US Car Accidents Severity Analysis

Wanyue Xiao, Yingxue Gao and Yiyuan Cheng
School of Information Studies, Syracuse University, Syracuse, NY 13244 USA

Introduction

The dataset is about countrywide car accidents collected from 49 states in United States from February 2016 to December 2019[1]. There are about 3.0 million accident records with 49 features (such as location, weather, traffic, etc.) in this dataset.

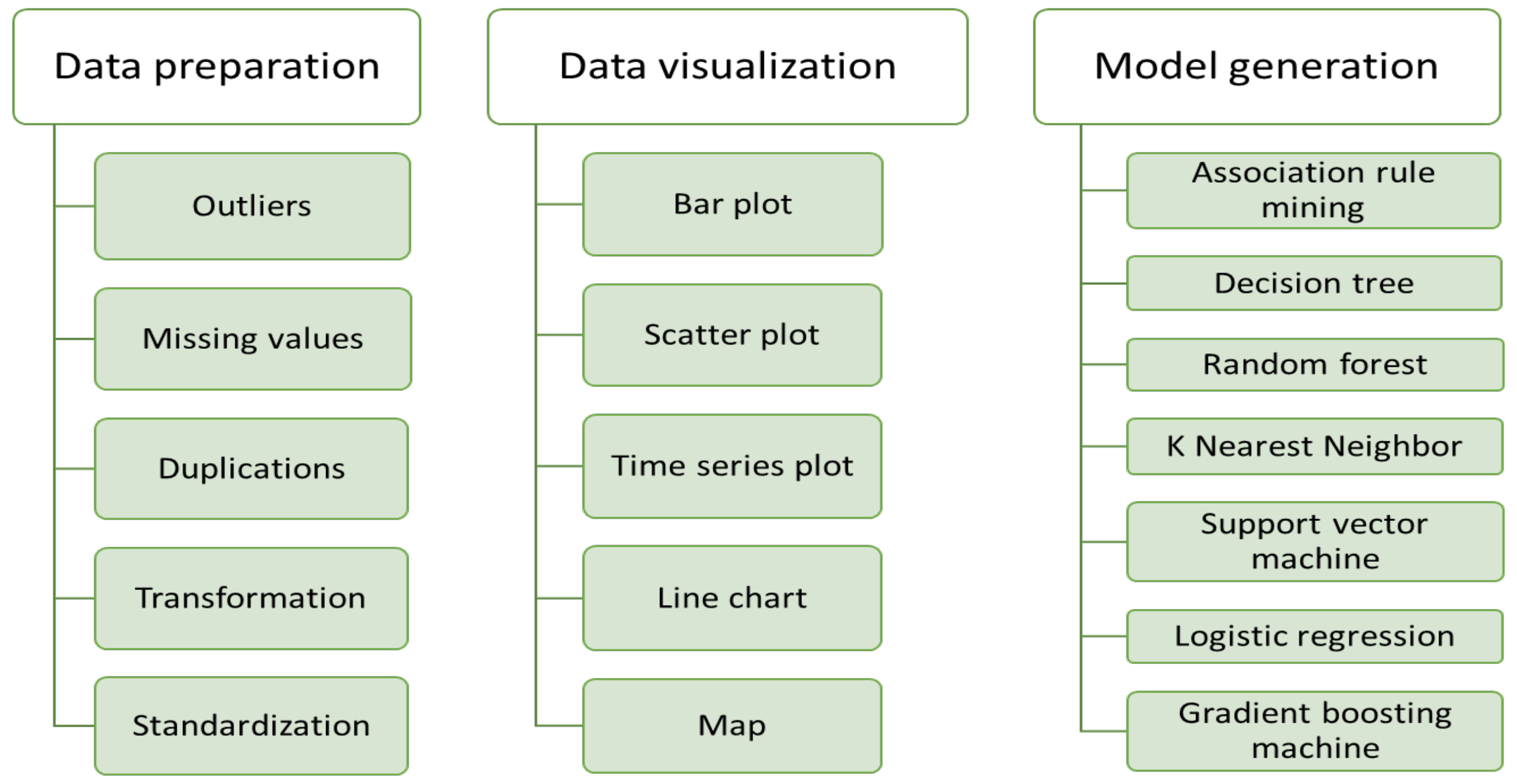
In order to know which features affect on the severity of accidents, we randomly sampled 20000 records from the original dataset for the analysis. Association rule mining was used to generated important rules related to accident severity. Decision tree, random forest, support vector machine and some other algorithms were used generate models for making prediction on the severity of accidents with other features.

Our results indicate that whether there is a traffic signal in the nearby location is a very important factor affecting the severity of the accident.

Objectives

- Identify important features and analyze how they affect the severity of accidents.
- Create models with different machine learning algorithms to make classification on the severity of accidents with these important features.
- Compare the prediction accuracies of different machine learning algorithms.

Approach



Data preparation
- Outliers
- Missing values
- Duplications
- Transformation
- Standardization

Data visualization
- Bar plot
- Scatter plot
- Time series plot
- Line chart
- Map

Model generation
- Association rule mining
- Decision tree
- Random forest
- K Nearest Neighbor
- Support vector machine
- Logistic regression
- Gradient boosting machine

Visualization

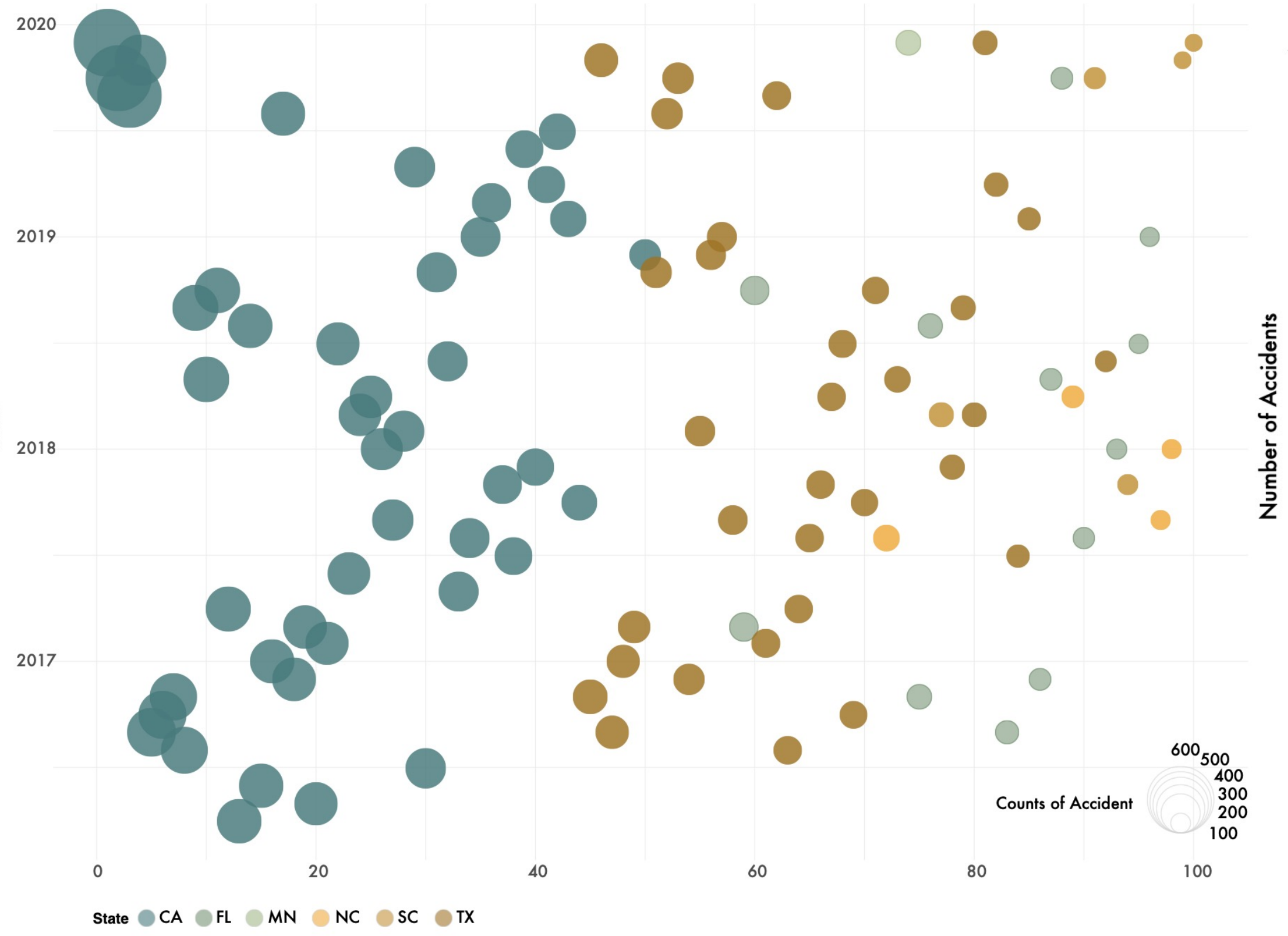


Figure 1: Top 100 Accidents

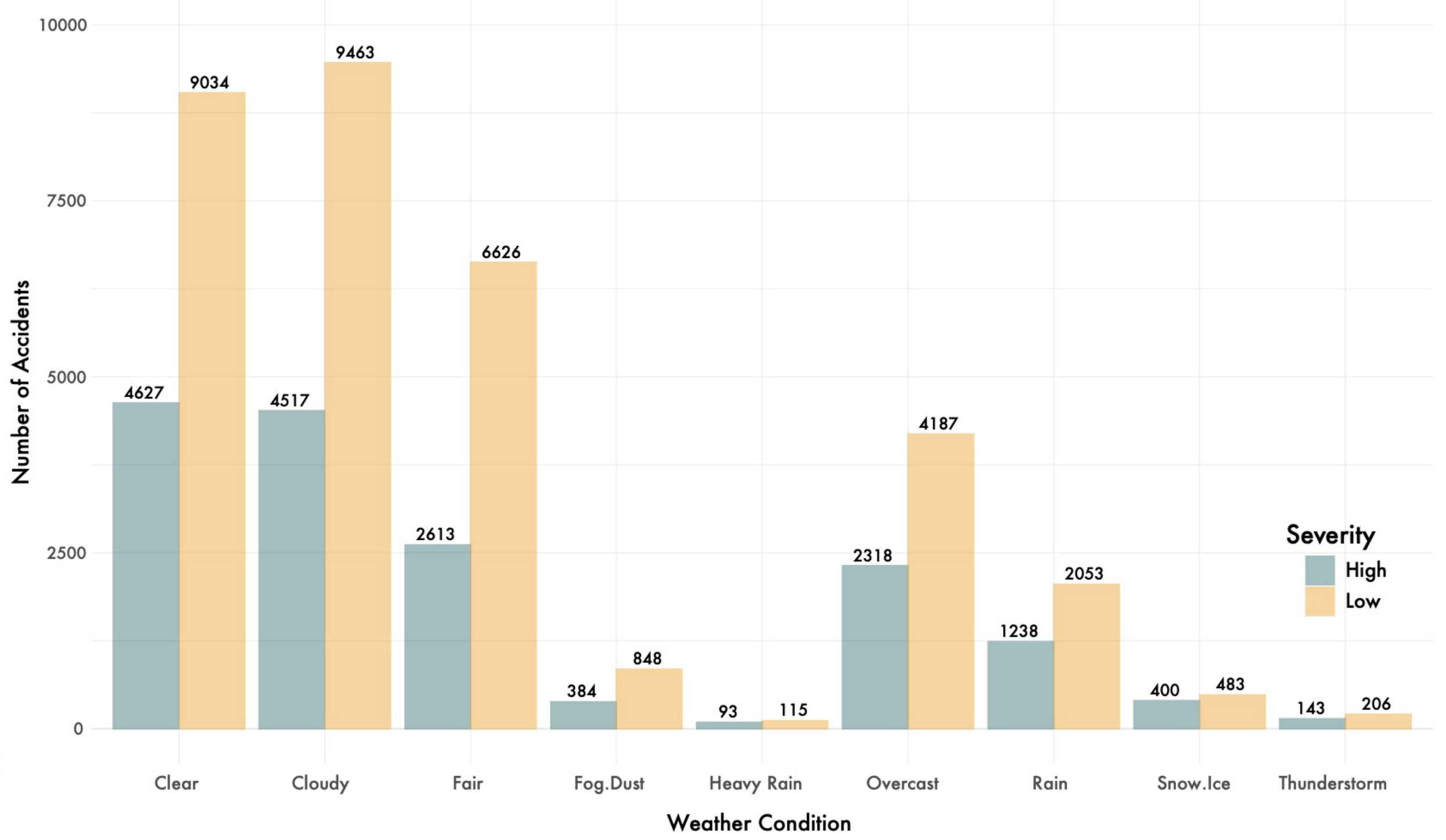


Figure 2: Accident count by Different Weather

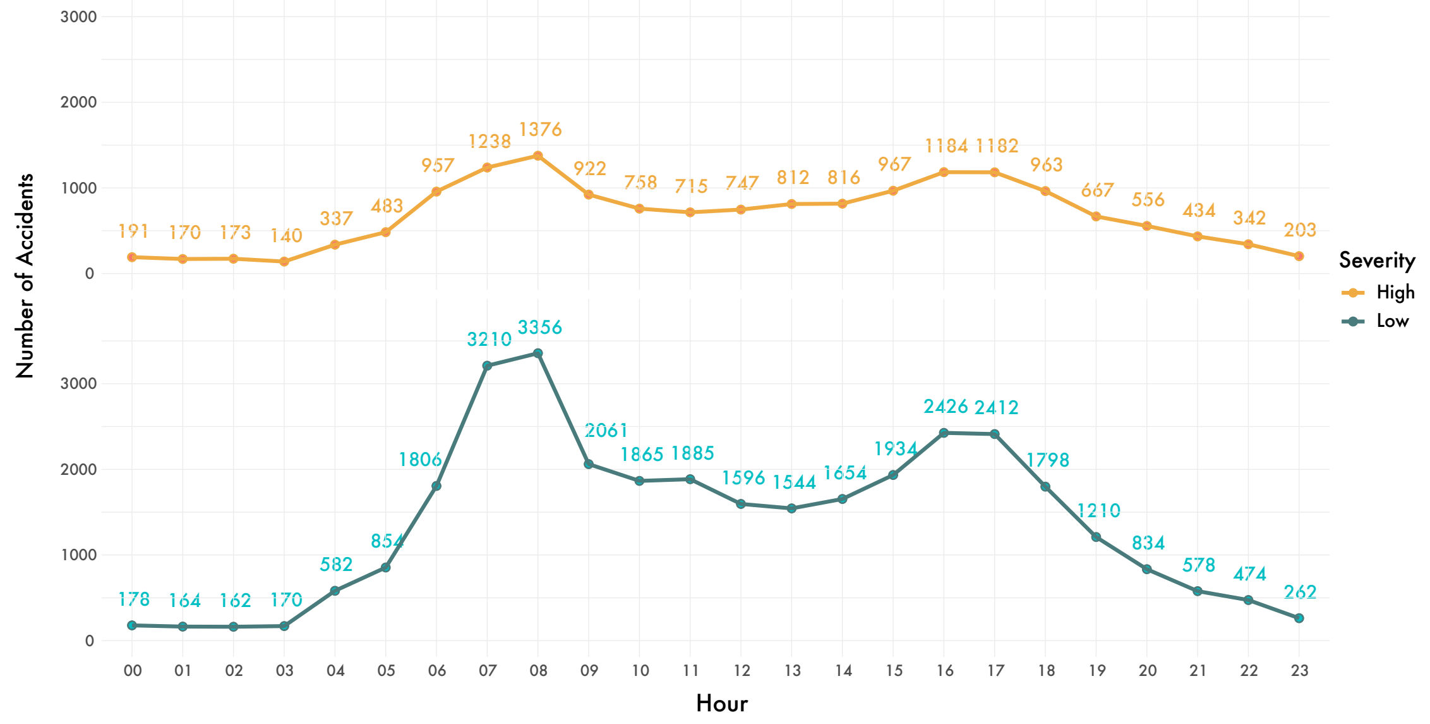


Figure 3: Accidents counts by Hour

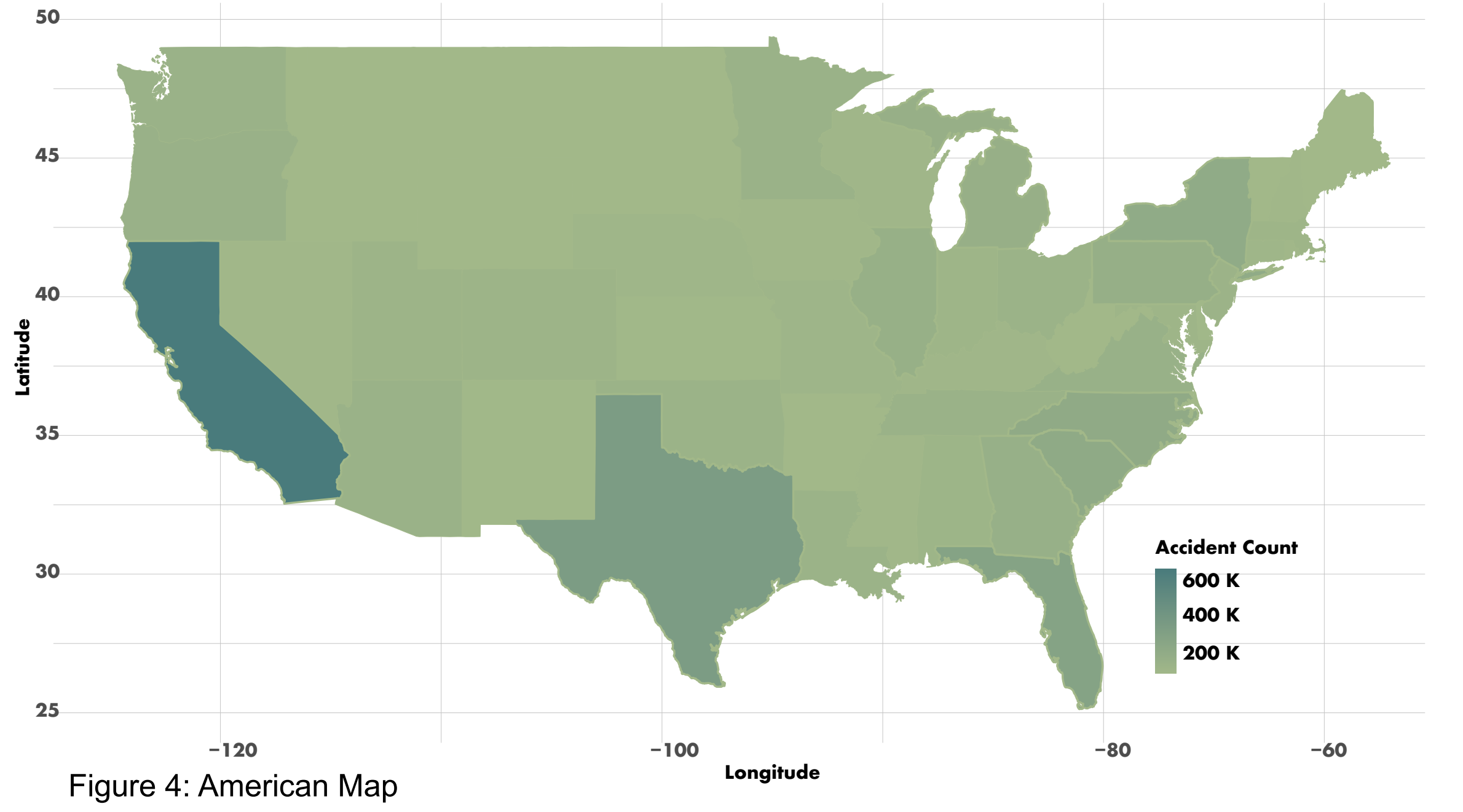


Figure 4: American Map

Methods

Association Rule Mining

Using the Apriori algorithm to derive frequent itemset and generate association rules.

$$Support(X) = \frac{Count(X)}{N}$$

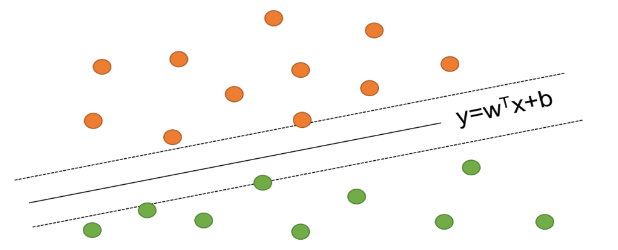
$$Confidence(X \rightarrow Y) = \frac{Support(X,Y)}{Support(X)}$$

$$Lift(X \rightarrow Y) = \frac{Support(X,Y)}{Support(X)Support(Y)}$$

Support Vector Machine

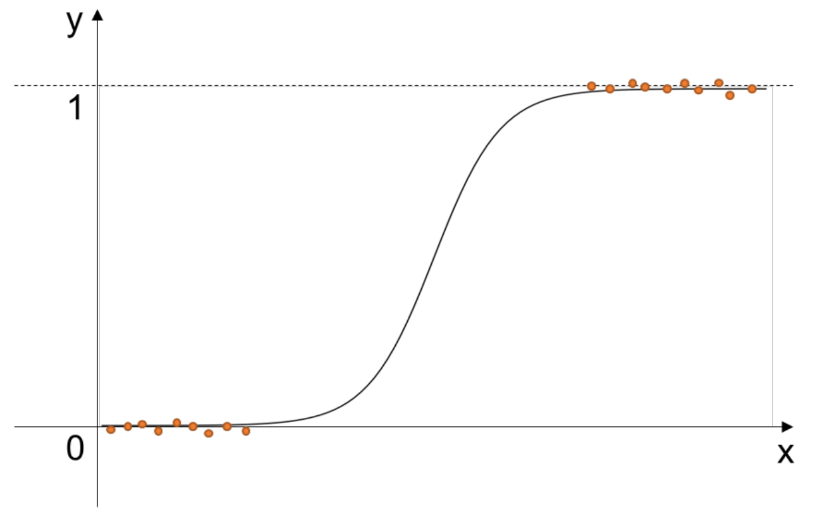
Classification based on a set of hyper-plane. In simple 2D, a linear hyper-plane can be defined as y = $w^T x+b$, with target distances (t_n) s.t:

$$w,b = \underset{w,b}{\arg\max}\left[\min_n\left(\frac{t_n\cdot\left(w^T x_n+b\right)}{\|w\|}\right)\right]$$



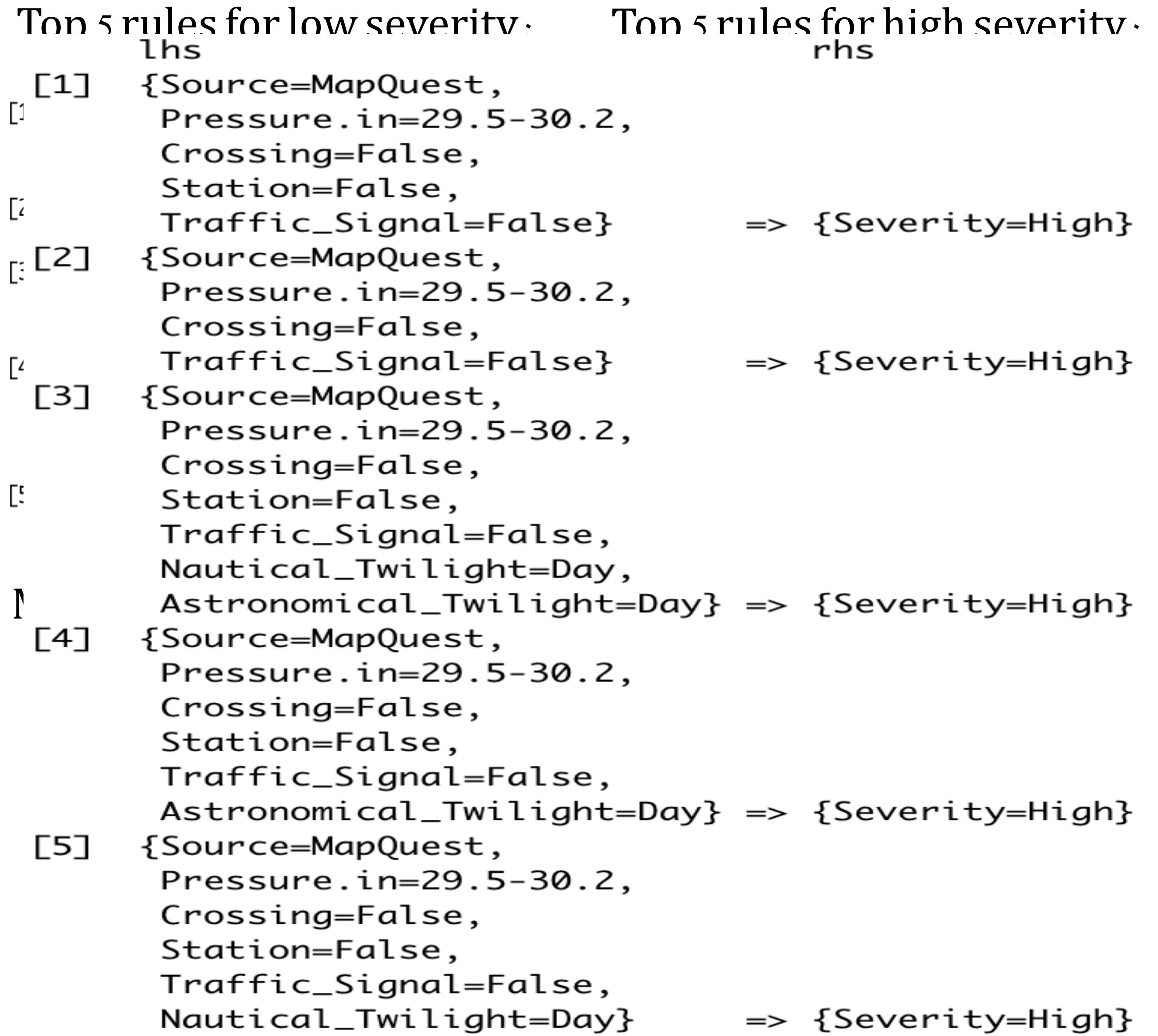
Logistic Regression

A statistical model that in its basic form uses a logistic function to model a binary dependent variable.

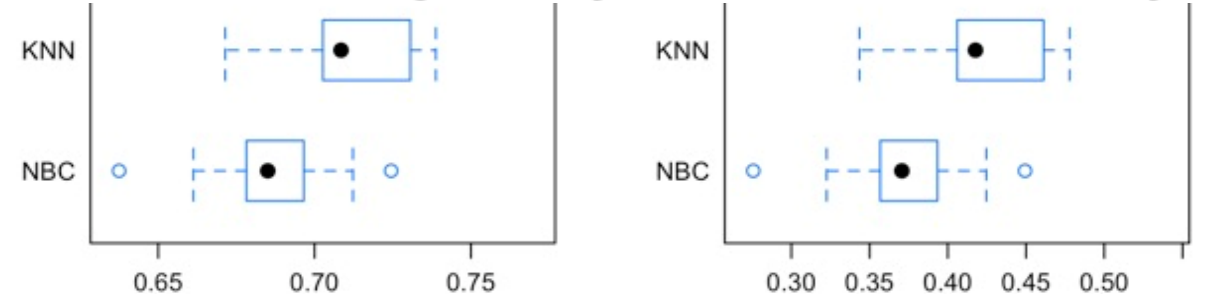


Results

Top 5 rules for low severity： Top 5 rules for high severity：



```
         lhs                                                rhs
 [1]   {Source=MapQuest,
        Pressure.in=29.5-30.2,
        Crossing=False,
        Station=False,
        Traffic_Signal=False}            => {Severity=High}
 [2]   {Source=MapQuest,
        Pressure.in=29.5-30.2,
        Crossing=False,
        Traffic_Signal=False}            => {Severity=High}
 [3]   {Source=MapQuest,
        Pressure.in=29.5-30.2,
        Crossing=False,
        Station=False,
        Traffic_Signal=False,
        Nautical_Twilight=Day,
        Astronomical_Twilight=Day} => {Severity=High}
 [4]   {Source=MapQuest,
        Pressure.in=29.5-30.2,
        Crossing=False,
        Station=False,
        Traffic_Signal=False,
        Astronomical_Twilight=Day} => {Severity=High}
 [5]   {Source=MapQuest,
        Pressure.in=29.5-30.2,
        Crossing=False,
        Station=False,
        Traffic_Signal=False,
        Nautical_Twilight=Day}           => {Severity=High}
```



Shiny app: https://yiyuan-cheng.shinyapps.io/IST707_Final/

Conclusion

- **Road condition**：No give way in the nearby location is an important factor to cause accident (either high or low severe) while It is more possible to be a high severe accident when there is no traffic signal.
- **Weather**：
- **Model**：XXX model has the best performance, which can be used to make prediction on the severity of accidents.

References

[1] U.S. Car Accidents dataset retrieved from the Kaggle：https://www.kaggle.com/sobhanmoosavi/us-accidents

Acknowledgement

We thank Prof. Ying Lin for his guidance and help.

xwanyue@syr.edu
ygao65@syr.edu
ycheng26@syr.edu