Predicting severity of car incident

Takeshi Oda

2020/5/4

1. Introduction

In this project, I built a classification model to predict serious car accident in the U.S. To do this, I used "US Accidents A Countrywide Traffic Accident Dataset (2016 - 2019)" in Kaggle dataset. (https://www.kaggle.com/sobhanmoosavi/us-accidents) This dataset contains 2974335 observations about car accidents from February 2016 to December 2019 across entire U.S. This data sets give us severity of car accidents along with more than 40 variables such as location, weather, conditions.

Using this data set, I attempted to present answer to the question:
How much would be the severity of the car accident when one car accident happens? This dataset provides response variable Severity i.e., a number between 1 and 4, where 1 indicates the least impact on traffic (short delay) and 4 indicates a significant impact on traffic. (long delay) **I defined binary classification variable Severity_bin i.e., "Low" for Seveirty 1 and 2, "High" for Severity 3 and 4 and built a predictive model to estimate probability to have "High" severity against specific input of a street such as weather, time and road structure.**

I assume this model is used by public relation sector in a regional government (county government) which is in charge of notifying car accident in their web site or social media. Using the model, public relation agents will be able to provide as accurate estimated time for recovery as possible helping citizens take alternative way to the destination.

To narrow the scope of the model, I extracted accident data of Illinois state in 2019 from original dataset. To enrich predictive variable, I downloaded population data for each county in Illinois from https://www.illinois-demographics.com/counties_by_population and combined it with car accident data set.

2. Analysis Process

Through preliminary analysis, I found several correlation of response variables with predictive variables. **Serious accident is most likely to happen in the afternoon while light accident is much more likely to happen in the morning. (Figure 1) Serious accident is more likely to happen on weekend. (Figure 2) Proportion of no traffic signal is much higher in High severity than Low severity. (Figure 3) Accidents are frequent in right side of the street. But serious accident is much more frequent in right side than light accident. (Figure 4)**

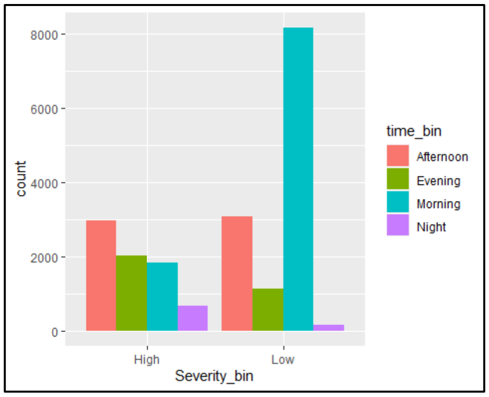

Figure 1

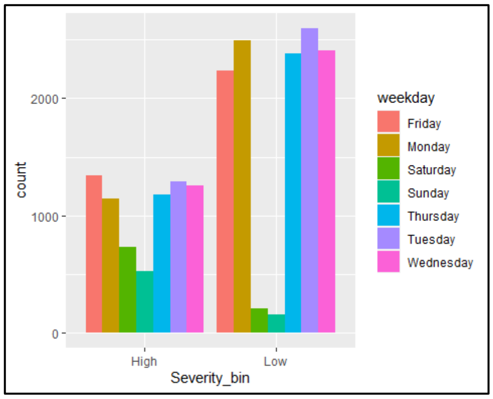

Figure 2

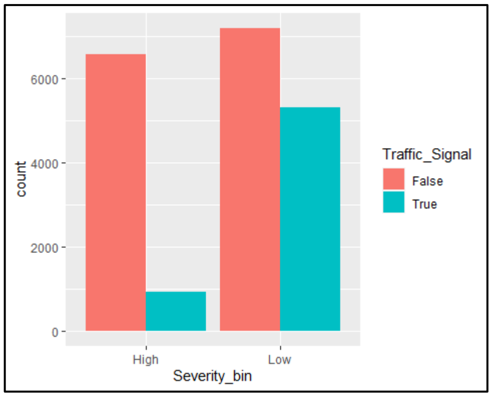

Figure 3

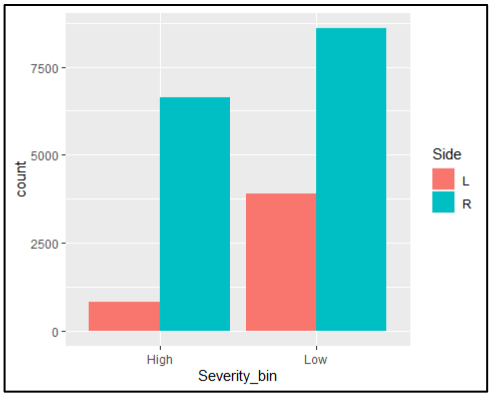

Figure 4

I took an approach to **balance between predictive accuracy and model interpretation.** While the model should provide high capability to predict serious car accident, it should enable local government to gain insight into hidden factors of serious car accidents. Therefore, I chose three classification methods, **Logistics Regression, Bagging and Random Forest**.

Through model fitting and validation on training data set, **Random Forest** achieved least **cross validation error rate 0.22** restricting **9 variables** per tree. This means the model explain around 78 percent of variation of severity. Following five variables are identified as strong predictor.

- Time
- Presence of Traffic Signal near by the location
- Day of Week
- Population
- Relative side of the street

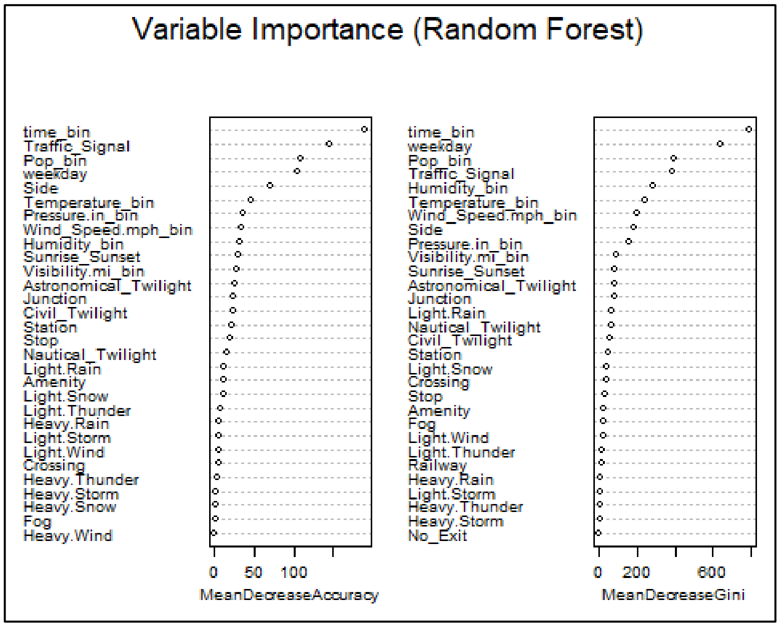

Figure 5 Variance Importance Plot

3. Model Evaluation and Recommendation

As a final step of predictive modeling, prediction error rate is measured. Random forest model is fitted on all the training data set and prediction is performed on validation set. **Final model achieved prediction error rate 0.21 on validation data set**. This result suggests that local public relation team should utilize this model in the following reasons. First, they will be able to serve better to drivers with accurate information. Since they will be able to present accurate information of severity at 79 percent confidence, drivers will be able to make decision whether they drive another route or wait for the accident to be recovered. Second, this model helps streamline accident notification process and increase efficiency. By collecting pre-defined survey items, they will reduce back and forth investigation to estimate severity of an accident.