Predicting collisions severity and probability

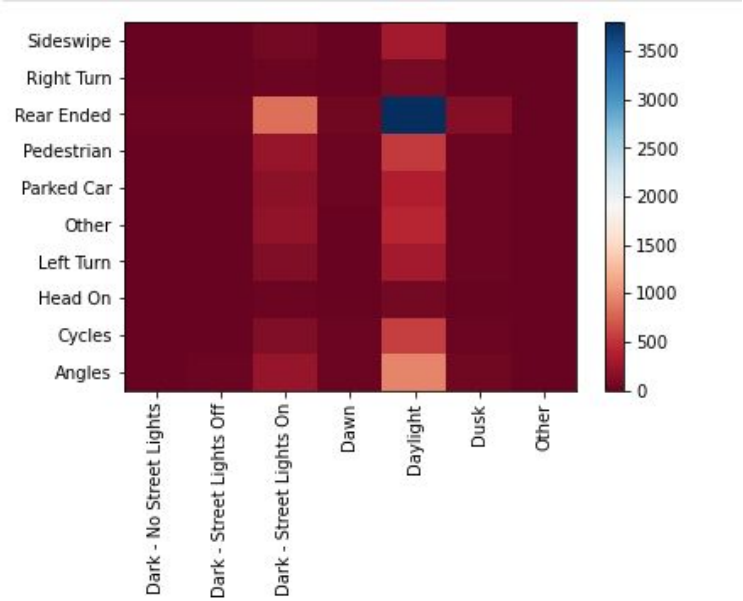
Predicting collision severity and probability is valuable for:

- Transportation and logistics companies
- Travellers
- Taxi drivers
- This prediction could be used in driver support systems to help prevent accidents and reduce number of people's deaths

Data acquisition and cleaning

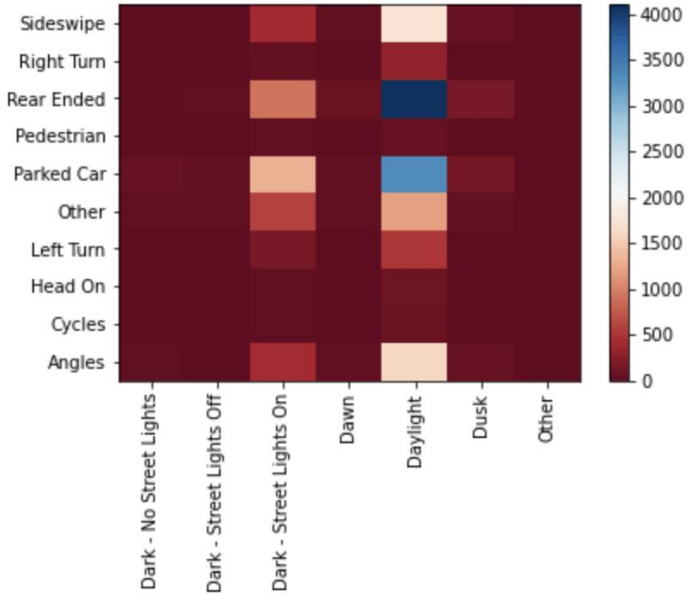
- The dataset in this project has been taken from coursera cloud storage (Data-Collisions.csv)
- Dataset has contains data in term 2004/01/01 to 2020/05/20
- The instant dataset included 192031 rows and 53 features
- Duplicate, highly similar, features with null values were dropped
- Cleaned data contains 16 features

Collisions with injures



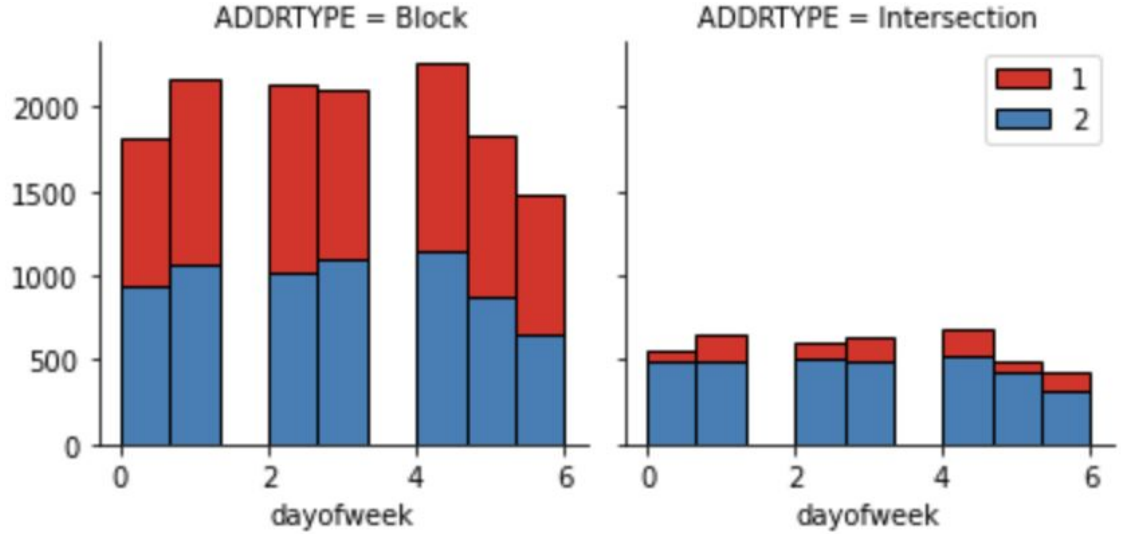
Most of collisions with injuries (severity code - 2) occure at daylight, clear weather, dry road on mid-block road with rear ended type of collision and without influence of drugs or alcohol ($p<0.001$).

Collisions with property damage



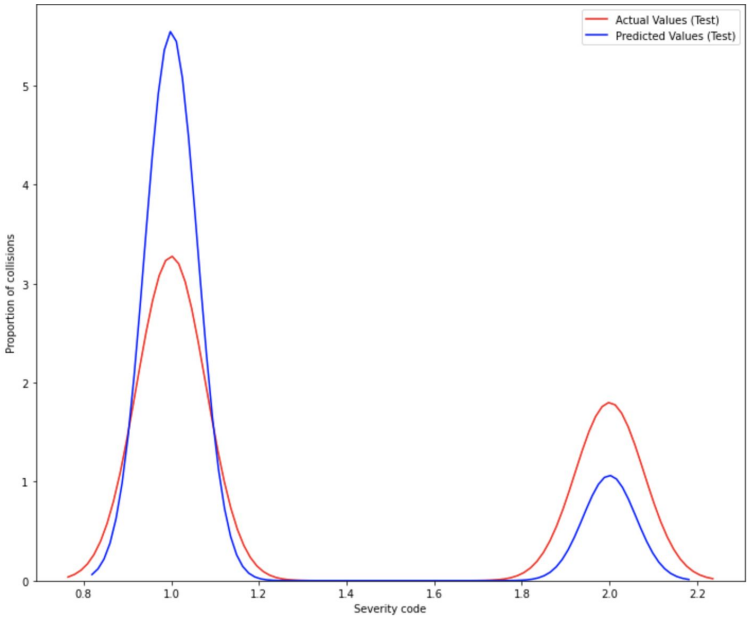
Most of collisions with property damage (severity code - 1) occure at daylight, clear weather, dry road on mid-block road with rear ended and parked car types of collision, without influence of drugs or alcohol ($p<0.001$).

Collisions distribution by days of week and address type of collision

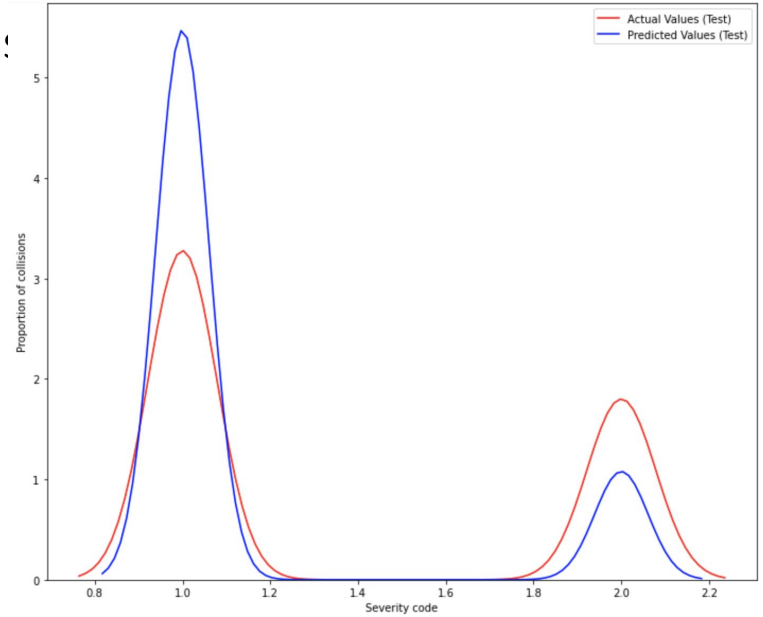


1 - property damage
2 - injuries
Also examined other features and hypotheses, including:
○ The number of people involved in a collision (significant correlation with severity code)
○ The number of vehicles (significant correlation)
○ The number of bicycles (significant correlation)
○ pedestrians count (strong correlation)

Classification models: distribution of predicted and real data

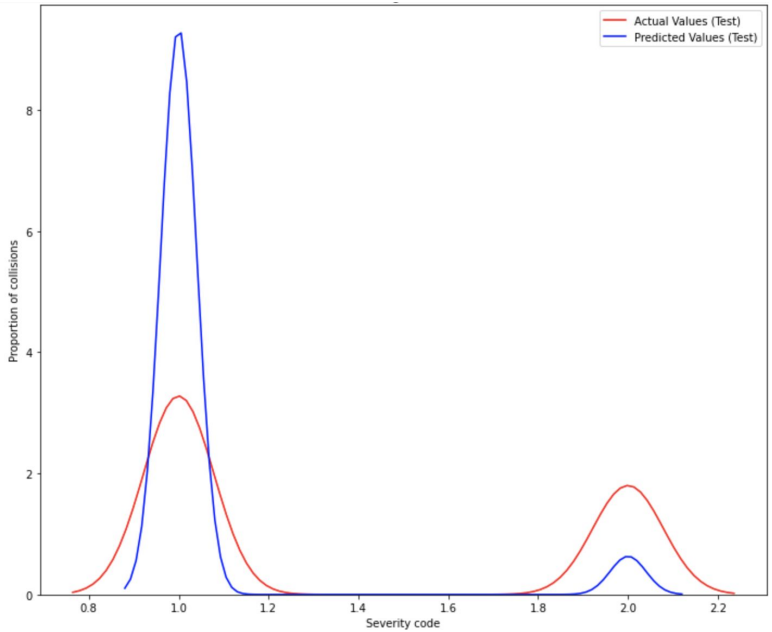


SVM

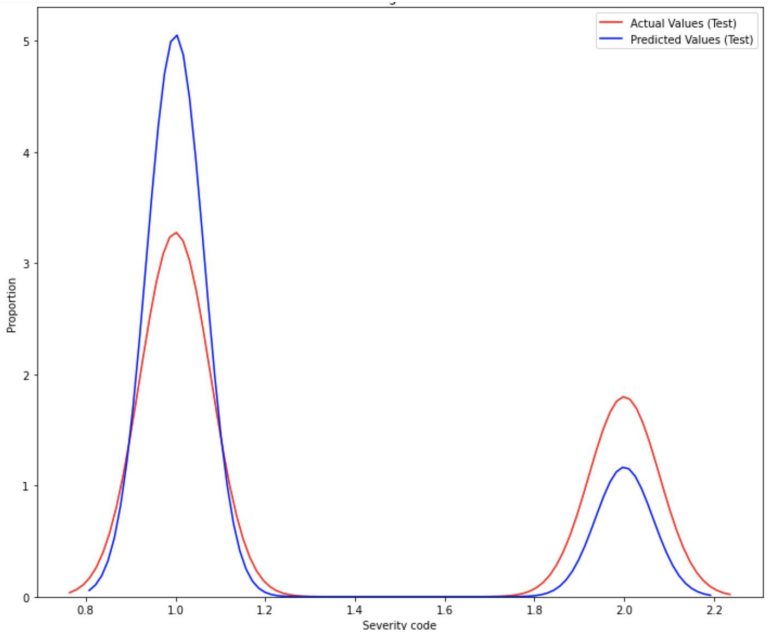


Logistic Regression

Classification models: distribution of predicted and real data



Decision Tree



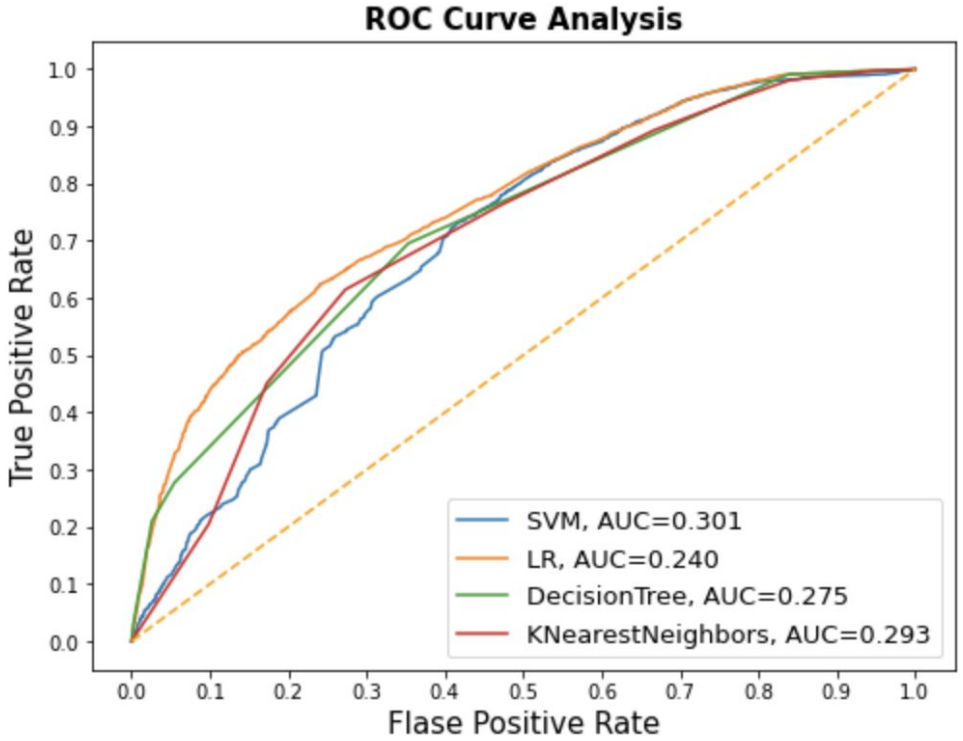
K-Nearest neighbors

Classification models performance

---	SVM	Logistic Regression	Decision Tree	K-nearest neighbors
F1-score	0.68	0.68	0.62	0.67

---	SVM	Logistic Regression	Decision Tree	K-nearest neighbors
Log Loss	0.59	0.54	0.57	1.77
Jaccard Similarity Index	0.71	0.71	0.70	0.69
No. of True Negatives	5012	4994	5356	4827
No. of False Positives	393	411	49	578
No. of False Negatives	2015	2002	2488	1978
No. of True Positives	953	966	480	990

ROC curve analysis



SVM and Logistic regression have very similar performance, other models have good performance as well. As seen from ROC curve analysis Logistic regression is the best model for prediction of collision probability (AUC=0.24)

Conclusion and future directions

- Built useful models to predict collisions severity and probability
- Accuracy of the models is quite high
- Capture more features
- Ideas include: car makes, car conditions, manufacturing date of a car