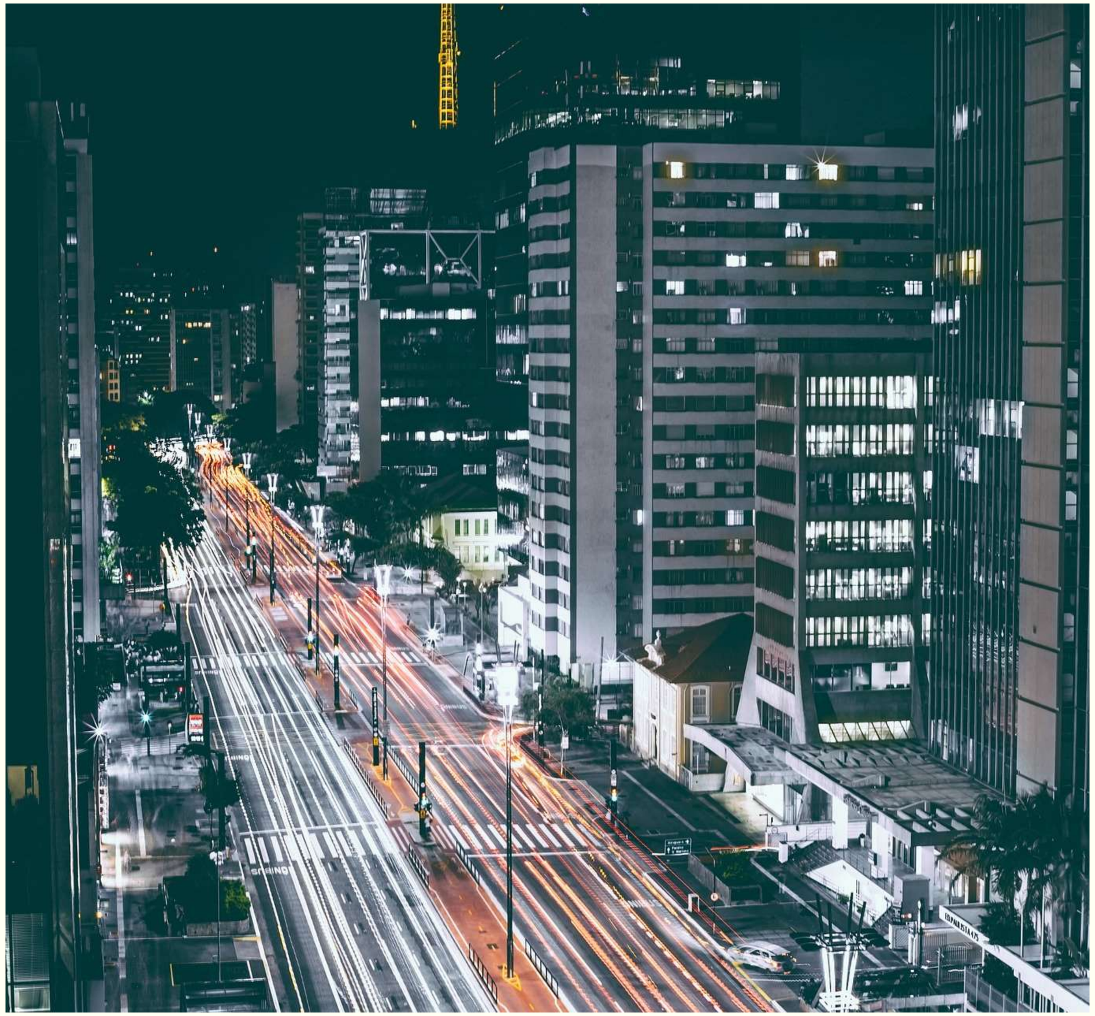
COURSERA CAPSTONE PROJECT

Car Accident Severity

By Michael Fretwell
2 September 2020

Table of Contents



- Introduction
- Data
- Methodology
- Results
- Discussion / Conclusion

Introduction

- According to the Association for Safe International Road Travel (ASIRD) approximately 38,000 people die in road crashes in the United States each year.
- The fatality rate is 12.4 deaths per 100,000 inhabitants.
- An additional 4.4 million people are injured seriously enough to require medical attention.
- Road crashes are the leading cause of death in the U.S. for people aged 1-54.
- The ability to leverage available data to make predictions related to automobile accidents would be valuable to a wide range of industry groups.

Introduction (Cont.)

- Scope of project includes:
  - Identification of factors (such as speed, time of day, and weather conditions) that contribute to accidents.
  - Build regression classification models to predict severity of accidents with out of sample data.
- Use of regression classification modeling provides for measures of probability, given various conditions, that can be considered as risk of accidents.
- This information will be useful to those in the traffic safety industry for the purpose of implementing actions to reduce risks of accidents.
- Information will also be of use to the insurance industry.

Data

- Source data is the Seattle “Collisions – All Years” dataset produced by the Seattle SDOT Traffic Management Division, Traffic Records Group.
- Point of contact for the data is the SDOT GIS Analyst at DOT_IT_GIS@seattle.gov.
- Period covered by the data is January 2004 to 19 May 2020.
- There are 194,673 observations and 37 attributes.

Methodology

- Duplicate or unnecessary features were removed.

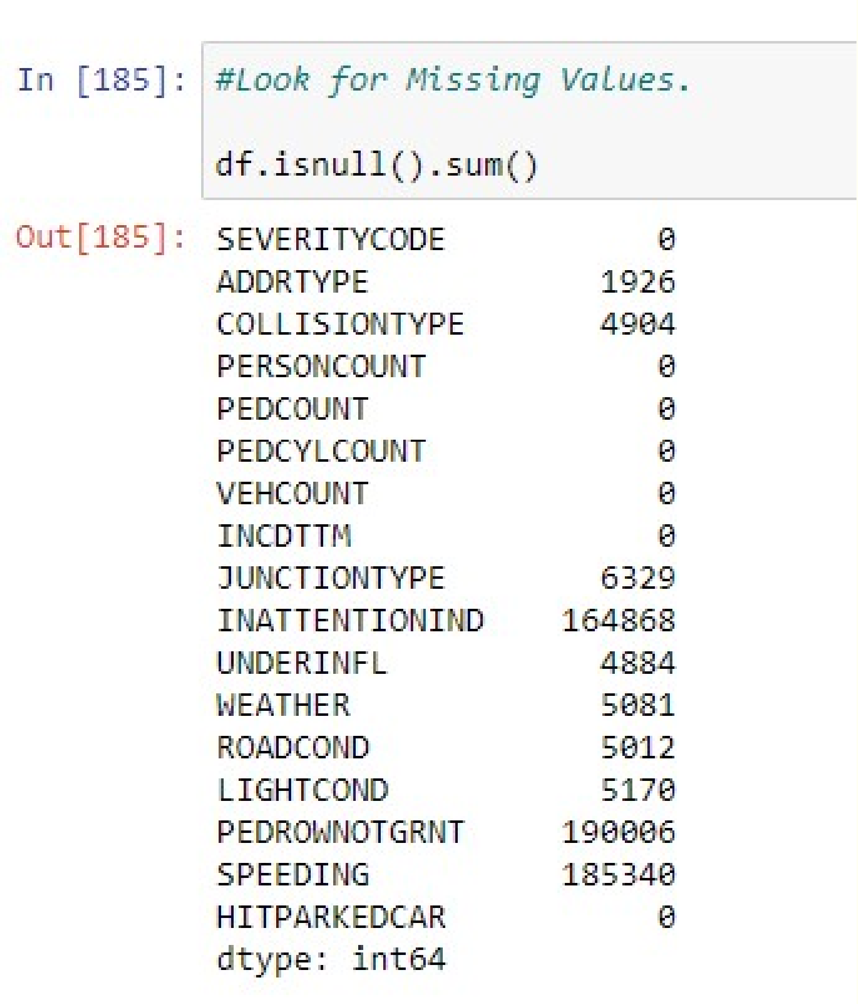
X	Y	INCKEY
COLDETKEY	REPORTNO	STATUS
SDOTCOLNUM	LOCATION	SEVERITYCODE.1
SEVERITYDESC	ST_COLDESC	EXCEPTRSNCODE
EXCEPTRSNDESC	ST_COLCODE	SDOT_COLCODE
SDOT_COLDESC	INTKEY	INCDATE
OBJECTID	SEGLANEKEY	CROSSWALKKEY

- The label , or target, for models is “SEVERITYCODE” which indicates the severity of the collision (“1” for property damage or “2” for injury).

- Label contains unbalanced data which required balancing to preclude bias in the model.

Methodology (Cont.)

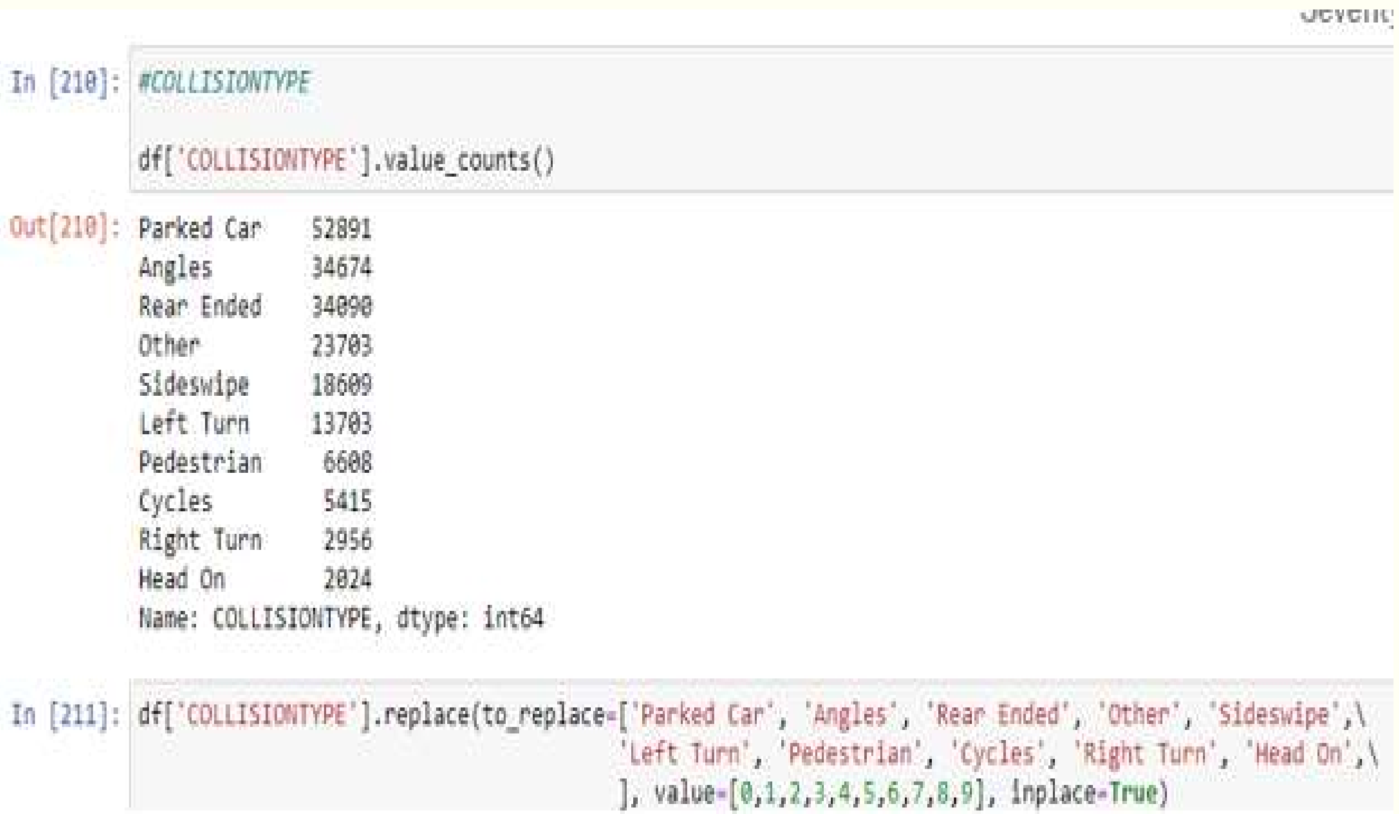
- Attributes with missing values were identified.
- Analysis of each attribute was done to determine appropriate steps to resolve missing information.
- Replaced missing information with the most frequently occurring value.
- Replaced null values with “0”.



```
In [185]: #Look for Missing Values.

          df.isnull().sum()

Out[185]: SEVERITYCODE            0
          ADDRTYPE             1926
          COLLISIONTYPE        4904
          PERSONCOUNT             0
          PEDCOUNT                0
          PEDCYLCOUNT             0
          VEHCOUNT                0
          INCDTTM                 0
          JUNCTIONTYPE         6329
          INATTENTIONIND     164868
          UNDERINFL            4884
          WEATHER              5081
          ROADCOND             5012
          LIGHTCOND            5170
          PEDROWNOTGRNT      190006
          SPEEDING           185340
          HITPARKEDCAR            0
          dtype: int64
```

Methodology (Cont.)

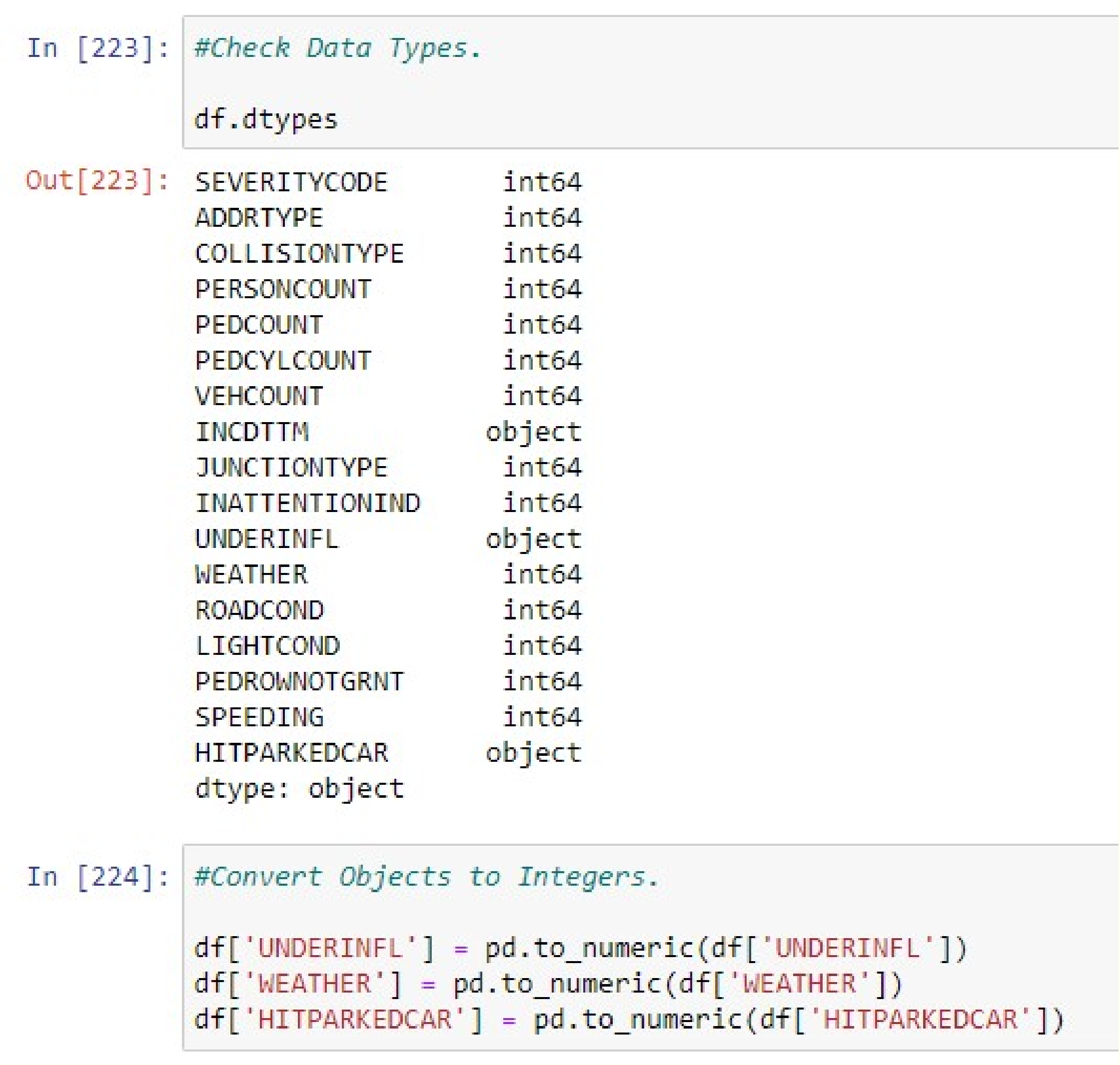
- Categorical features were converted to numbers.



```
In [210]: #COLLISIONTYPE

          df['COLLISIONTYPE'].value_counts()

Out[210]: Parked Car     52891
          Angles         34674
          Rear Ended     34090
          Other          23703
          Sideswipe      18609
          Left Turn      13703
          Pedestrian      6608
          Cycles          5415
          Right Turn      2956
          Head On         2024
          Name: COLLISIONTYPE, dtype: int64

In [211]: df['COLLISIONTYPE'].replace(to_replace=['Parked Car', 'Angles', 'Rear Ended', 'Other', 'Sideswipe',\
                                                  'Left Turn', 'Pedestrian', 'Cycles', 'Right Turn', 'Head On',\
                                                  ], value=[0,1,2,3,4,5,6,7,8,9], inplace=True)
```

Methodology (Cont.)

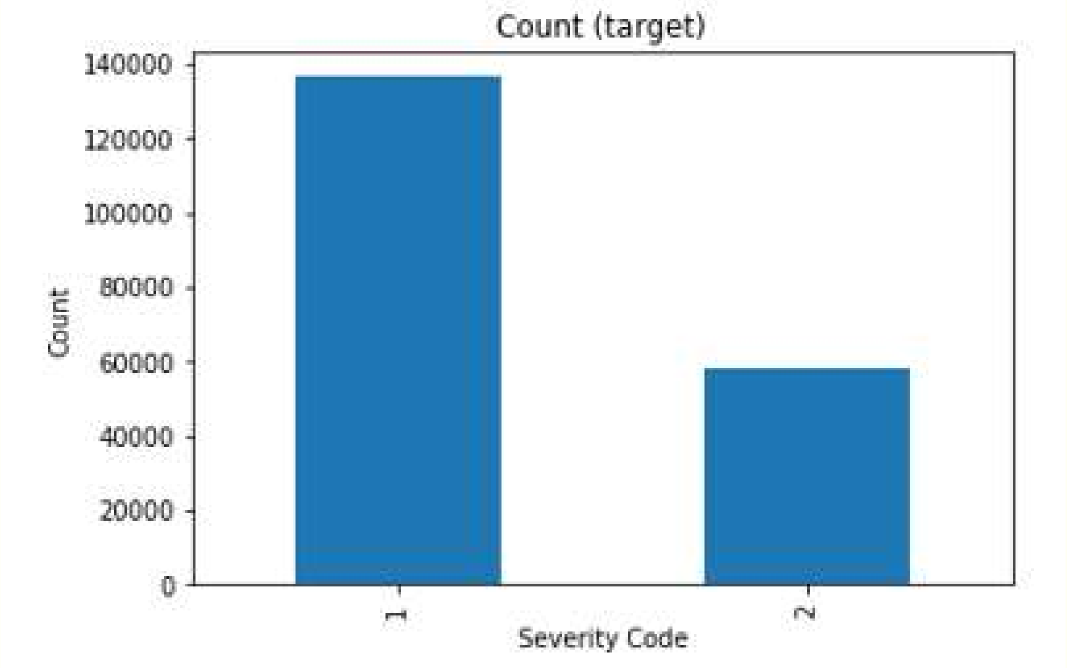
- Object data types were converted to integers.



```
In [223]: #Check Data Types.

          df.dtypes

Out[223]: SEVERITYCODE        int64
          ADDRTYPE            int64
          COLLISIONTYPE       int64
          PERSONCOUNT         int64
          PEDCOUNT            int64
          PEDCYLCOUNT         int64
          VEHCOUNT            int64
          INCDTTM            object
          JUNCTIONTYPE        int64
          INATTENTIONIND      int64
          UNDERINFL          object
          WEATHER             int64
          ROADCOND            int64
          LIGHTCOND           int64
          PEDROWNOTGRNT       int64
          SPEEDING            int64
          HITPARKEDCAR       object
          dtype: object

In [224]: #Convert Objects to Integers.

          df['UNDERINFL'] = pd.to_numeric(df['UNDERINFL'])
          df['WEATHER'] = pd.to_numeric(df['WEATHER'])
          df['HITPARKEDCAR'] = pd.to_numeric(df['HITPARKEDCAR'])
```

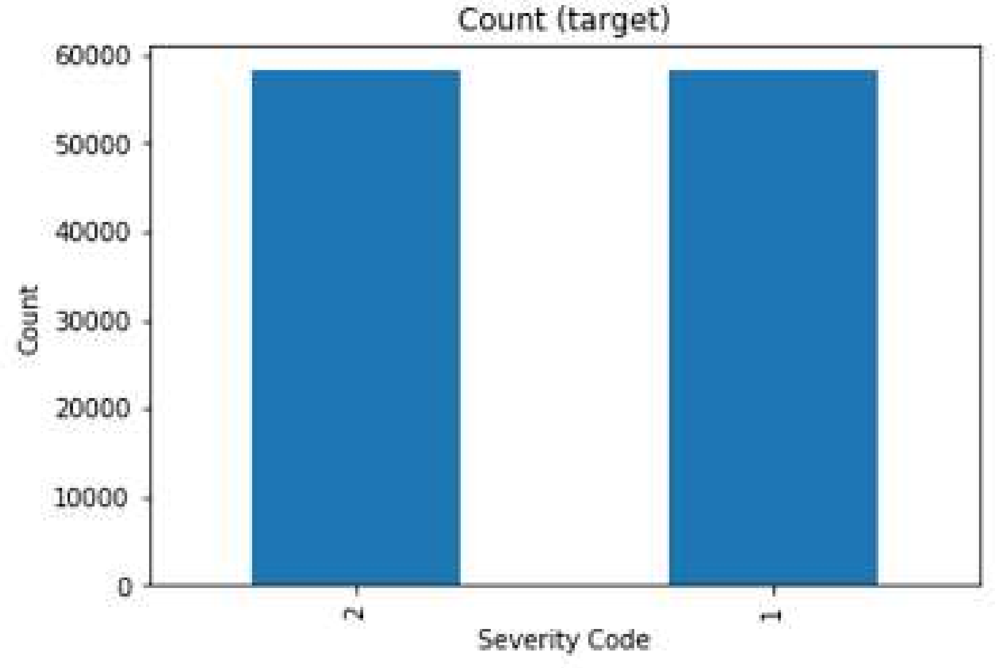
Methodology (Cont.)

- Target for models was Severity Code.
  - 1 – Property Damage
  - 2 – Injury
- Target data was unbalanced.
- Would create bias in models if not adjusted.



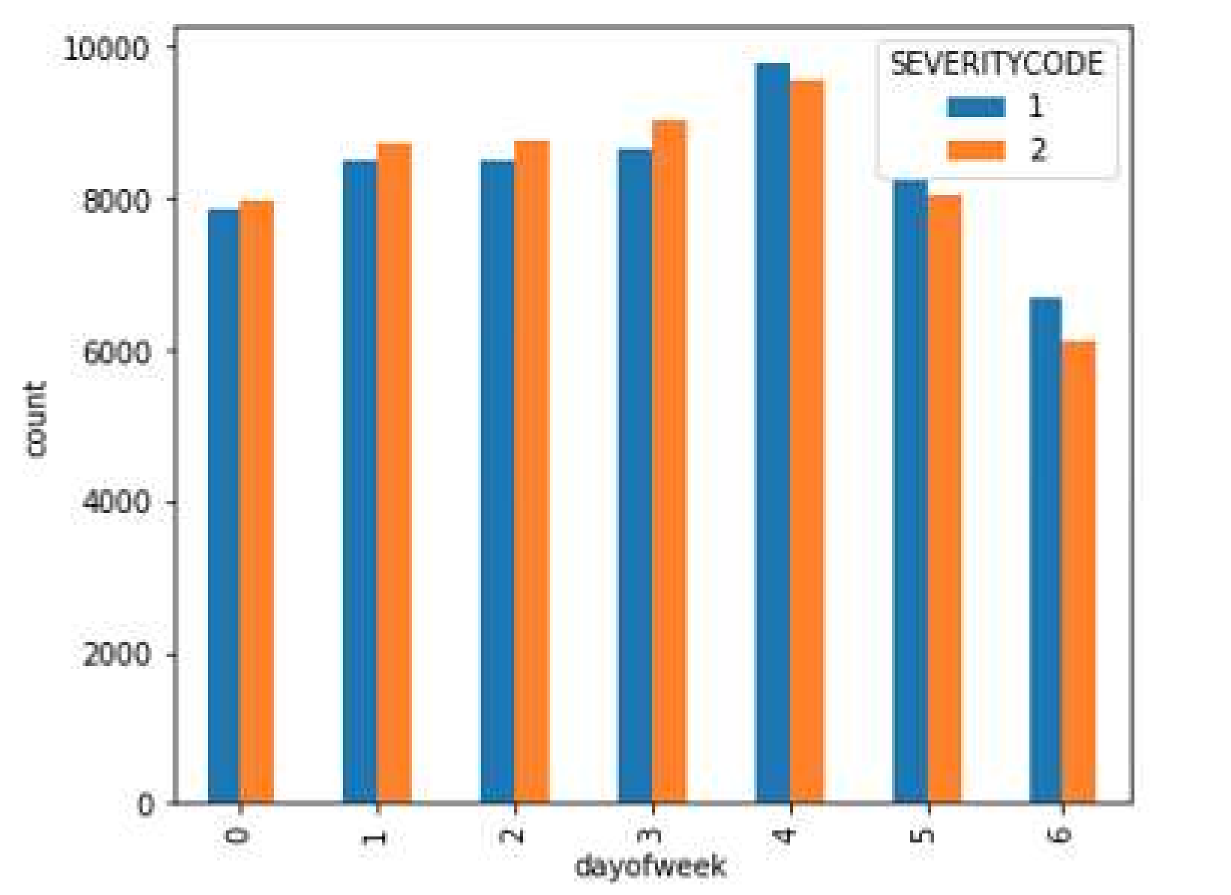
Methodology (Cont.)

- Steps were taken to balance the dataset.
- Balanced dataset consists of 116,376 records.



Methodology (Cont.)

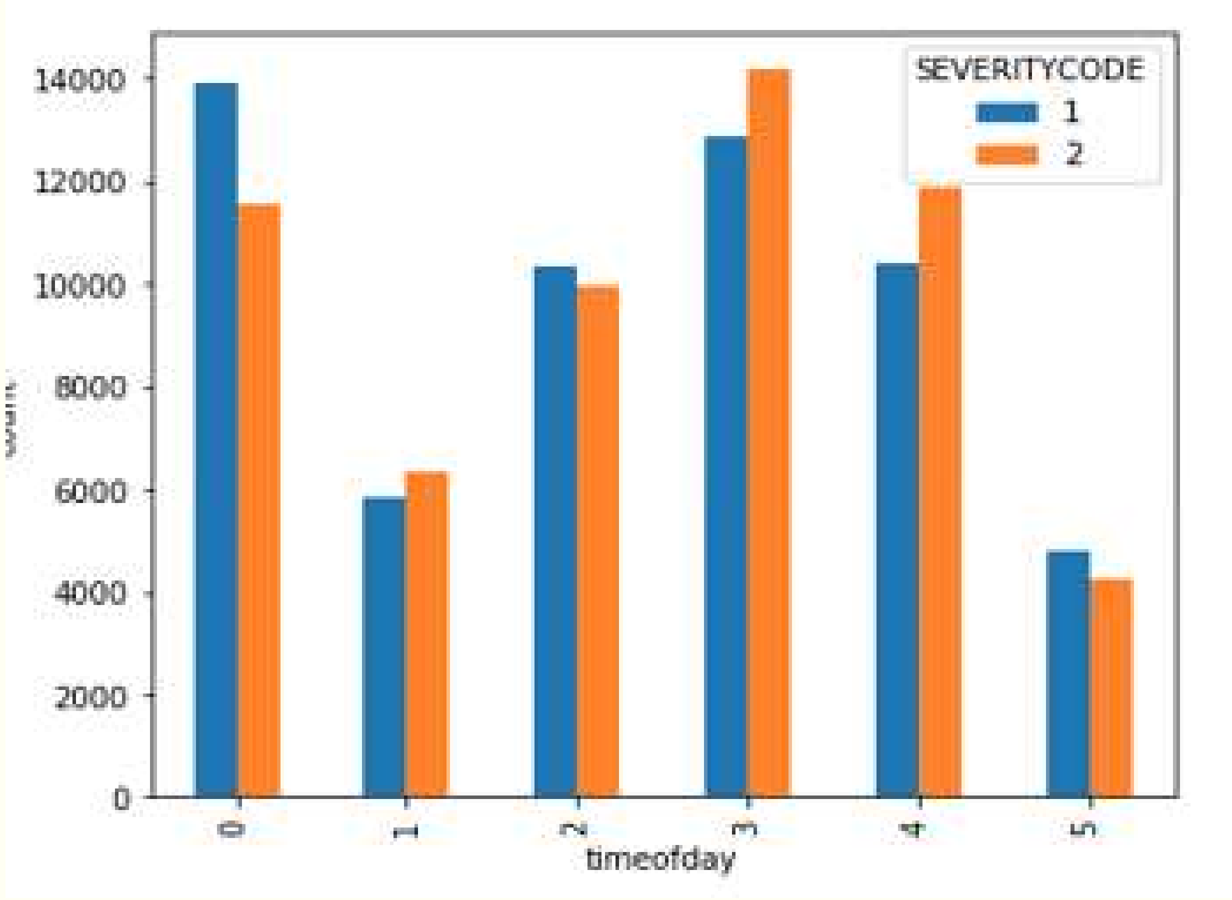
- Incident date time feature (INCDTTM) was an object.
- Conversion to date time format allowed for calculations.
- Day of week analysis shows mostly minor variations in accidents by day.
- Noticeable drop in number of accidents on Saturdays.



Methodology (Cont.)

- Time of day grouped into 6 bins for analysis.
- Fewer number of accidents during the periods 0400-0800 and 2000-0000.

Time of Day	Bin
0000 to 0400	0
0400 to 0800	1
0800 to 1200	2
1200 to 1600	3
1600 to 2000	4
2000 to 0000	5



Methodology (Cont.)

- 15 features were selected for use in the models from the initial attributes.
- 2 additional features were derived from the INCDTTM field:
  - dayofweek
  - timeofday

ADDRTYPE	COLLISIONTYPE	PERSONCOUNT
PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
JUNCTIONTYPE	INATTENTIONIND	UNDERINFL
WEATHER	ROADCOND	LIGHTCOND
PEDROWNOTGRNT	SPEEDING	HITPARKEDCAR
dayofweek	timeofday	

Methodology (Cont.)

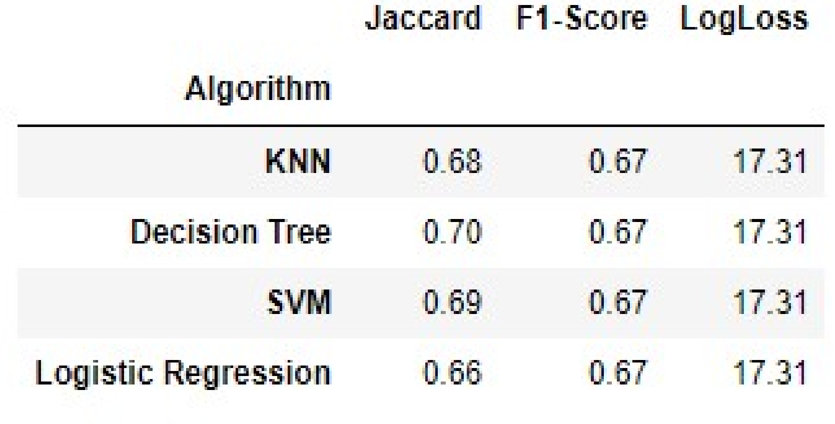
- Objective of models is to predict severity of accidents; thus classification models were used.
- K Nearest Neighbor
- Decision Tree
- Support Vector Machine
- Logistic Regression

Methodology (Cont.)

- Since all data was manually converted to numerical, uncertain if sklearn preprocessing to further standardize data would have an impact.
- Models were run with and without preprocessing – variances were minor.
- Data was split into Train and Test groups
  - Train set: (93,100, 17) (93,100,)
  - Test set: (23,273, 17) (23,276,)

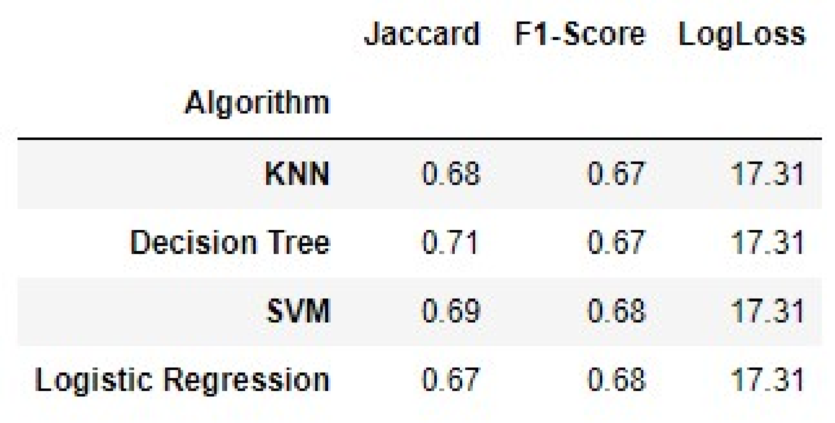
Results

- Accuracy of models was evaluated using sklearn metrics library:
  - Jaccard
  - F-1
  - LogLoss
- Models were run with and without sklearn preprocessing resulting in minor differences.
- In both cases, decision tree model produced slightly better scores.



	Jaccard	F1-Score	LogLoss
Algorithm			
KNN	0.68	0.67	17.31
Decision Tree	0.70	0.67	17.31
SVM	0.69	0.67	17.31
Logistic Regression	0.66	0.67	17.31

Scores Without Preprocessing

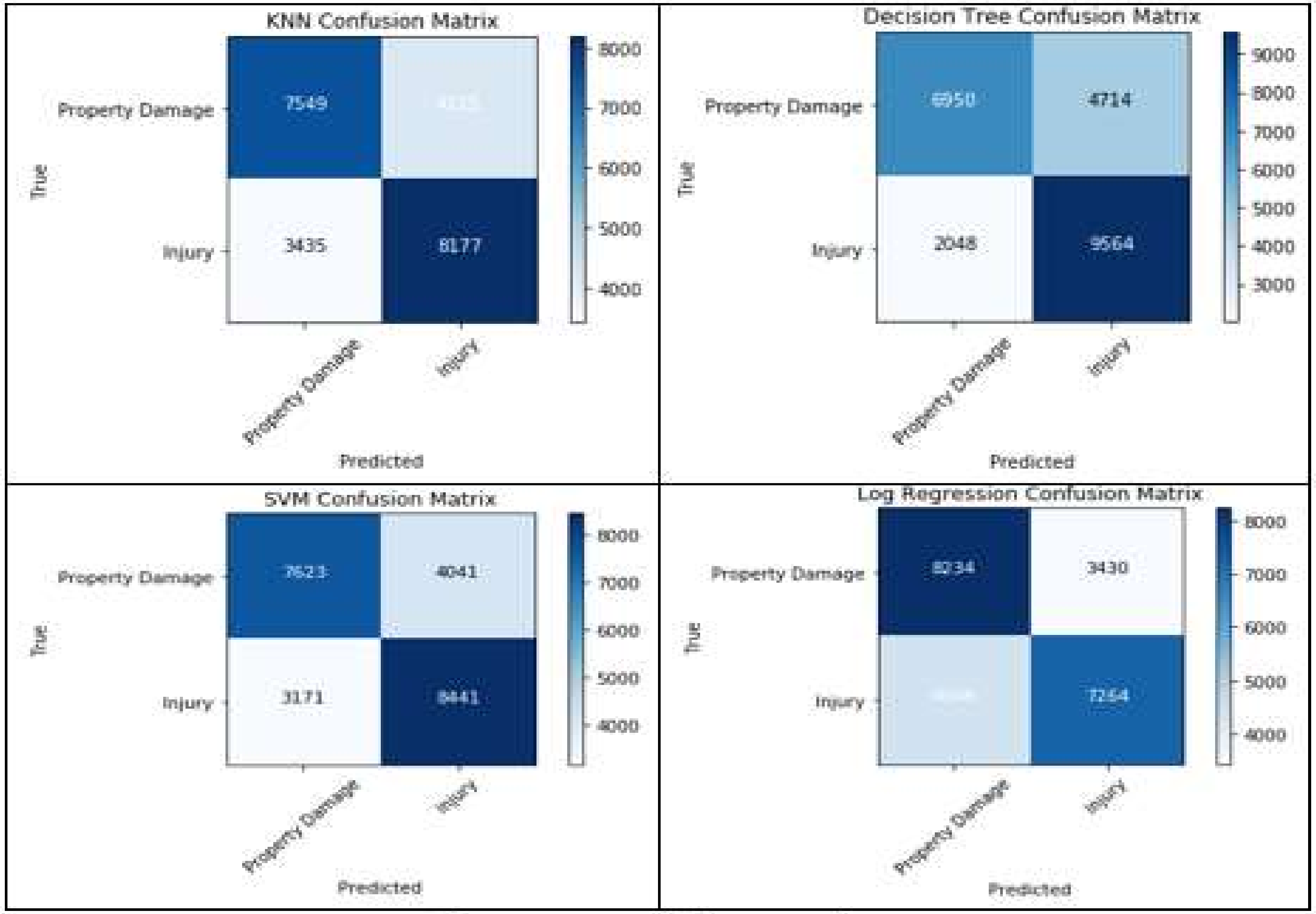


	Jaccard	F1-Score	LogLoss
Algorithm			
KNN	0.68	0.67	17.31
Decision Tree	0.71	0.67	17.31
SVM	0.69	0.68	17.31
Logistic Regression	0.67	0.68	17.31

Scores With Preprocessing

Results (Cont.)

- Confusion matrices for each model.
- Decision tree best at predicting accidents with injuries.
- Logistics Regression best at predicting accidents with property damage.



Discussion / Conclusion

- Using Seattle accident information dataset we processed the data and built classification models that predicting accident severity with reasonable accuracy.
- We observed a noticeable drop in number of accidents occurring on Saturdays.
- We also noted significantly fewer accidents occurred during the time periods 0400-0800 and 2000-0000.
- There is room for improvement in accuracy of models.
- Future efforts could perform more feature evaluation to determine best features for use in models.
- Additional datasets could be added to include information such as nationwide data and population information.
- Categorical target data could be expanded to additional classifications such as fatality information.