PREDICTING ACCIDENT SEVERITY

AUTO FACTS

- approximately 17 thousand traffic crashes each day nationwide
- around 36,560 motor vehicle deaths in 2018
- Incurred losses for auto insurer has been climbing steadily

Incurred Losses For Auto Insurance, 2015-2019 (1)
($000)

	2015	2016	2017	2018	2019
Private passenger auto					
Liability	$79,098,617	$88,249,238	$90,495,835	$91,736,331	$96,189,924
Physical damage	48,564,511	55,738,221	57,052,411	58,766,743	62,637,686
Commercial auto					
Liability	13,587,152	14,987,073	15,528,570	17,810,709	20,434,568
Physical damage	3,902,124	4,279,414	4,874,748	4,999,100	5,407,130
Total	**$145,152,404**	**$163,253,946**	**$167,951,564**	**$173,312,883**	**$184,669,308**

PROJECT: PREDICTING ACCIDENT SEVERITY

- there are a lot of attributes that could cause a server car crush
- accurately predicting the possibility of getting into a car accident and how severe it would be
- potentially change people's driving attitude and behaviors
- Goal:

Reduce cost for auto insurers + Reduce fatalities rate

DATA SOURCE

- Seattle traffic collisions data
- provided by SPD and recorded by Traffic Records
- 2004 – present
- https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

DATA WRANGLING

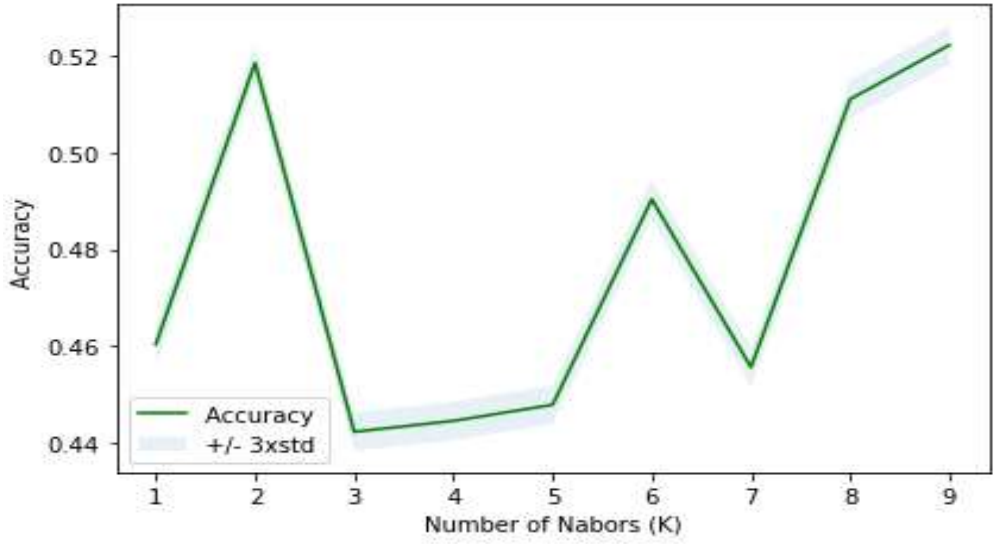
- This project aims to predict accident severity base on objective features: **collision address, weather conditions, road conditions, and light conditions**.
- Cleaned data contains features: inattention, drugs or alcohol influence, speeding
- Dropped missing values
- Use under-sampling method to balance data
- In total, 39,899 rows and 5 columns in modified dataset.

BUILD MODELS

- Classification algorithms
- 4 various models
- KNN, Decision Tree, SVM, Logistic Regression
- Find the one with the highest accuracies
- Apply the same split set of training and testing data

MODEL 1: K-NEAREST NEIGHBORS

- Find the correct K value
- Set the range of K from 1 to 10. Repeated the process by increasing K to find the highest accuracy




```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=None, n_neighbors=9, p=2,
           weights='uniform')
```

MODEL 2: DECISION TREE

- set the depth value as 8 to build the model.

```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=8,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=1,
            splitter='best')
```

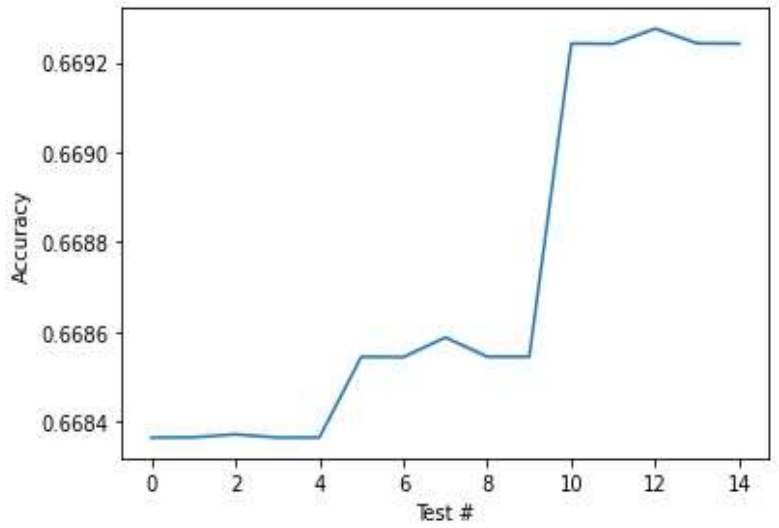
MODEL 3: SUPPORT VECTOR MACHINE

- Four types of kernelling in SVM model: linear, polynomial, rbf, Sigmoid
- Ran each model to find which kernel would build the model with highest accuracies
- polynomial kernel !

```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
  kernel='poly', max_iter=-1, probability=False, random_state=None,
  shrinking=True, tol=0.001, verbose=False)
```

MODEL 4: LOGISTIC REGRESSION

- 12 different tests with different c value and solver
- accuracy graph

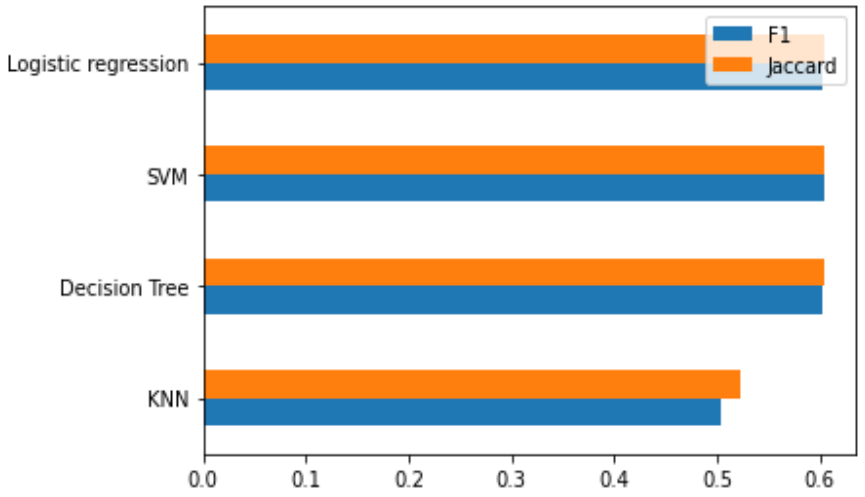



```
LogisticRegression(C=0.001, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='liblinear',
          tol=0.0001, verbose=0, warm_start=False)
```

F1-SCORE & JACCARD INDEX

- Log loss - logistic regression = 0.669.

	F1	Jaccard
KNN	0.5040	0.52350
Decision Tree	0.6028	0.60457
SVM	0.6035	0.60526
Logistic regression	0.6030	0.60470



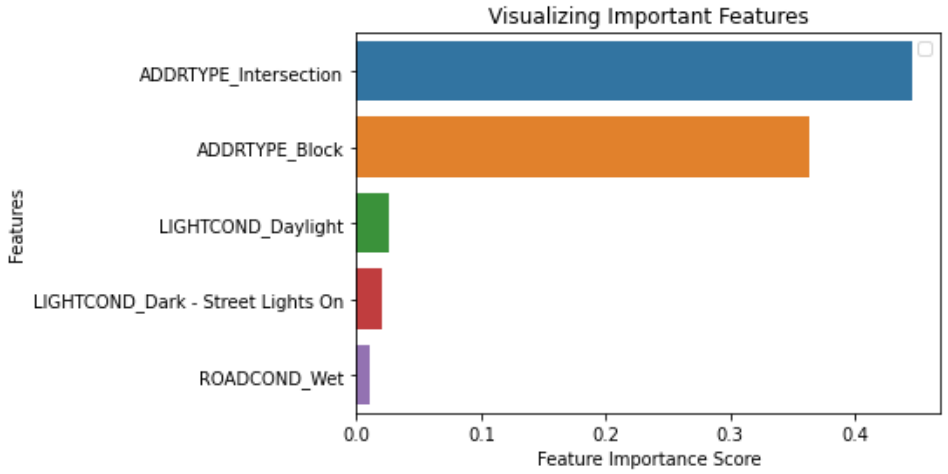
RESULT

- Pick logistic regression to build the model
- Predict the probability of an unknown collision belonging to severity 1 or 2

```
LogisticRegression(C=0.001, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='liblinear',
          tol=0.0001, verbose=0, warm_start=False)
```

TOP 5 WEIGHTED FEATURES

- "Intersection" and "Block" are the most important features of predicting a car accident, far ahead from the rest
- Drivers should pay extra attention when they are driving close to intersection or block. Speed detectors or warning signs could be placed to reduce cars' speed.



IMPROVEMENT OF THE MODEL

- logistic regression model has accuracy of 0.669
- For more accurate prediction, the dataset can be added more detailed features like gender, age, visibility, etc.

REFERENCE DATA

- fact+statistics: auto insurance
- Https://www.iii.org/fact-statistic/facts-statistics-auto-insurance
- Wiki
- https://en.wikipedia.org/wiki/Motor_vehicle_fatality_rate_in_U.S._by_year