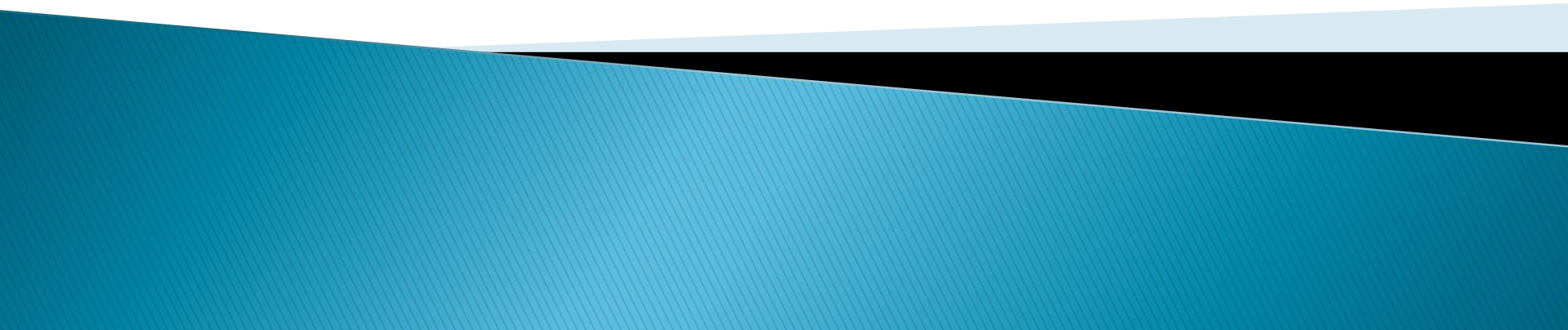


Accident Severity Coursera Capstone

By: Sumair Ajanee



Business problem

- ▶ Can accident severity be predicted?
- ▶ Applications:
 - City design
 - Insurance Pricing
- ▶ Will be taking the view of Insurer in this problem

Data Understanding

► Collision data provided by the SOPD

Column Name	Non Null Elements	Description	Used
SEVERITYCODE	194673	Target	Y
X	189339	Not used as it is not defined	N
Y	189339	Not used as it is not defined	N
OBJECTID	194673	SPD identifier key	N
INCKEY	194673	Incident key. Not giving much value	N
COLDETKEY	194673	Secondary key. Not much value	N
REPORTNO	194673	Not used as it is not defined	N
STATUS	194673	Not used as it is not defined	N
ADDRTYPE	192747	Collision type, like alley, block, intersection	Y
INTKEY	65070	Not used as it is too specific	N
LOCATION	191996	Not used as it is too general and not expandable	N
EXCEPTRSNCODE	84811	Not defined	N
EXCEPTRSNDESC	5638	Not defined	N
SEVERITYCODE.1	194673	Not defined	N

Data Understanding cont

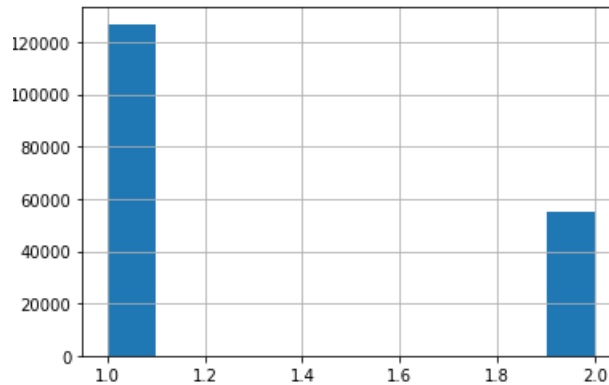
Column	Non_Null Values	Description	Used
COLLISIONTYPE	189769	Type of the collision	Y
PERSONCOUNT	194673	Number of people	Y
PEDCOUNT	194673	Number of pedestrians	Y
PEDCYLCOUNT	194673	Number of cyclists	Y
VEHCOUNT	194673	Number of vehicles	Y
INCDATE	194673	Date of the incident. Not used as feature	N
INCDTTM	194673	Date and time of the incident	N
JUNCTIONTYPE	188344	Not used as ti is not well defined	N
SDOT_COLCODE	194673	SOPD code not used	N
SDOT_COLDESC	194673	SOPD code not used	N
INATTENTIONIND	29805	Was the driver not paying attention	Y
UNDERINFL	189789	Weather the driver was under influence	Y
WEATHER	189592	Weather condition	Y
ROADCOND	189661	Road condition	Y
LIGHTCOND	189503	Lighting condition	Y
PEDROWNOTGRNT	4667	Did pedestrian have right of way	Y
SDOTCOLNUM	114936	SDOT id, not used	N
SPEEDING	9333	Was driver speeding	Y
ST_COLCODE	194655	SDOT code, not used	N
ST_COLDESC	189769	SDOT code, not used	N
SEGLANEKEY	194673	SDOT code, not used	N
CROSSWALKKEY	194673	Key for cross walk, not used as too specific	N
HITPARKEDCAR	194673	Hit a parked car, not used	N

Dealing with missing values

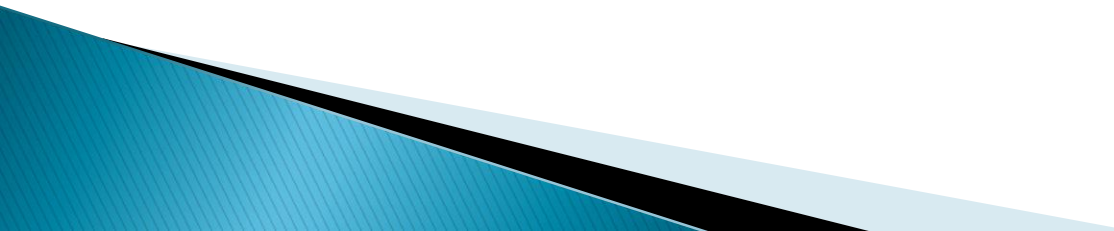
Column Name	Non Null Objects	How NA values are dealt with
SEVERITYCODE	194673	No NA values
ADDRTYPE	192747	Dropped
COLLISIONTYPE	189769	Dropped
PERSONCOUNT	194673	No NA values
PEDCOUNT	194673	No NA values
PEDCYLCOUNT	194673	No NA values
VEHCOUNT	194673	No NA values
INATTENTIONIND	29805	NA values assumed to be N as the only unique is Y
UNDERINFL	189789	NA values assumed to be N
WEATHER	189592	NA values defaulted to be clear
ROADCOND	189661	NA mapped to unknown road
LIGHTCOND	189503	NA values mapped to Daylight, as I will assume that is when most of the driving is doen
PEDROWNOTGRNT	4667	NA mapped to N as the only other is Y
SPEEDING	9333	NA mapped to N, as the only other is Y

Severity Balance

- ▶ Unbalanced data so models are evaluated based on F1 Score



Models Used

- ▶ K nearest neighbours
 - Looks for the closest type of data on a new set
 - ▶ Support vector
 - Tries to create a hyperplane
 - ▶ Decision Tree
 - Looks to separate classes based on features
 - ▶ Logistic regression
 - Determines probability using the sigmoid function
- 

Model evaluation

- ▶ Logistic regression and SVM show up as best model

Model	Jaccard Index	F1 Score	Log Loss
KNN	0.67	0.68	NA
Decision Tree	0.73	0.69	NA
SVM	0.74	0.70	NA
Logistic Regression	0.73	0.70	0.55

Conclusion

- ▶ Assumption made
 - Probabilities of certain features are known on the prior and can be used
 - May change based on geographic area
- ▶ Logistic regression recommended
 - Similar score to SVM, but has faster runtime