IBM
APPLIED DATA SCIENCE CAPSTONE - DATA COLLISIONS

1)INTRODUCTION

I examined data before suggest to business problem. Fill
NaN values with correct values that the way I think it is
right. This infos been in own cells. Data has accident by
Traffic Records. I evaluate here how accident is
happened, how many people dead or get injured, how
many vehicle included accident, collisions type of
vehicles, as so many parameters in that data. There are
some factors about driver that get drunk or alcohol.
Weather, road and light condition about accident.
Location of accident. All that parameters will be examined in detail. If an accident occur anywhere ,
data has information about there. I said "there" is that mean all info about cars, peoples,weather
condition namely all factors may affect the accident.

2)BUSINESS PROBLEM

where did the accident occur ? why did the accident happen ? which factors affect the accident
weather, light or road condition or all of them ? was driver drunk? was driver too speedly? I think
ours target is that questions answer. I want to see factors that affect the accident. Collisions type and
collisions address is interest each other. If interest , how we explain this ? which time interval is
have high accident rate and why is that? then maybe find some solution about accident problem.
Maybe it is road problem or psychological problem or we don't discover yet that any new problem
that we will discover with this data.

3)DATA

Data include accident that have some factors about driver, pedestrian, road cond, weather cond,
road type, collisions type, accident time, location, severity, injuries, dead. How occur accident and
its explanation as briefly. I did some adjustment about unknown situations. I will do some feature
engineering and add some new attributes in that data. Maybe detect the location in that occur
accident frequently. And below I explain data in detail and do some changing about missing values
and thought to be wrong values.

3.1) Data Understanding

The dataset has 194673 rows with a total of 38 different columns, each detailing
some variable inherent to the accident. A summary information about variables been in notebook. I
adjusted "ADDRTYPE", "ROADCOND", "LIGHTCOND", "WEATHER", "UNDERINFL",
"COLLISIONTYPE", "JUNCTIONTYPE" variables. I did some explanation about that adjustment
in notebook. This adjustment generally about fixed some deficient infos in independent variables.
And I prepared this independent variables for machine learning algorithm as correctly.

3.2)Data Preparation

There are some wrongly writing values. For example; "N" that means no and as same
meaning "0" is different be written and so I adjust this values for "0" other one is "1". And then I

use that values with dummy variables for machine learning. Then I want to categorize some values in independent variables as "good condition", "bad condition" and "normal condition". Sometimes that names be different because of independent variables's values. I did this adjustemnt because I want prepare independent variable for machine learning algorithms. And then I realised that independr variables in dataset is generally categorical. So I will use machine learning algorithm according to that.

4) METHODOLOGY

The model chosen was the decision tree for some reasons, such as easy implementation, good description and it is highly interpretable. I explained data understanding and preparation above. And now I am explaining the algorithm which I used for detect which severity is that accident. I used decision tree with dummy variables that I adjust. ex. "LIGHTCOND", "ROADCOND" etc. thats are dummy variable "STATUS", "ADDRTYPE", "COLLISIONTYPE", "JUNCTIONTYPE", "SDOT_COLCODE", "WEATHER", "ROADCOND", "LIGHTCOND", "UNDERINFL". And concat that with numeric variables and train the algorithm.

5) RESULTS - CONCLUSIONS

I wonder factors that effect the accident. And done some coding for explain that my wonder. I did some crosstab metods for look values for collisins type and addrtype. I examine correlation between independents variables. I did some visualize for good sense for dataset and how can I explain that. And I did decision tree. I saw the conclusion. This set prediction for "1" is 84% probability and "2" is 44% probability. Maybe later may do some trimming for improve the probability.