Applied Data Science Capstone

Car Accident Severity

Aug 2020

Problem Statement

- 20-30 million people involved in a road accident
- 10% lose their life

- Old accident record data for Seattle along with accident severity

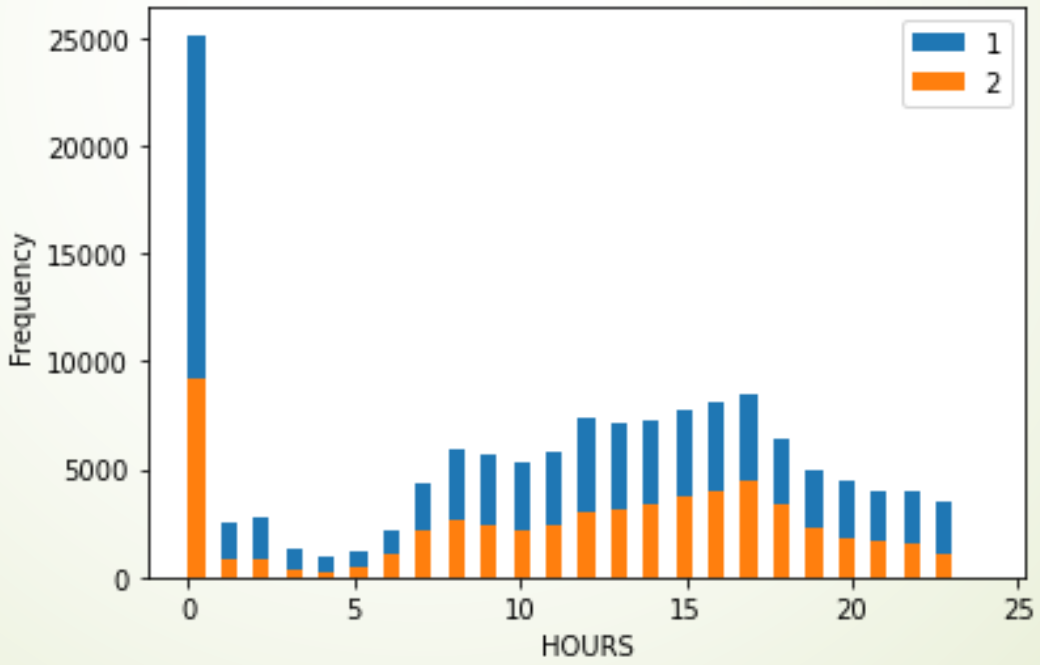
- GOAL:
- Develop a machine learning model to predict the future severity of the accident

Data Analysis

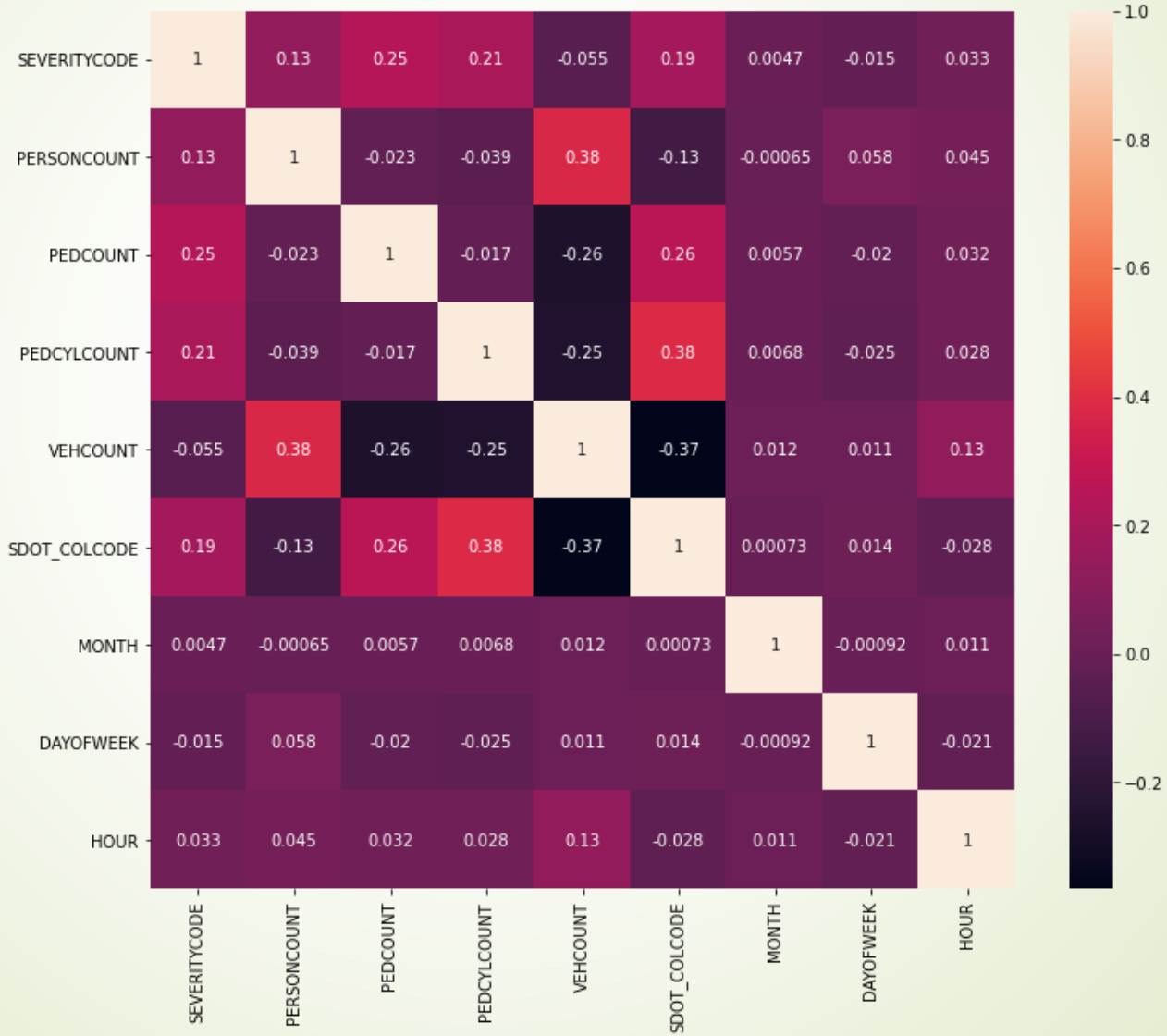
- Shape: (194673, 38)
- If data is balanced: 136485 rows for type 1 compare to 58188 rows for type 2
- Selected Features:
- 'SEVERITYCODE','ADDRTYPE','PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','VEHCOUNT', 'INCDTTM','SDOT_COLCODE','INATTENTIONIND', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SPEEDING', 'ST_COLCODE'
- Fill Nan values with Mode
- Change Categorical Data to Dummy Variables

Visualization

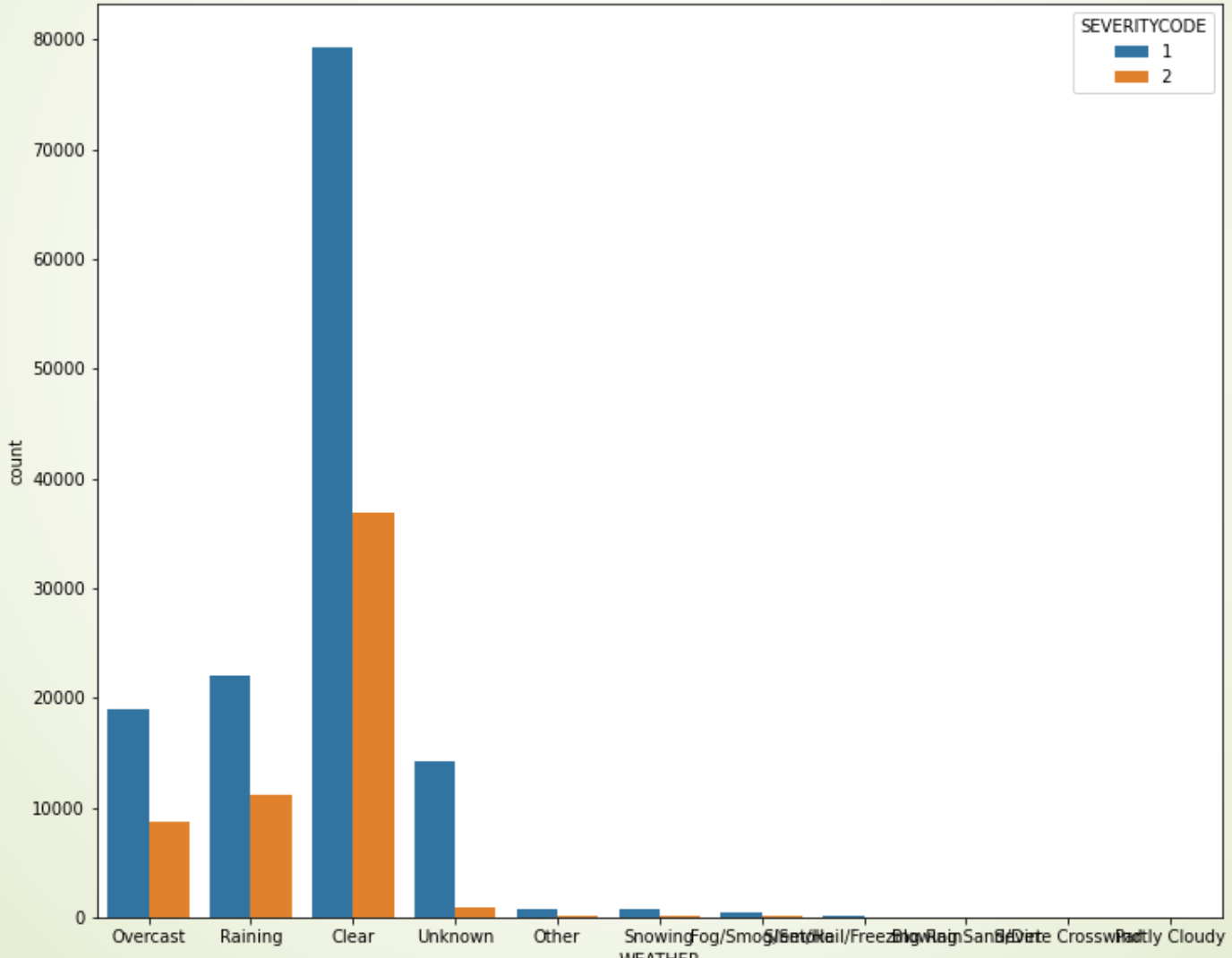
- The most frequent accidents occurred between 7 am to 7 pm, peak at 12 am!
- No relation for Day and Month



Correlation between features



Weather condition



Machine Learning Model

- Data was splitted to train (80%) and test (20%) set
- Normalized data
- Binary Classification: Decision Tree Classifier: criteria: “entropy”

- dt = DecisionTreeClassifier(criterion="entropy")

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='entropy', max_depth=None,
 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='deprecated’,
random_state=None, splitter='best')
```

Conclusion

- The decision tree model was applied to predict the severity of the car accident
- The model had the f1 score of 0.81
- The model needs to be updated itself based on new generated data