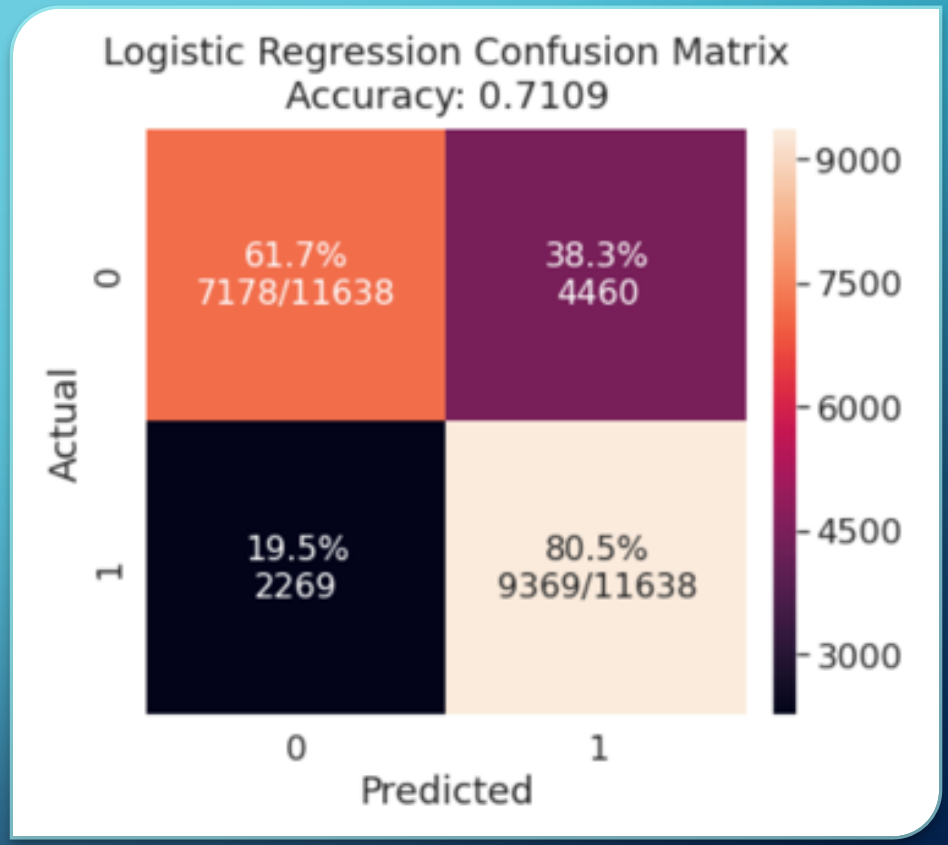
3RD PLACE DECISION TREE

- General Accuracy - 0.685513
- Jaccard - 0.593356
- F1 - 0.667581
- Log Loss - 0.551461
- Top 5 Features
  - Collision Type = Parked Car
  - Collision Type = Pedestrian
  - Collision Type = Angles
  - Collision Type = Sideswipe
  - Person Count



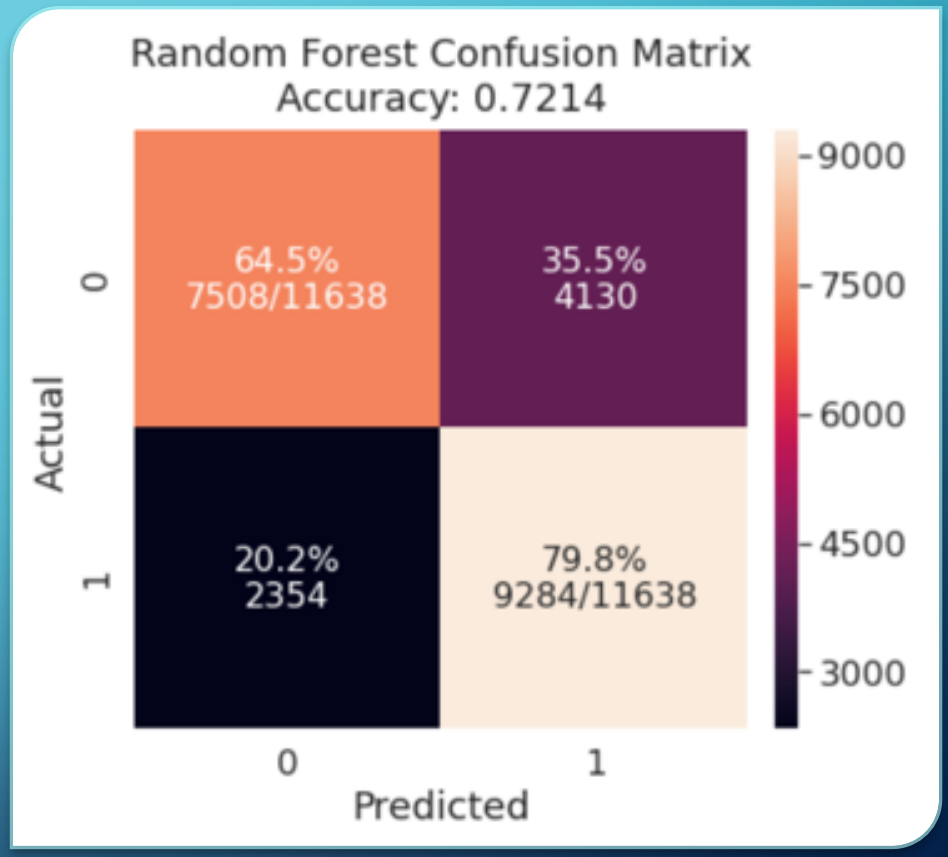
4TH PLACE
LOGISTIC REGRESSION

- General Accuracy - 0.710904
- Jaccard - 0.581998
- F1 - 0.708319
- Log Loss - 0.534020



5TH PLACE
RANDOM FOREST

- General Accuracy - 0.721430
- Jaccard - 0.588787
- F1 - 0.719798
- Log Loss - 0.528355
- Top 5 Features
  - Collision Type = Parked Car
  - SDOT Collision Code
  - Person Count
  - Collision Type = Pedestrian
  - Vehicle Count



CONCLUSION

- Using the current data it is possible to predict Injury Collisions with 95% accuracy using an XBT-based model
- There is little variation in accuracy between all models and no single model significantly standouts as the best in terms of accuracy. XGB and GBT where the best at predicting Injury Severity
- COLLISIONTYPE == Parked Car is the most significant feature. Followed by COLLISIONTYPE == Pedestrian, then ROADCOND == Unknown, COLLISIONTYPE == Pedestrian Not Granted Right of Way, PERSONCOUNT and VEHCOUNT and JUNCTIONTYPE == Intersection
- SDOT_COLCODE, SPEEDING, HITPARKEDCAR, INATTENTIONIND did not have a significant effect on results
- Excluded attributes including latitude (Y), longitude (X), INCDTTM. ST_COLCODE did not effect results and provided to be the right call to leave out
- Based on this study is it possible to offer a prediction service that could be used to provide external applications with predictive capabilities however more work needs to be done to improve the overall performance as well as collect more data including weather. Incorporating vehicle and driver details would also be helpful to improve the overall performance and yield better results