Car Accident Severity Analysis in Seattle

Sep 13th 2020

ByoungGue

Table of Contents

1. Background / Business problem

1-1. Background

When you driving to another city, you may come across a terrible traffic jam on the highway, it may be an car accident due to bad weather such as rainy and windy. Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

1-2. Business Problem

The business problem is to predict severity of an accident using relative variables such as weather and road condition, based on prediction score we can set up the warning system.

If we could warn the severity level before driving the area to the people, people can plan to reduce or avoid to drive any high severity area.

2. Data summary

2-1. Data source

Data Set Basics	
Title	Collisions—All Years
Abstract	All collisions provided by SPD and recorded by Traffic Records.
Description	This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present.
Keyword(s)	SDOT, Seattle, Transportation, Accidents, Bicycle, Car, Collisions, Pedestrian, Traffic, Vehicle
Data name/format	Data-Collisions.csv
Contact Information	
Contact Organization	SDOT Traffic Management Division, Traffic Records Group
Contact Person	SDOT GIS Analyst
Contact Email	DOT_IT_GIS@seattle.gov

2. Data summary

2-2. Key attribute information

Attribute	Data type,length	Description
SEVERITYCODE	int64	A code that corresponds to the severity of the collision:
		• 3—fatality
		• 2b—serious injury
		• 2—injury
		• 1—prop damage
		• 0—unknown
ADDRTYPE	object	Collision address type:
		• Alley
		• Block
		• Intersection
LOCATION	object	Description of the general location of the collision
SEVERITYDESC	object	A detailed description of the severity of the collision
PERSONCOUNT	int64	The total number of people involved in the collision
VEHCOUNT	Double	The number of vehicles involved in the collision; This is entered by the state.
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.

2. Data summary

2-3. Data read & review

- Data contains 38 columns and 194673 rows.
  After cleansing missing data 187525 rows are left.
- Extract 6 columns related to severity
  'SEVERITYCODE’
  'ADDRTYPE’
  'ROADCOND’
  'LIGHTCOND’
  'WEATHER’
  'PERSONCOUNT’
  'VEHCOUNT’
- Compare to each attributes with severity, ADDRTYPE is highly related to severity but the severity level was decided based on number of person and car impacted on collisions

3. Modeling

3-1. Model selection

The expected outcome is to predict severity level which is categorical value. The other independent variables are also categorical variables which contains text values.

Mulitiple categorial variables can selected as condition for final result and not much numerical correlation calculated.

Based on this I selected **Decision Tree Model**.

3-2. Data pre-processing

Convert text values to numeric values

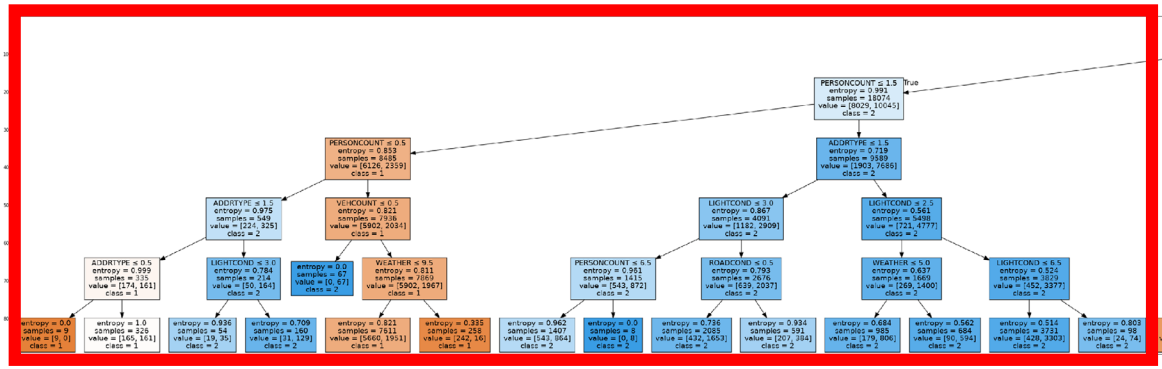
3. Modeling

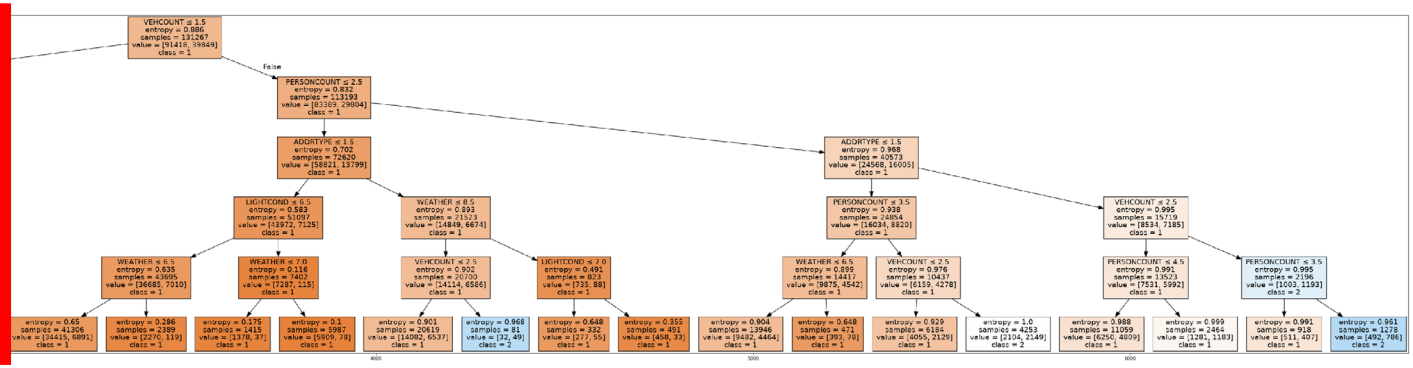
3-3. Setting Modeling

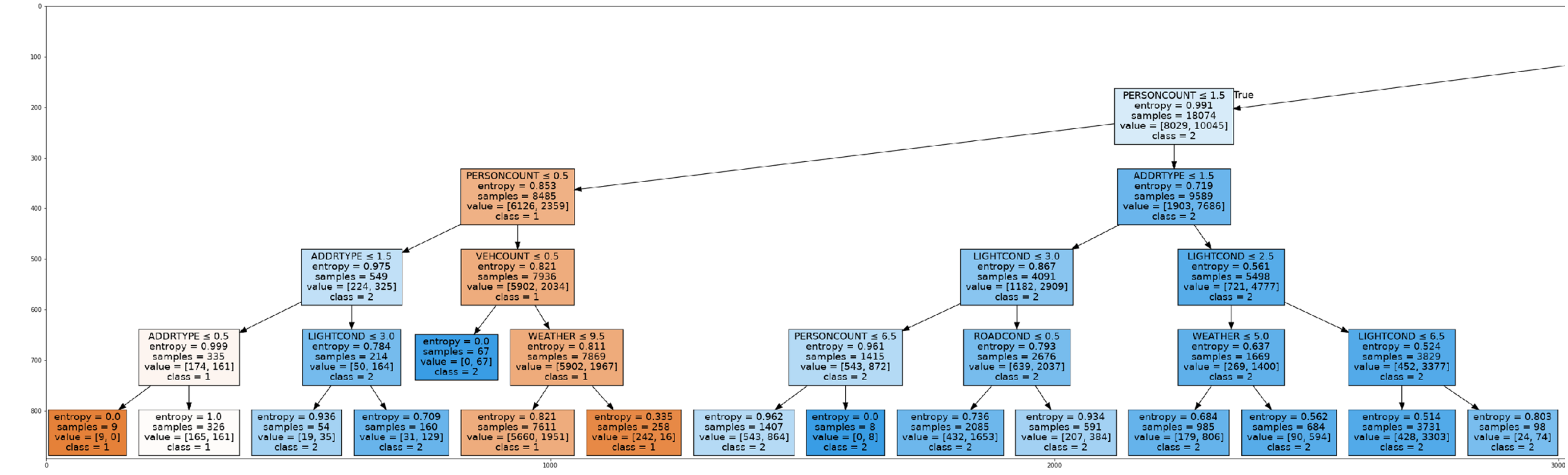
- Test model setting with test_size=0.3, random_state=3
- Modeling : Decision tree
- Evaluation
  - DecisionTrees's Accuracy: 0.74
  - DT Jaccard index: 0.74
  - DT F1-score: 0.70
- Tested scenarios
  - scenario 1 : Mixture of 'ADDRTYPE','ROADCOND','LIGHTCOND','WEATHER' => No significant outcome, all potential possibility is severitycode 1, not well predicted
  - scenario 2 : Mixture of single independent attribute => Beside 'ADDRTYIPE', other attributes have no significant impact on severity.
  - scenario 3 : Mixture of other attiribute 'JUNCTIONTYPE', 'Location' added to scenario 1, but not significant impact on severity.
  - scenario 4 : Mixture of numeric value 'PERSONCOUNT','VEHCOUNT' in scenario 1 => showing severity impact. although it is the the result of severity, need to see the consequence of other attribute.
  - Finally Selected scenario 4

3. Modeling

3-4. Visualization

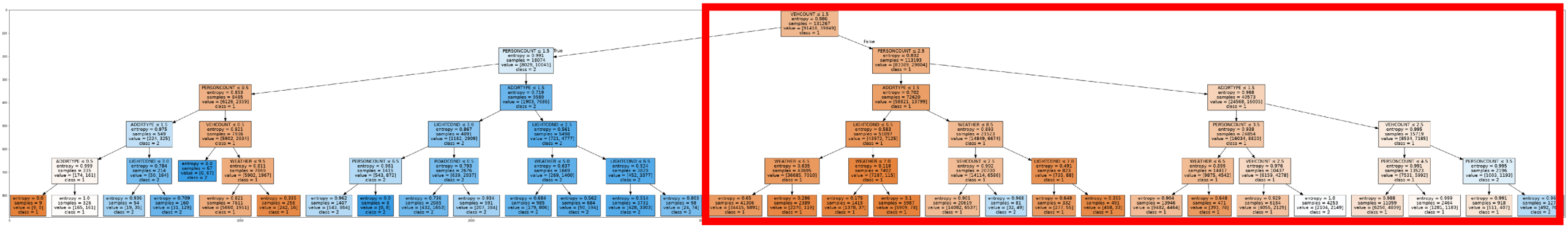


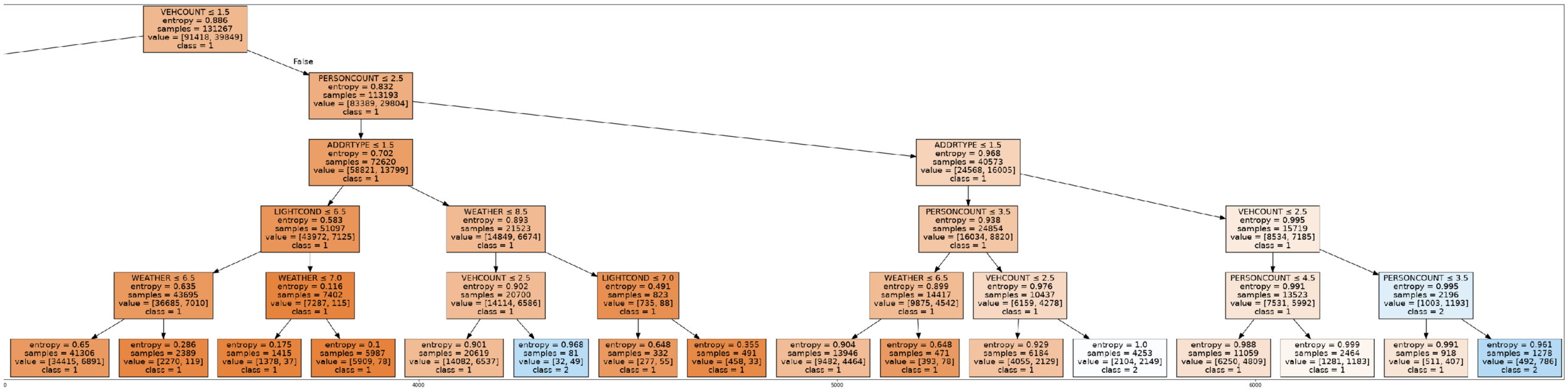




3. Modeling

3-4. Visualization





4. Result

- Severity is decided mainly from the number of vehicles involved in the collision and The total number of people involved in the collision, but not exactly having linear relationship.
- Impacted Number of peope, Number of Vehicle are depending on Addtype, Light and Weather.
  For example, the situation in Alley or Block, in Dark time, under Smog or Cloudy weather are highly predicted severity 2.
- Based on those two numeric factors, we define the condition of 4 attribute, ADDRTYPE, LIGHTCOND, WEATHER, ROADCOND
- ADDRTYPE is highly related to severity level, Weather and Roadcondition have less impact

5. Discussion

- Each columns doesn’t have significant relationship to severity. therefore we need to do more validation.
- Number of People and number of Vehicle are key factors of severity, but we cannot forecast those factors in this exercise and also those factors are not showing linear relationship.
- Further consideration on more data cleansings such as others or unknown categories.
- What is the optimal depth for decision tree model generation?
- Can we consider other model, random forest or hiarachical clustering

6. Conclusion

- Severity is decided mainly from the number of vehicles involved in the collision and The total number of people involved in the collision, but not exactly having linear relationship.
- Through the decision tree model, we defined some conditions have been increased severity such as intersection and block situation, night-time and cloudy weather.
- But we need to find highly correlated factors on severity to define better decision.