Car Accident Severity Prediction in Seattle

- V. Sai Sudheer

1. Introduction

1.1 Background

Car Collision occurs when a vehicle collides with another vehicle, pedestrian, animal, road debris, or other stationary obstruction, such as a tree, pole or building. Traffic collisions often result in injury, disability, death, and property damage as well as financial costs to both society and the individuals involved. They also impact economy with increased commuting times, increased delivery times of products and costs, and pollution, due to the massive number of cars waiting for the road to be cleared, or heavily slowed down.

A number of factors contribute to the risk of collisions, including vehicle design, speed of operation, road design, road environment, driving skills, impairment due to alcohol or drugs, and behavior, notably distracted driving, speeding and street racing.

1.2 Problem

Building a model to predict the severity of car accident using traffic accident data in Seattle city of United States. This can help commuters, health workers, Administration of Seattle, professional drivers or logistic planners to reduce the personal and/or business impact of car accidents**.**

1.3 Target Audience

This model can help commuters, professional drivers, or logistic planners to reduce the personal and/or business impact of car accidents.

Seattle Government: Accident-prone areas can have Interventions such as speed breakers, Traffic signs, establish new Police check-posts for checking drunken driving, etc.. can help reduce accidents

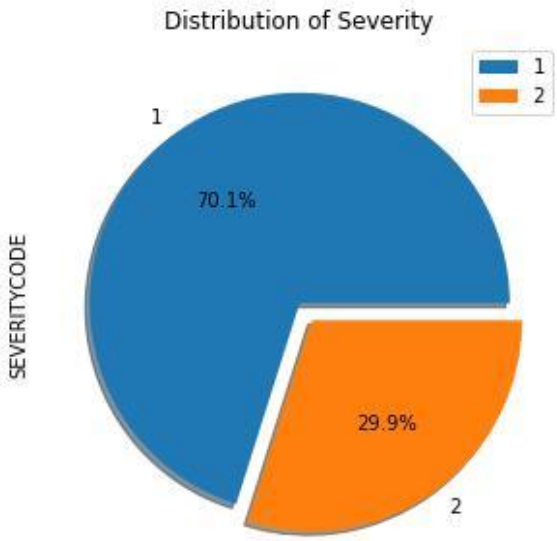
Car Owners: Owners staying in the Areas where parked cars are prone to get hit by other vehicles should concentrate on parking spots and can pay more insurance in order to decrease the loss.

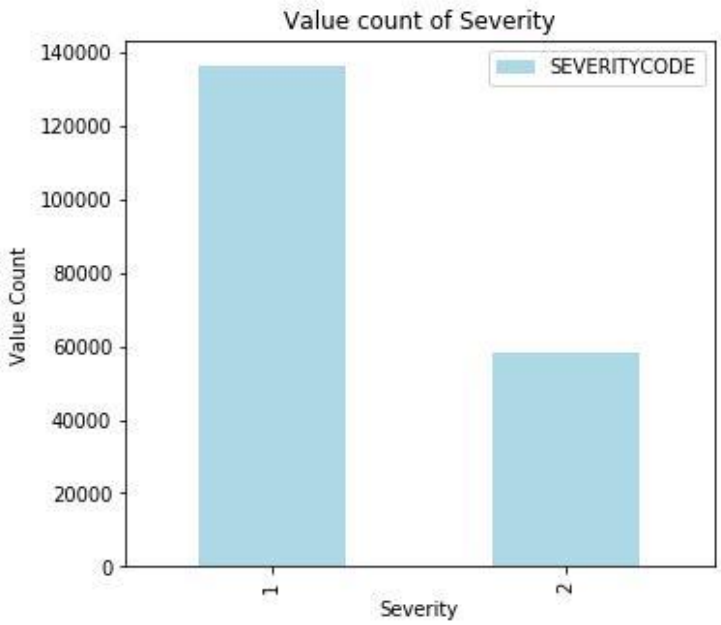
Health care and Emergency services in Seattle: On predicting the severity of accidents, they can take necessary actions and can potentially save lives.

2. Data requirements

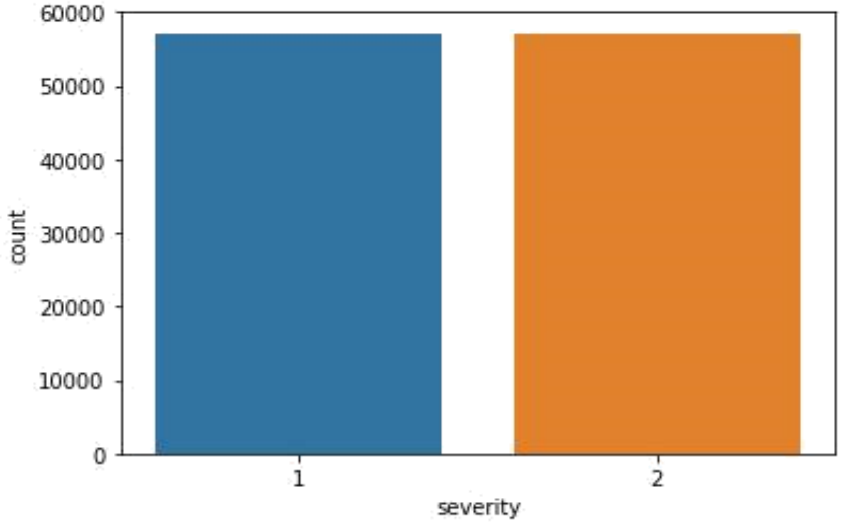
2.1 Data

The traffic accidents data is obtained for the Seattle city in this link. The metadata description of the columns of the data is available in this link. The data period is from January 1, 2004 to May20, 2020. The column 'SEVERITYCODE' is target label that the model should predict. It is notated as 1and 2which represent property damage and injury respectively. The distribution of the severity level of the data is presented as follow. About 70.1% of accidents were with level 1 and 29.9% were with level2.





Note that the distribution between classes is not even, so under-sampling the data will help to avoid biased classification model.



After doing the under-sampling, both the classes have equal count and can potentially decrease bias.

The column INCDATE is converted into date object in the dataset.

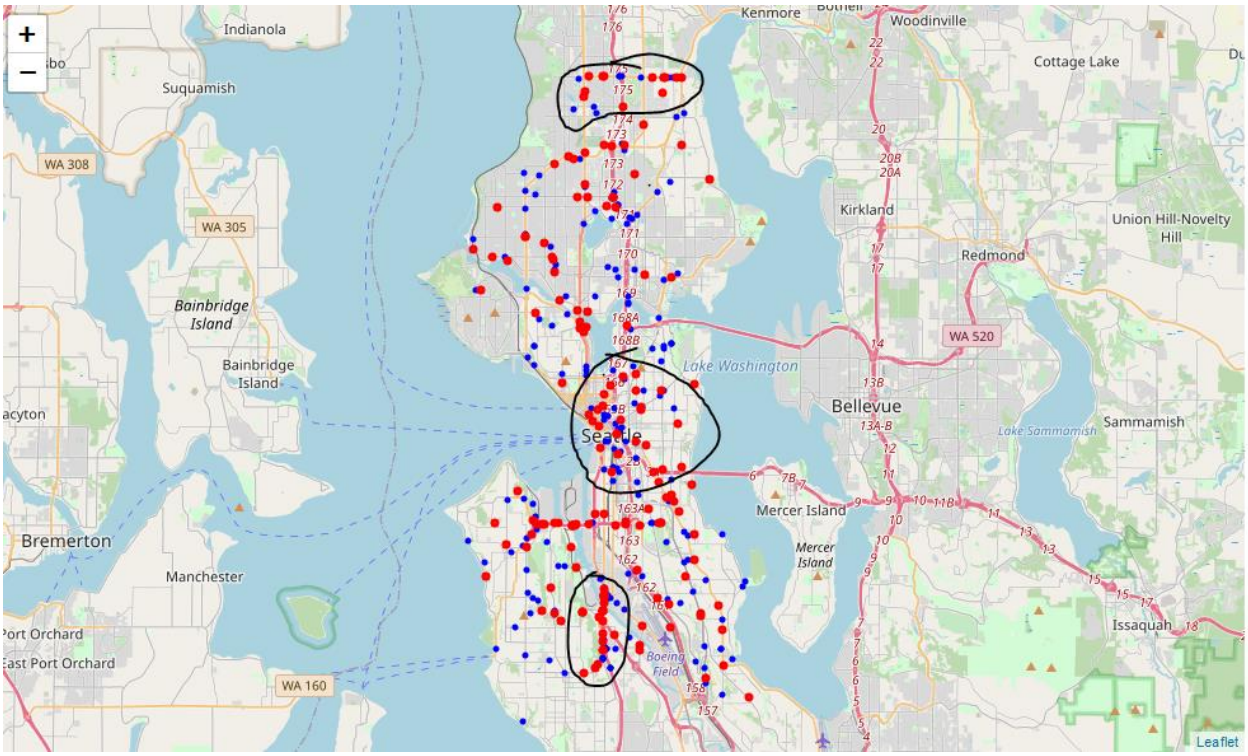
2.2 Feature Selection

Out of 38 features, selected only 17 necessary features that are related to the problem statement. The selected features are listed below.

1. ADDRTYPE: Collision address type (Alley, Block, Intersection)
2. PERSONCOUNT: The total number of people involved in the collision
3. PEDCOUNT: The number of pedestrians involved in the accident
4. PEDCYLCOUNT: The number of bicycles involved.
5. VEHCOUNT: Number of vehicles involved.
6. UNDERINFL: Whether or not driver involved was under influence of drugs
7. WEATHER: Weather condition at the time of accident
8. ROADCOND: Road condition
9. LIGHTCOND: Light condition
10. PEDROWNOTGRNT: Whether pedestrian right of way was not granted
11. SPEEDING: Whether or not speeding was a factor of collision
12. HITPARKEDCAR: Whether collision involved hitting a parked car
13. Date: Date of accident
14. Year: Year of collision
15. Weekday: day of the week accident happened
16. Hour: Time of accident
17. Month: Month of accident

3. Data Analysis

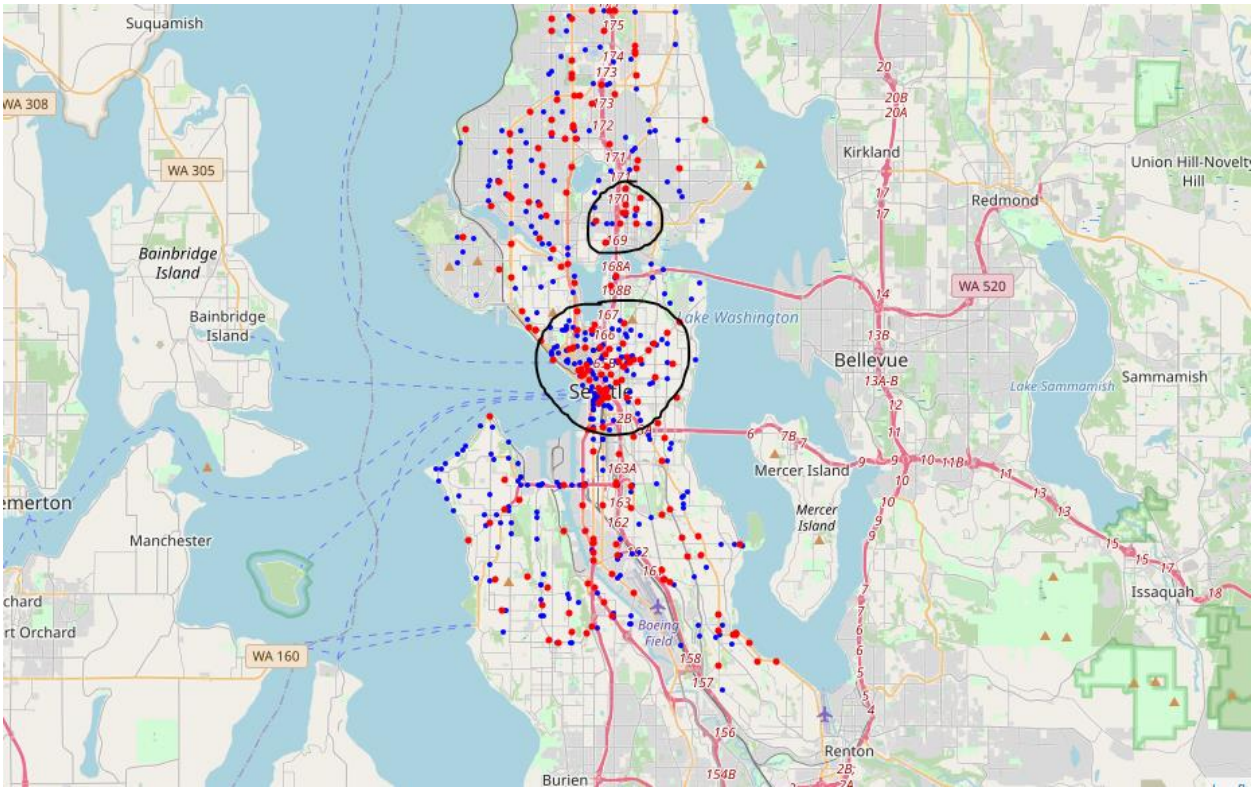
Speeding vs accident severity map



Certain roads have a lot of accidents as shown in the black circles, so govt can put speed breakers on those roads in order to decrease accidents due to over speed.

Under Drug influence and accident severity

Police can introduce check-ups on vehicles that are entering the black circles.



3.1 Feature Selection

After doing feature engineering, I selected the 19 most relevant features out of 38 features.

SL No.	Feature	Description	Reason for Selecting
1	ADDRTYPE	Collision at Alley, Block, Intersection	Gives the likelihood of collision at these places
2	PERSONCOUNT	Number of people involved in the collision	Gives an indication of severity
3	PEDCOUNT	Number of pedestrians involved in the accident	Gives an indication of severity
4	PEDCYLCOUNT	Number of cyclists involved in the accident	Gives an indication of severity
5	VEHCOUNT	Number of vehicles involved in the accident	Gives an indication of severity
6	INATTENTIONIND	Whether the person was not paying attention	Not paying attention can result in accident
7	UNDERINFL	Whether the person was driving under influence	DUI can cause accidents
8	WEATHER	Weather conditions	Bad weather can cause accidents
9	ROADCOND	Road conditions	Wet roads can cause skidding
10	LIGHTCOND	Light conditions	Light conditions affect visibility
11	PEDROWNOTGRNT	Pedestrian right of way was granted or not	
12	SPEEDING	Whether speeding or not	Speeding causes accidents
13	COLLISIONTYPE	Collision Type	Type of collision gives severity of accident
14	HITPARKEDCAR	Whether or not the collision involved hitting a parked car.	Hitting a parked car causes property damage
15	Year	Year of accident	Did one year have a lot of accidents
16	Month	Month of Accident	Does month affect number of accidents
17	Day	Day of accident	Day of month
18	Hour	Time of accident	Are accidents caused majorly at night
19	Weekday	What day of the week accident happened	Are accidents caused more on certain days of the week

We are going to build the model using these 19 features. In order to do that, we should first encode all the object type features to int or float type and this process is called feature encoding.

3.2 Feature Encoding

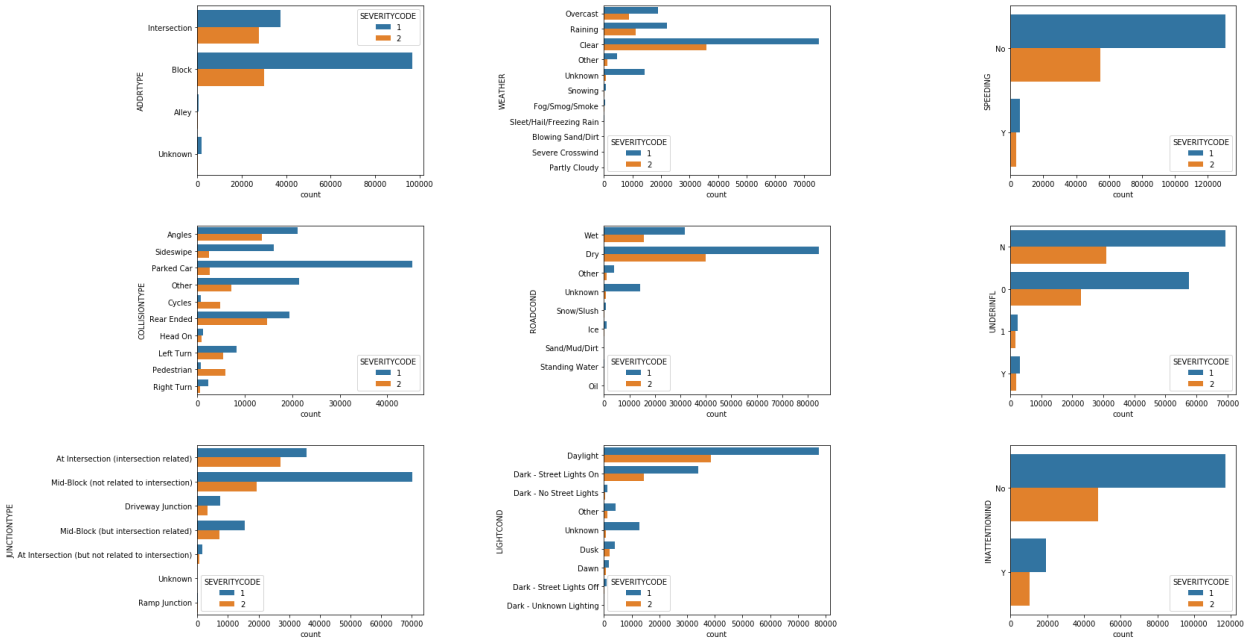
Data types of selected features are given below:

- Nominal categorical variables such as ADDRTYPE, WEATHER, ROAD COND, LIGHTCOND, COLLISIONTYPE are replaced with unique integers for each category value.

- Binary categorical variables such as INATTENTIONIND, UNDERINFL, SPEEDING, HITPARKEDCAR, PEDROWNOTGRNT can be replaced with 1 and 0

```
ADDRTYPE          object
PERSONCOUNT        int64
PEDCOUNT           int64
PEDCYLCOUNT        int64
VEHCOUNT           int64
INATTENTIONIND    object
UNDERINFL         object
WEATHER           object
ROADCOND          object
LIGHTCOND         object
PEDROWNOTGRNT     object
SPEEDING          object
COLLISIONTYPE     object
HITPARKEDCAR      object
Year               int64
Month              int64
Day                int64
Weekday            int64
Hour               int64
SEVERITYCODE       int64
```

3.3 Count plots



From the count plot, we can say the most common value of every feature. Let's see each feature type that has the most accidents below.

The most common address type is "Block" that has most of the accidents. "Clear" is the most common weather type. Most accidents don't have a speeding condition. "Parked car" is the most common collision type. "Dry" is the most common type of road condition. Most accidents are not due to alcohol consumption. Most accidents are occurring at "Mid-Block(not related to intersection)", and in "Daylight".

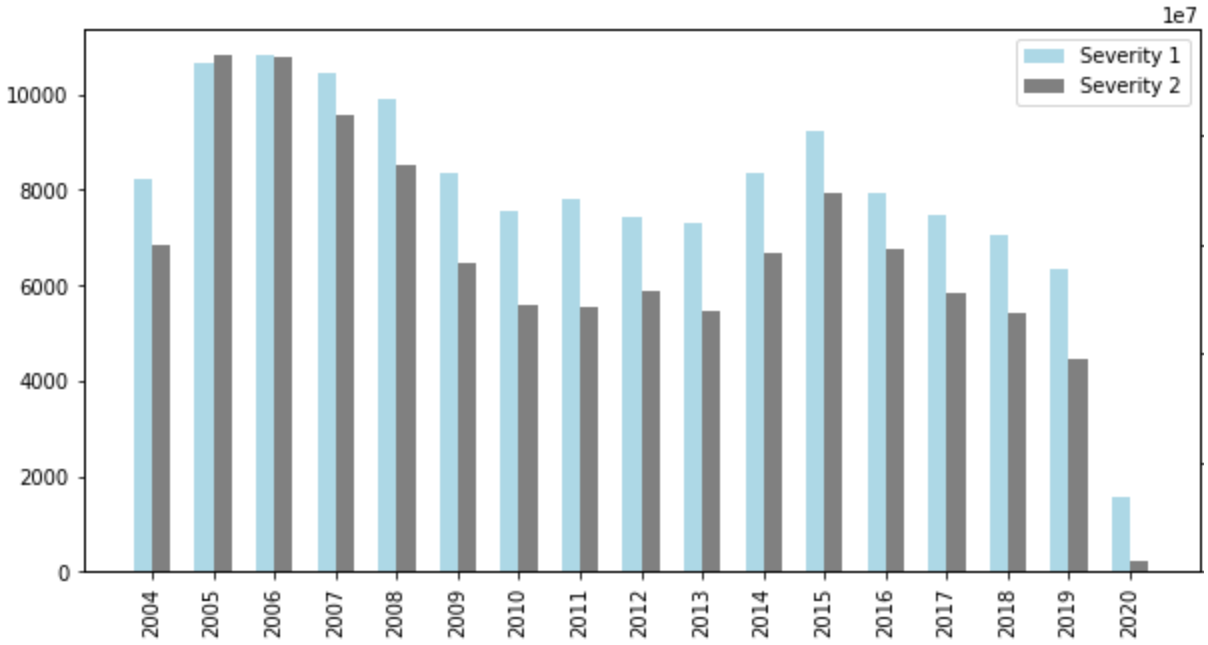


Fig above showing the count of accident severity each year. The year 2005,2006 has the most number of accidents and from the year 2015 accidents have been decreasing.

4.Model Development and Evaluation

After cleaning and resampling the data, we split the data into training(70%) and testing(30%) samples. Trained with different classifiers and results are tabulated below. The optimization step has also been done.

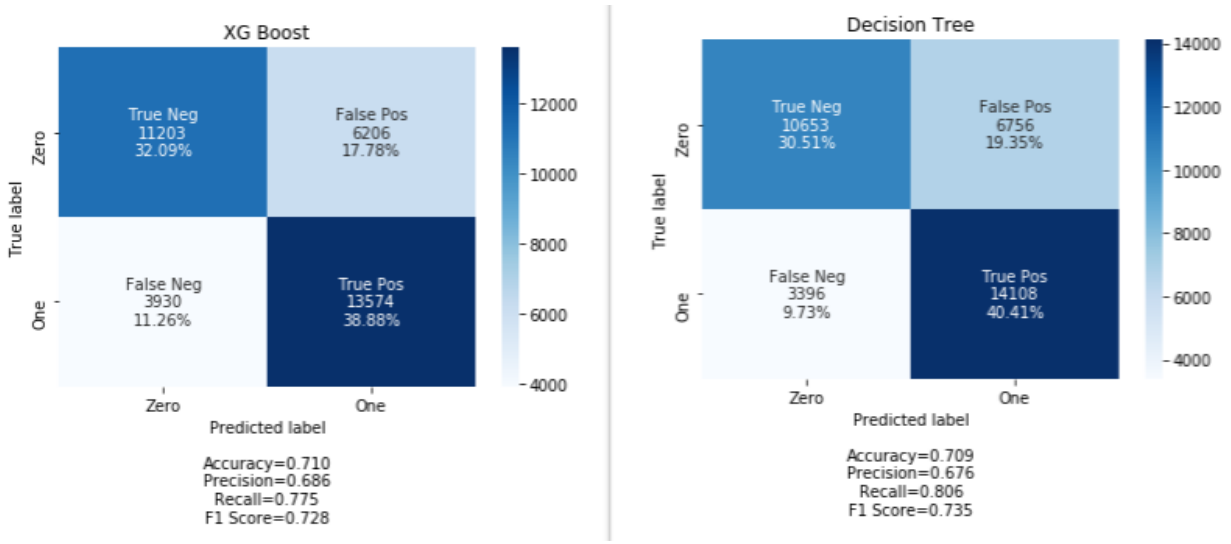


Figure above shows the confusion matrix plot for XG boost and decision tree classifiers.

Classifier	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.662	0.693	0. 583	0.633
K Means	0.649	0.660	0.619	0.639
Random Forest	0.695	0.684	0.728	0.706
XG Boost	0.710	0.688	0.767	0.728
Decision Tree	0.709	0.676	0.806	0.735

From the above table, we can say that the Decision tree and XG boost classifier performs well on the data and has high f1 scores compared to other classifiers

5.Conclusion

The data-set has been used to classify the severity of the accidents based on certain select features. The exploratory data analysis shows the density of accidents based on geography-based on Speeding, Driving Under Influence, In-attention, and Hitting Parked Cars. From a machine learning standpoint. The most important

features were: Collision Type, Person Count, Vehicle Count, and Address Type.The Decision Tree algorithm performed the best.AI in self-driving cars can use such models to assess the risk of accidents and change routes or ask the driver to be vigilant during auto-pilot.