Business Report: Predicting Survival in Car Crashes Using Machine Learning

Extended Project Report

Submitted to

greatlearning
Learning for Life

By

Subhadeep Seal

In Partial Fulfillment of PDP-DSBA

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

CONTENTS

PROBLEM STATEMENT

1.1. Context

Car crashes are a leading cause of injury and death worldwide, and improving vehicle safety is a critical concern for car manufacturers. With advancements in technology and engineering, manufacturers are continuously seeking ways to design safer vehicles to reduce fatalities and severe injuries in the event of a crash. Despite these efforts, understanding the precise factors that contribute to survival in car crashes remains a complex challenge.

The problem arises from the nature of car accidents, where various elements such as impact speed, the use of safety features, the type of collision, and the demographics of the occupants all play significant roles. Each crash is unique, and even minor variations can significantly affect the outcome for the occupants. This complexity necessitates a detailed analysis to identify which factors are most influential in determining survival outcomes.

Solving this problem is essential for several reasons:

1. Safety Regulations
2. Design Improvements
3. Public Health
4. Consumer Confidence

1.2. Problem Definition

Car accidents remain a major public health concern, with a 15% year-over-year increase in urban incidents. This report aims to analyze five years of historical crash data to identify key survival determinants and develop predictive models. The ultimate goal is to inform safety regulations and design enhancements for automotive manufacturers.

1.3. Objective

Analyze historical car crash data to uncover patterns related to survival. Develop machine learning models to predict survival outcomes. Identify and interpret critical factors influencing survival. Provide actionable recommendations for improving vehicle safety and road regulations.

1.4. Data Description

The data contains the different attributes of car crashes, with the outcome variable being whether the occupant was deceased during the crash or not. The detailed data dictionary is given below.

Data Dictionary

- caseid: character, created by pasting together the population sampling unit, the case number, and the vehicle number. Within each year, use this to uniquely identify the vehicle.
- speed_range: factor with levels (estimated impact speeds) 1-9 km/h, 10-24 km/h, 25-39 km/h, 40-54 km/h, 55+ km/h
- wei
- ght: Observation weights, albeit of uncertain accuracy, are designed to account for varying sampling probabilities. (The inverse probability weighting estimator can be used to demonstrate causality when the researcher cannot conduct a controlled experiment but has observed data to model)
- seatbelt: a factor with levels none or belted
- frontal_impact: a numeric vector; 0 = non-frontal, 1=frontal impact
- sex: a factor with levels f: Female or m: Male
- age_of_occ: age of occupant in years
- year_of_acc: year of accident
- model_year: Year of model of vehicle; a numeric vector
- airbag: Did one or more (driver or passenger) airbag(s) deploy? This factor has levels deploy, nodeploy, and unavail.
- occ_role: a factor with levels driver or pass: passenger
- deceased: the target variable with levels no (survived) or yes (not survived / deceased)

Dataset consists of crash records with attributes including case ID, speed range, occupant weight, seat-belt usage, type of impact, occupant demographics, year, vehicle model, airbag deployment, occupant role, and survival outcome.

2. DATA OVERVIEW

- We will view the first 5 & last 5 rows of the dataset.

	caseid	speed_range	weight	seatbelt	frontal_impact	sex	age_of_occ	year_of_acc	model_year	airbag	occ_role	deceased
0	02:13:02	55+ km/h	27.07800	none	1	m	32	1997	1987	unavail	driver	yes
1	02:17:01	25-39 km/h	89.62700	belted	0	f	54	1997	1994	nodeploy	driver	yes
2	0.138206019	55+ km/h	27.07800	belted	1	m	67	1997	1992	unavail	driver	yes
3	0.138206019	55+ km/h	27.07800	belted	1	f	64	1997	1992	unavail	pass	yes
4	04:58:01	55+ km/h	13.37400	none	1	m	23	1997	1986	unavail	driver	yes

	caseid	speed_range	weight	seatbelt	frontal_impact	sex	age_of_occ	year_of_acc	model_year	airbag	occ_role	deceased
11212	82:107:1	25-39 km/h	3179.68800	belted	1	m	17	2002	1985	unavail	driver	no
11213	82:108:2	10-24 km/h	71.22800	belted	1	m	54	2002	2002	nodeploy	driver	no
11214	82:110:1	10-24 km/h	10.47400	belted	1	f	27	2002	1990	deploy	driver	no
11215	82:110:2	25-39 km/h	10.47400	belted	1	f	18	2002	1999	deploy	driver	no
11216	82:110:2	25-39 km/h	10.47400	belted	1	m	17	2002	1999	deploy	pass	no

Table 1: First 5 & last 5 rows of the dataset

2.1. Shape of the Dataset

- The dataset contains 11217 rows & 12 columns.

2.2. Check the type of data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11217 entries, 0 to 11216
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   caseid          11217 non-null  object
 1   speed_range     11217 non-null  object
 2   weight          11217 non-null  float64
 3   seatbelt        11217 non-null  object
 4   frontal_impact  11217 non-null  int64
 5   sex             11217 non-null  object
 6   age_of_occ      11217 non-null  int64
 7   year_of_acc     11217 non-null  int64
 8   model_year      11217 non-null  int64
 9   airbag          11217 non-null  object
 10  occ_role        11217 non-null  object
 11  deceased        11217 non-null  object
dtypes: float64(1), int64(4), object(7)
memory usage: 1.0+ MB
```

Table 2: Data types

There are 7 object data types, 4 integer data types, and 1 float data type in the dataset. All these features could be good predictors for an outcome of an accident.

2.3. Check for missing values

	0
caseid	0
speed_range	0
weight	0
seatbelt	0
frontal_impact	0
sex	0
age_of_occ	0
year_of_acc	0
model_year	0
airbag	0
occ_role	0
deceased	0

dtype: int64

Table 3: Missing Values

- There are no missing values in the dataset.

2.4. New Variable creation

- New variable created veh_usage_duration, that Indicates the time period (in years) the vehicle has been in use.

2.5. Statistical summary of the dataset

	weight	frontal_impact	age_of_occ	year_of_acc	model_year
count	11217.00000	11217.00000	11217.00000	11217.00000	11217.00000
mean	431.40531	0.64402	37.42765	2001.10324	1994.17794
std	1406.20294	0.47883	18.19243	1.05681	5.65870
min	0.00000	0.00000	16.00000	1997.00000	1953.00000
25%	28.29200	0.00000	22.00000	2001.00000	1991.00000
50%	82.19500	1.00000	33.00000	2001.00000	1995.00000
75%	324.05600	1.00000	48.00000	2002.00000	1999.00000
max	31694.04000	1.00000	97.00000	2002.00000	2003.00000

Table 4: Statistical summary

- In the above table we can see the counts, mean, standard deviation, minimum value and maximum value of numerical features.

3. EXPLORATORY DATA ANALYSIS (EDA)

3.1. Univariate Analysis

- Revealed distributions of deceased, weight, age_of_occ, speed_range, airbag, seatbelt, frontal_impact, sex, model_year, occ_role & veh_usage_duratiuon. Bar plots & Histogram-Box plots for each distribution are as follows:

Observations on deceased:

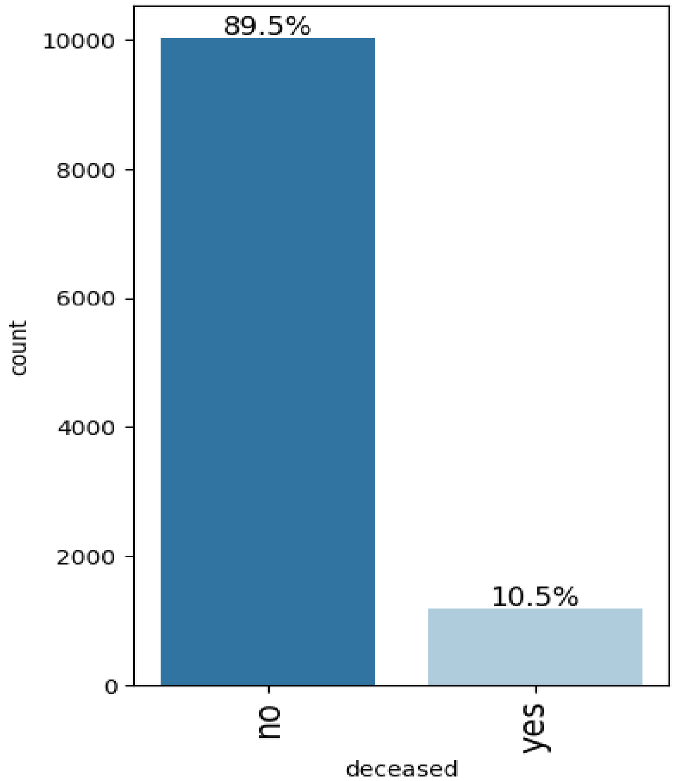

Fig-1

- It shows that rate of survival in accident is 89.5% and deceased is 10.5%.

Observations on weight:

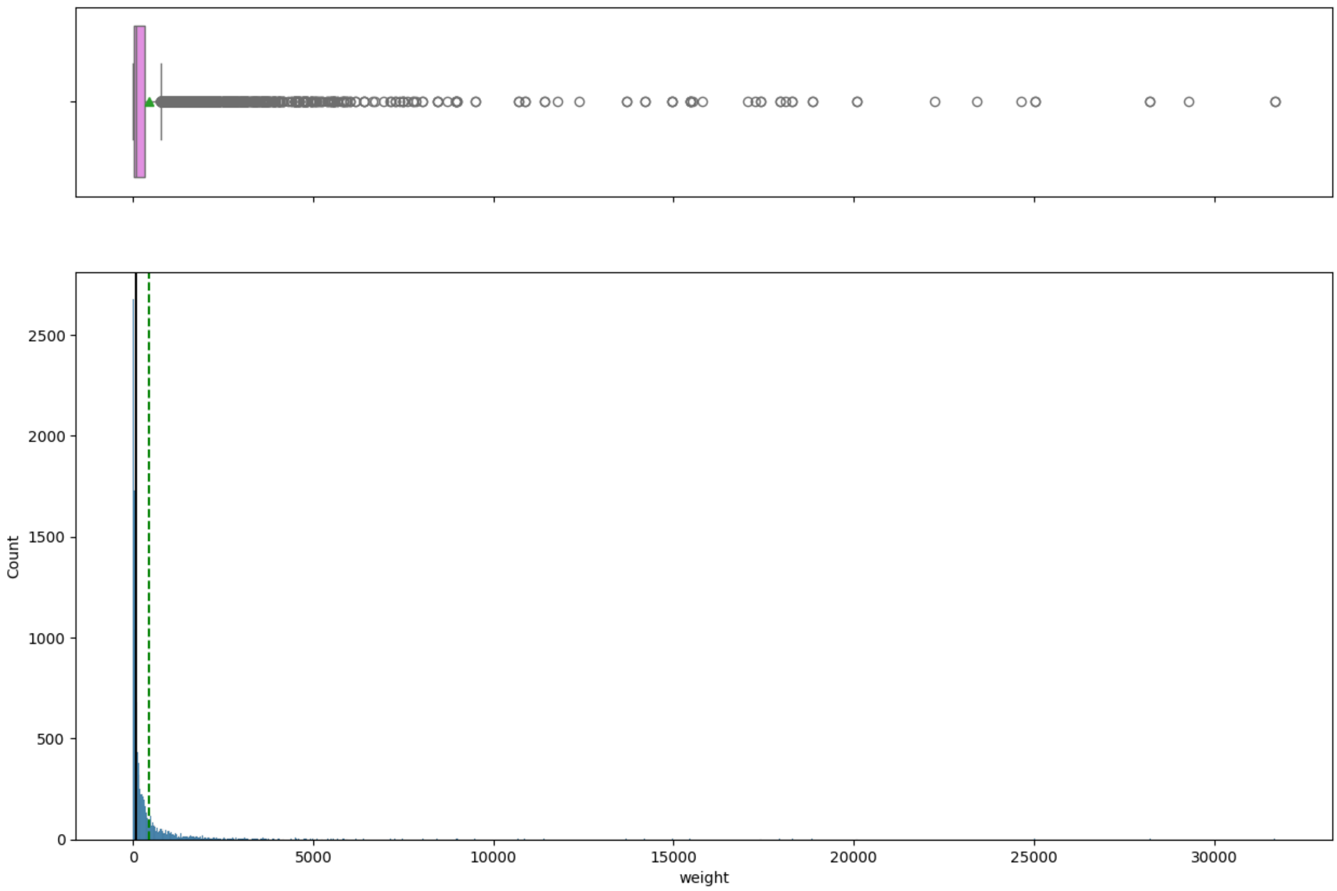

Fig-2

- There are huge outliers present in the distribution of weight.

Observations on age_of_occ

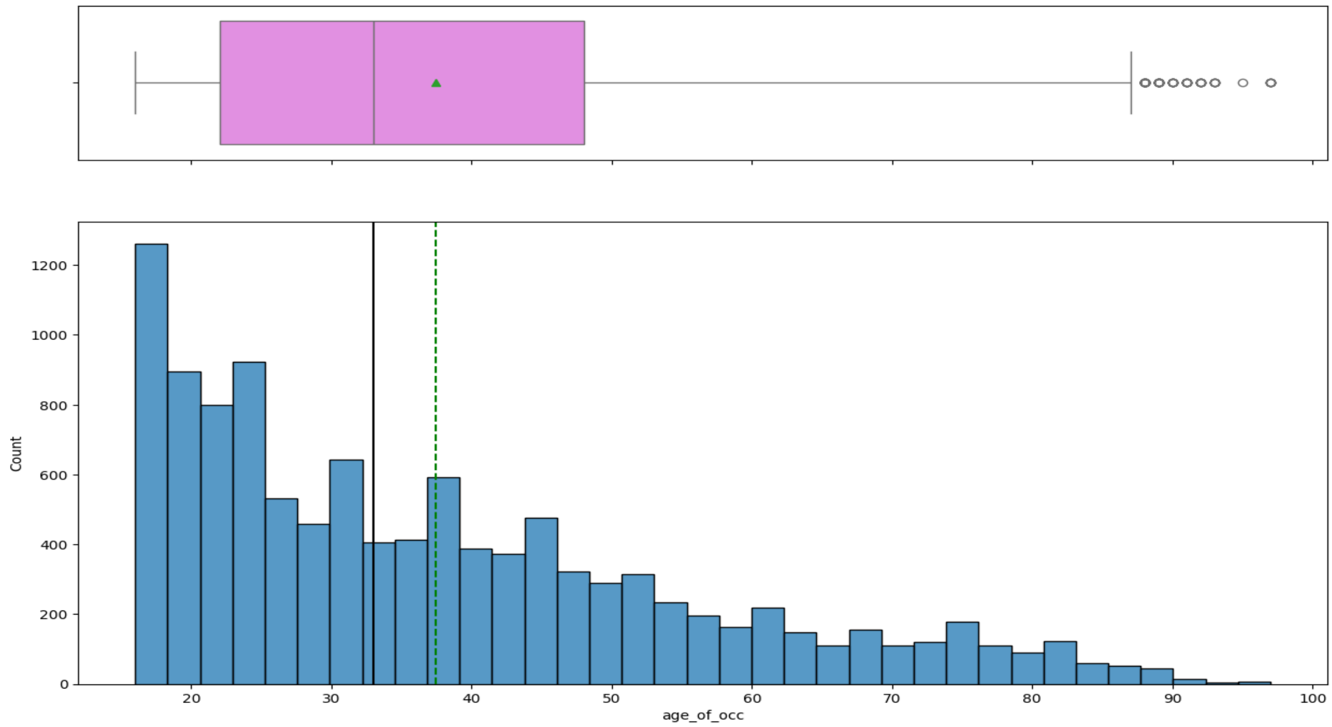

Fig-3

- Records for accidents between ages 10 to 15 are the highest.
- There are few outliers present in the distribution of age of occupants.
- The data is slightly skewed towards right.

Observations on speed_range

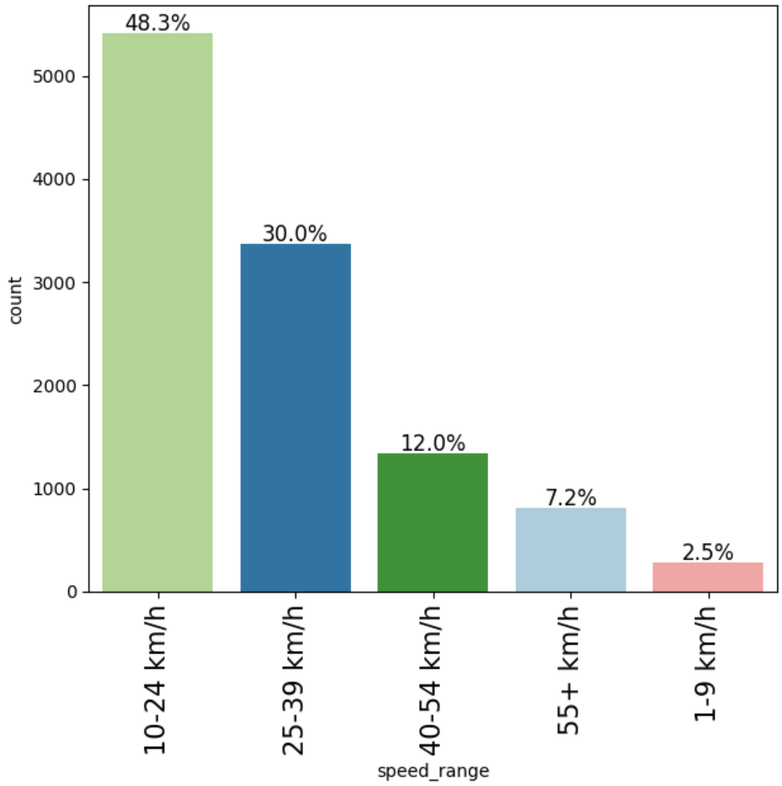

Fig-4

Observations on airbag

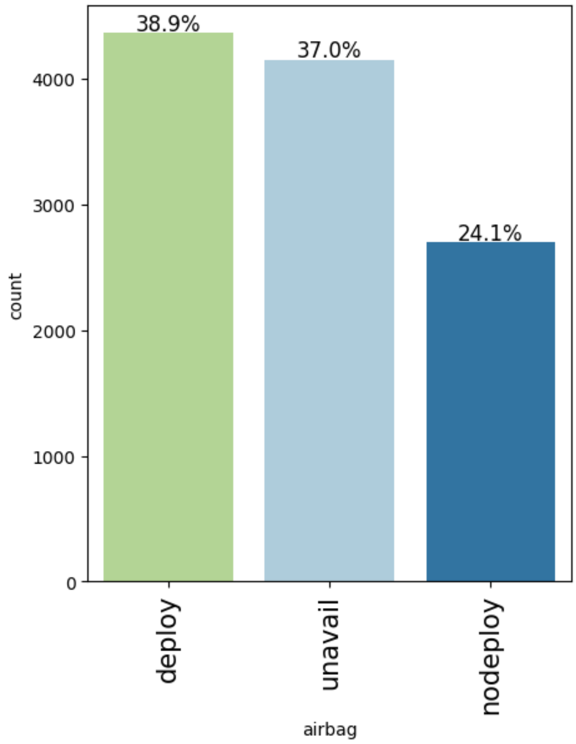

Fig-5

- Lack of airbag deployment contributes to fatalities. It is observed that around 24% accidents are caused due to no-deployment of airbags in cars and around 37% of the accidents are caused due to unavailability of airbag facility.

Observations on seatbelt

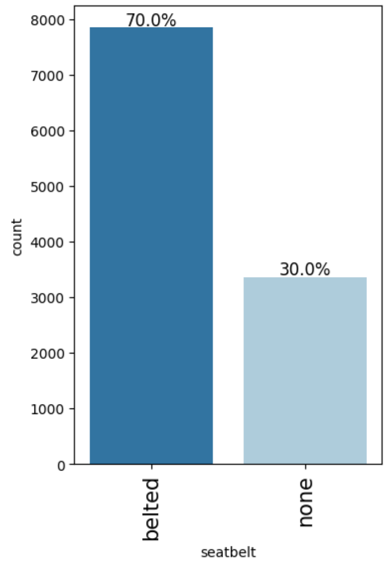

Fig-6

- Around 30% of the accident occurs for not wearing seatbelt.

Observations on frontal_impact

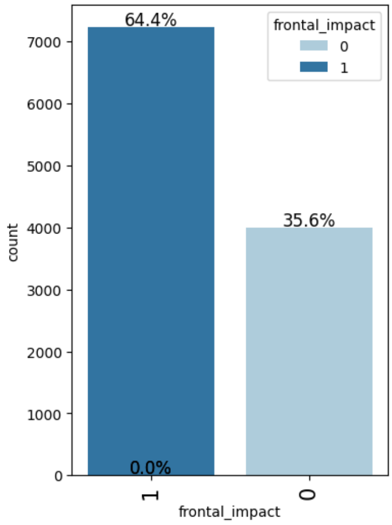

Fig-7

- Around 64.4% of the accidents are caused by the frontal impact, which contribute fatalities.

Observations on sex

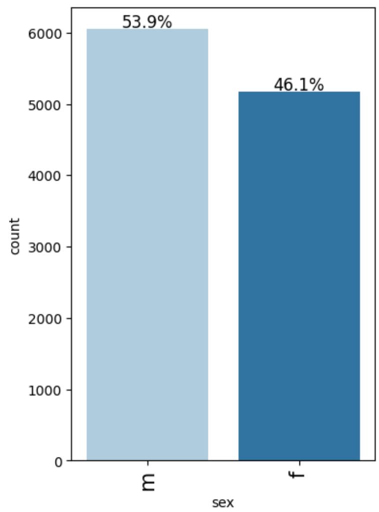

Fig-8

- Around 54% of the occupants are males and 46% are females.

Observations on model_year

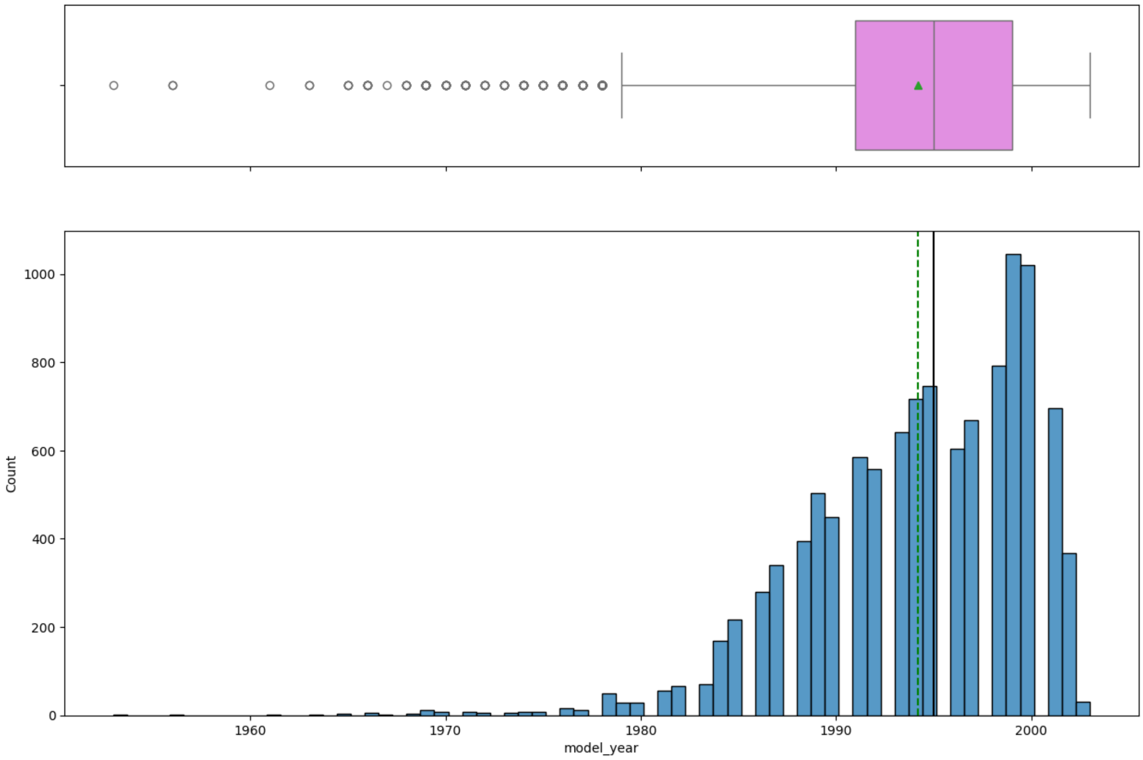

Fig-9

- Maximum cars were manufactured in the year 2000 as per the data.
- Outliers are observed from year 1960 to 1980.
- There is no skewness in data.

Observations on occ_role

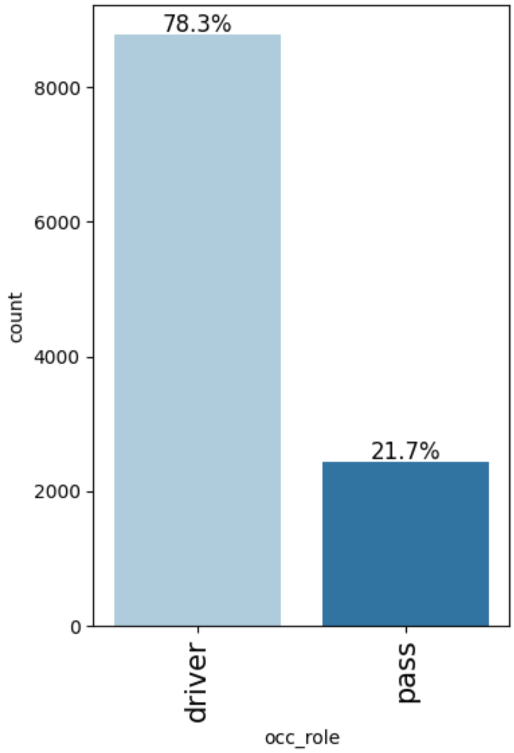

Fig-10

- Around 78% of the occupants were drivers and 22% are passengers.

Observations on veh_usage_duration

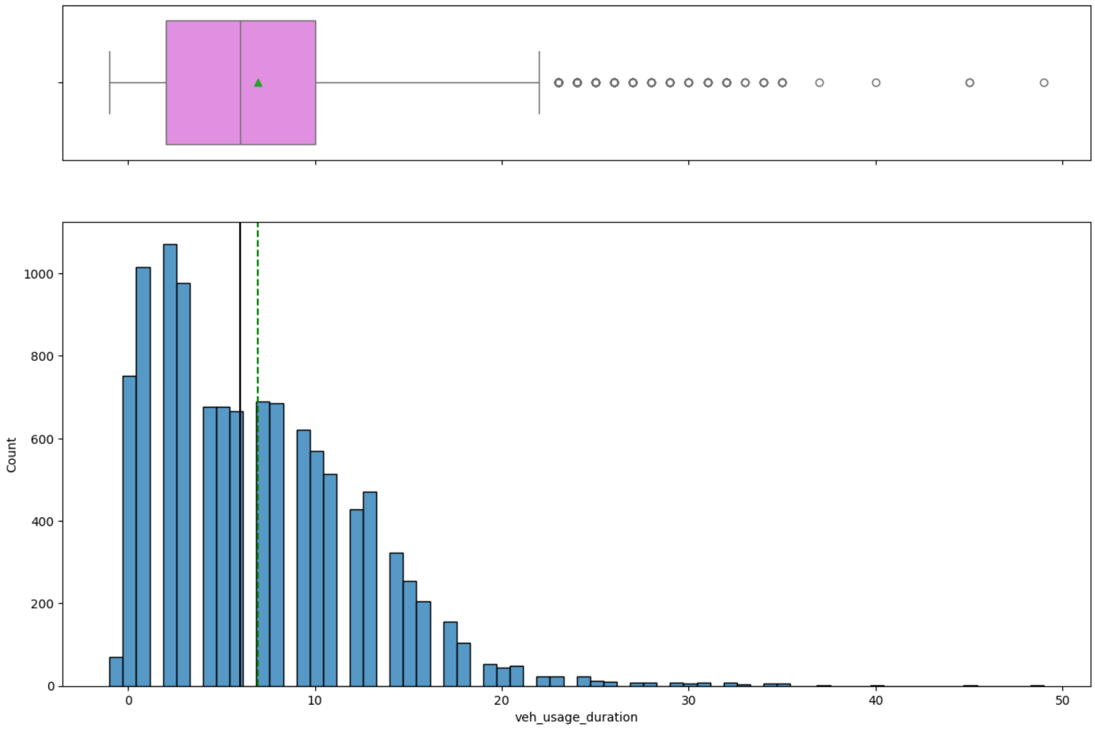

Fig-11

- Maximum usage duration of the car was recorded between 0 to 10 years.
- Outliers observed between 20-50 years.
- There is no skewness in the data.

3.2. Bivariate Analysis

- Used stacked bar-plot, distribution plots & correlation heatmaps to identify relationships with target variable deceased. Plots for each distribution are as follows:

Speed_range vs deceased

deceased speed_range	no	yes	All
All	10037	1180	11217
55+ km/h	394	415	809
40-54 km/h	1000	344	1344
25-39 km/h	3064	304	3368
10-24 km/h	5300	114	5414
1-9 km/h	279	3	282

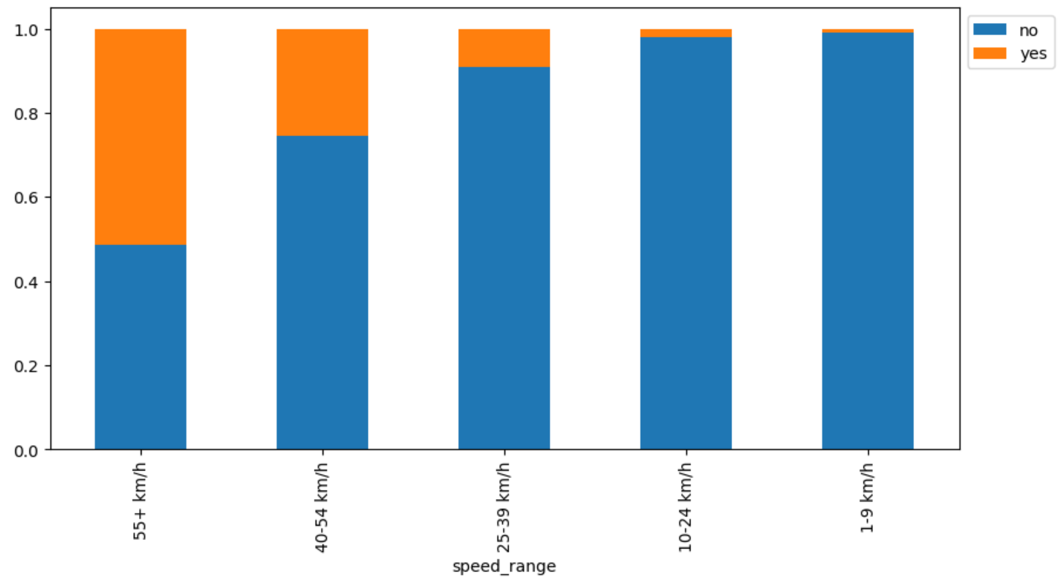

Fig-12

- Crashes frequent at high speeds (55+ km/h).

seatbelt vs deceased

deceased	no	yes	All
seatbelt			
All	10037	1180	11217
none	2688	680	3368
belted	7349	500	7849

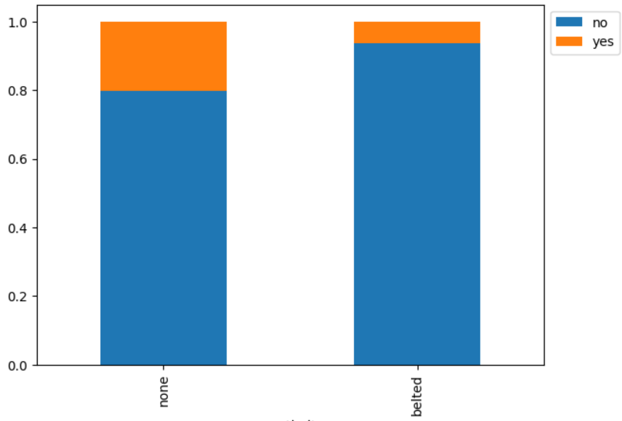

Fig-13

- Non-belted occupants have higher fatality rates.

frontal_impact vs deceased

deceased	no	yes	All
frontal_impact			
All	10037	1180	11217
0	3395	598	3993
1	6642	582	7224

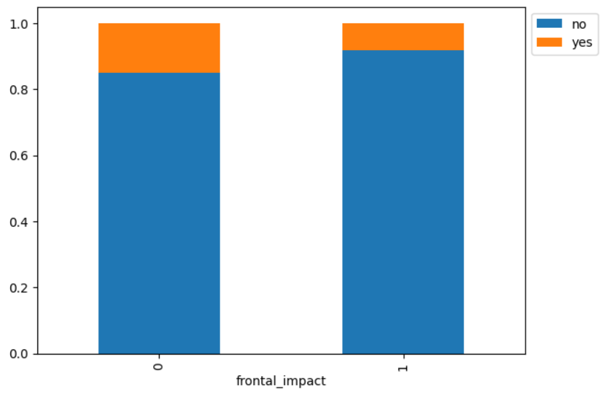

Fig-14

- Frontal impact contributes higher to fatalities.

sex vs deceased

deceased	no	yes	All
sex			
All	10037	1180	11217
m	5332	716	6048
f	4705	464	5169

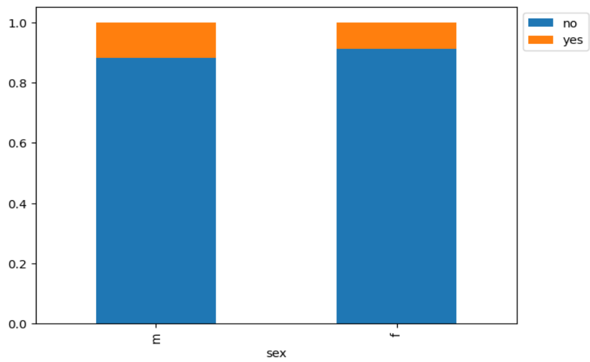

Fig-15

- Most deceased occupants are males.

airbag vs deceased

deceased	no	yes	All
airbag			
All	10037	1180	11217
unavail	3484	669	4153
deploy	3997	368	4365
nodeploy	2556	143	2699

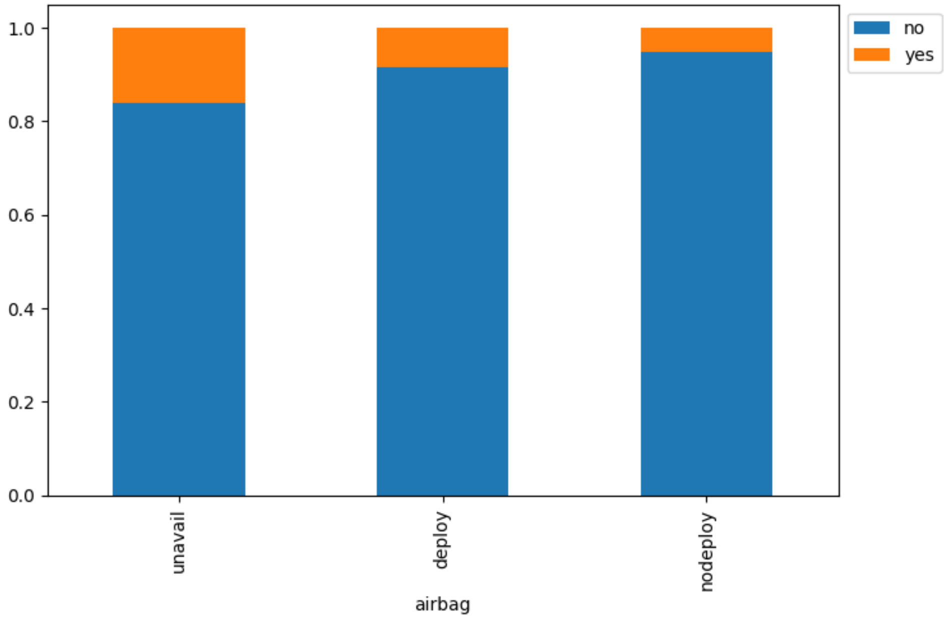

Fig-16

- Lack of airbag deployment contributes to fatalities.

occ_role vs deceased

deceased	no	yes	All
occ_role			
All	10037	1180	11217
driver	7895	891	8786
pass	2142	289	2431

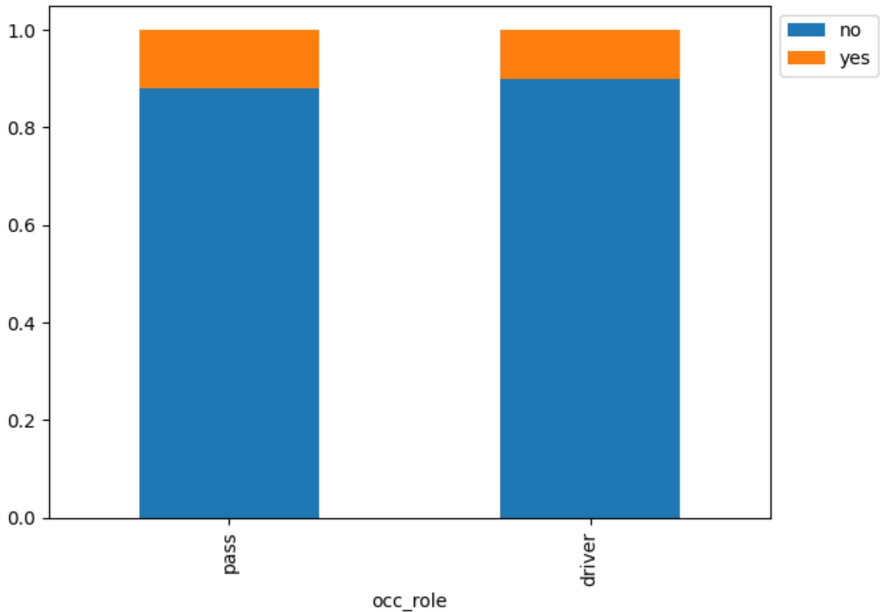

Fig-17

- Numbers of driver are more in the fatality count.

age_of_occ vs deceased

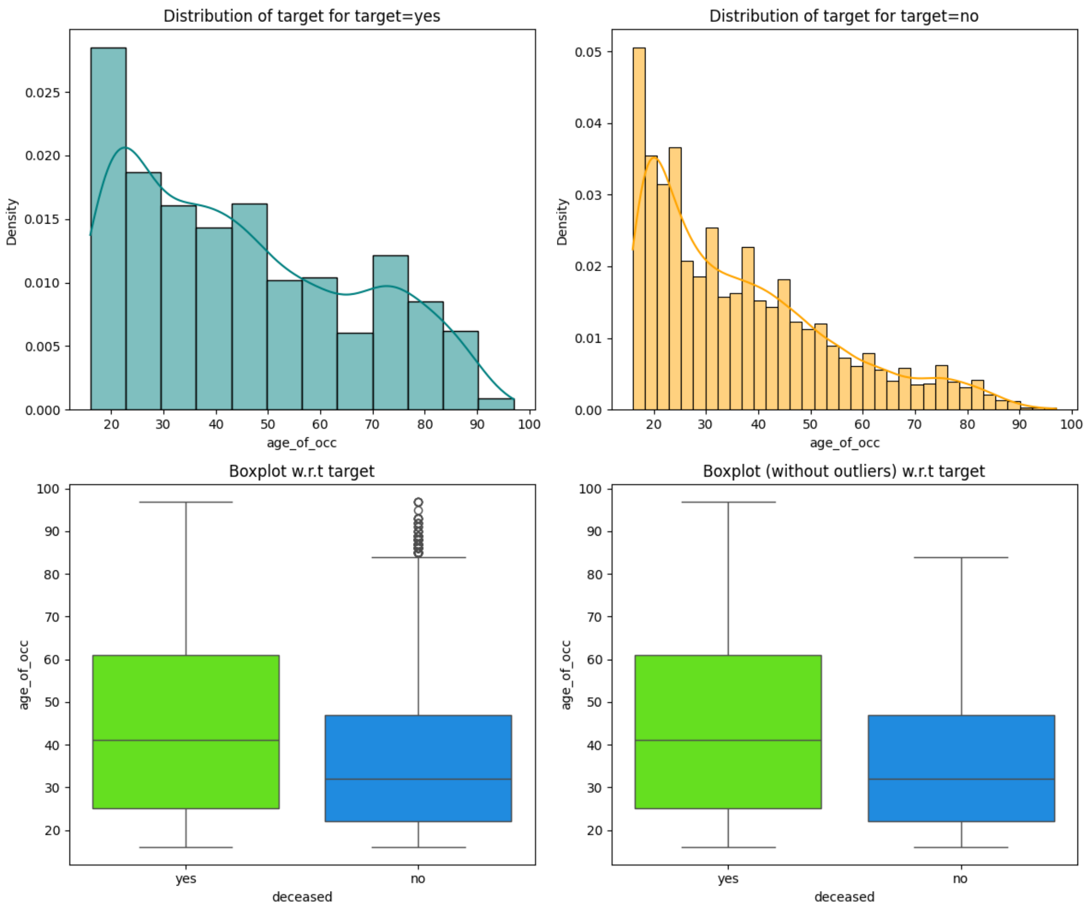

Fig-18

- Most occupants are between 15-45 years of age.

Veh_usage_duration vs deceased

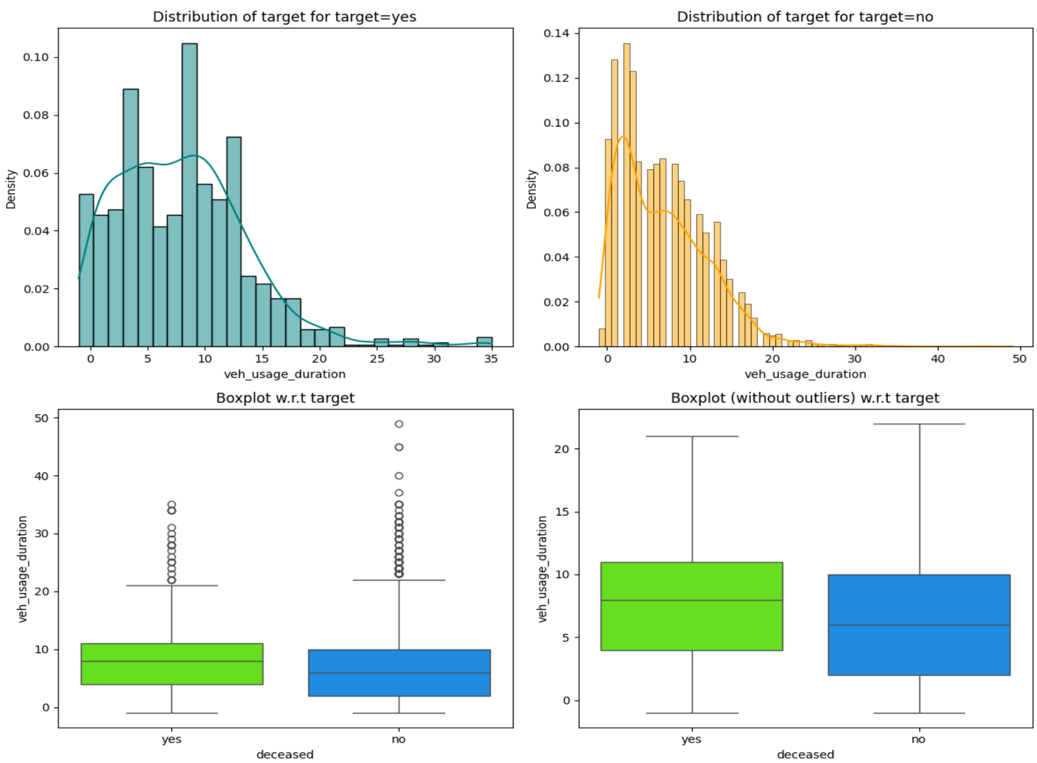

Fig-19

- Maximum vehicle usage duration is between 5-10 years.

Correlation Heatmap

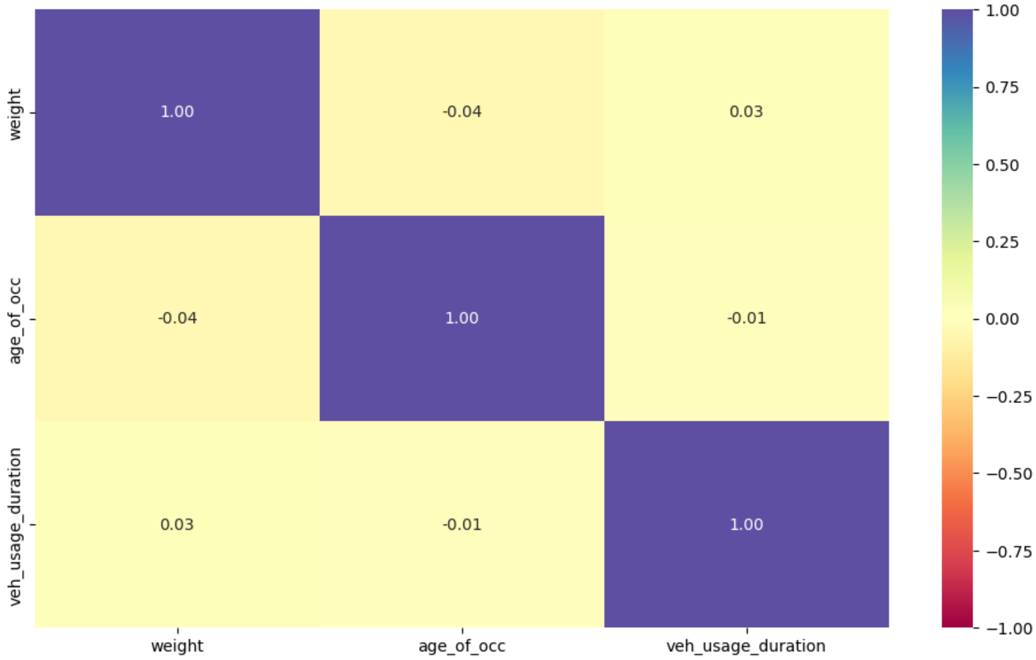

Fig-20

- We do not have very strong linear relationships between features. Except a few like weight and veh_usage_durion have a positive relationship. And age_of_occ and weight have a prominent negative relationship.

4. DATA PREPROCESSING

4.1. Outlier Check

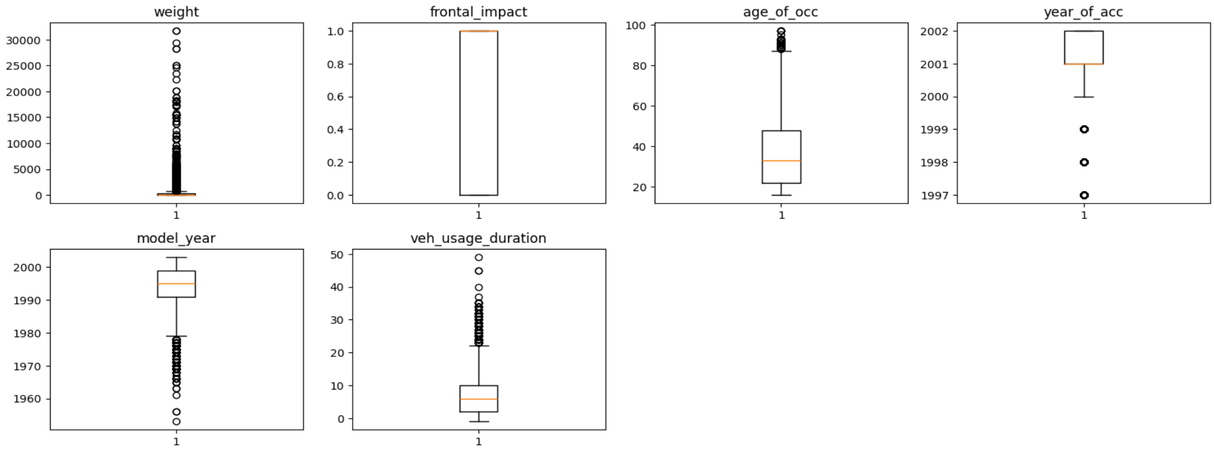

Fig-21

- We will not be treating outliers as it is not impacting our model building.

4.2. Data Preparation for modeling

- We will drop the unnecessary columns like caseid, year_of_acc & model_yearas these parameters don't contribute towards model building.

	speed_range	weight	seatbelt	frontal_impact	sex	age_of_occ	airbag	occ_role	deceased	veh_usage_duration
0	55+ km/h	27.07800	none	1	m	32	unavail	driver	yes	10
1	25-39 km/h	89.62700	belted	0	f	54	nodeploy	driver	yes	3
2	55+ km/h	27.07800	belted	1	m	67	unavail	driver	yes	5
3	55+ km/h	27.07800	belted	1	f	64	unavail	pass	yes	5
4	55+ km/h	13.37400	none	1	m	23	unavail	driver	yes	11

Table 5: Final dataset for modeling

- The data now looks clear and we are ready to build our prediction model.
- We have taken a test size of 30% and rest 70% is our train set.
- Data scaled and split into train-test (70:30).

5. MODEL BUILDING

5.1. Model evaluation criterion

- The model_performance_classification_sklearn function will be used to check the model performance of models.
- The confusion_matrix_sklearn function will be used to plot the confusion matrix.

Performance Metrics:

We will check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for the best performing model. We will Compare each model and write inferences, which model is best optimized.

5.2. Logistic regression

- A total accuracy score for train set in 91%.This is a good score for our prediction.
- A total accuracy score for test set is 91%.This is a good score for our prediction.

We observe that our model is able to generalize well as we have good and a balanced accuracy scores for train set and test set.

Following is the report of **train set** used and its **confusion matrix**.

	Accuracy	Recall	Precision	F1
0	0.91160	0.35645	0.63974	0.45781

Table 6

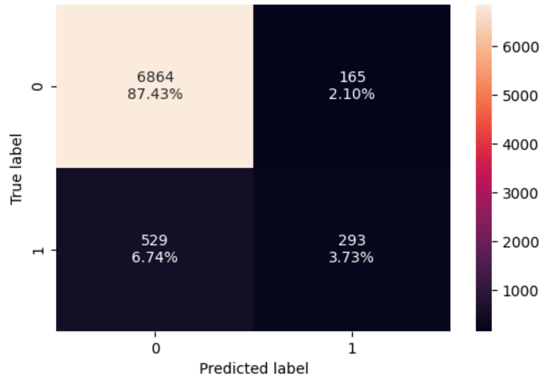

Fig-22

Following is the report of **test set** used and its **confusion matrix**.

	Accuracy	Recall	Precision	F1
0	0.91503	0.38268	0.67822	0.48929

Table 7

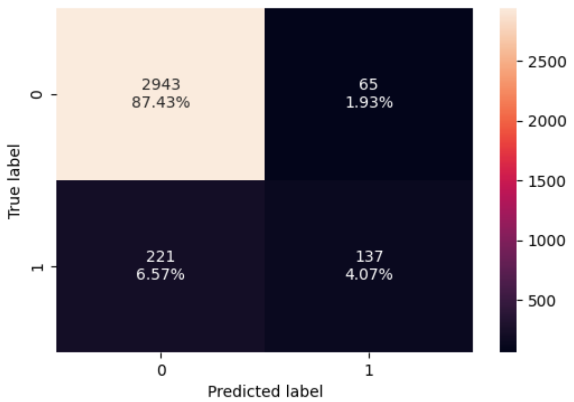

Fig-23

We can see in the confusion matrix that our model was able to predict 2943 plus 137 times right but it did not predict 221 plus 65 right.

This model seems very capable as the accuracy is very high.

5.2. Naive – Baye's Classifier

We have again taken the same data sets of train and test to build our model.

- A total accuracy score for train set is 81%, which is much lesser than logistic regression model.
- A total accuracy score for test set is 80%, which is much lesser than logistic regression model.

Following is the report of **train set** used and its **confusion matrix**.

	Accuracy	Recall	Precision	F1
0	0.81136	0.81022	0.33451	0.47352

Table 8

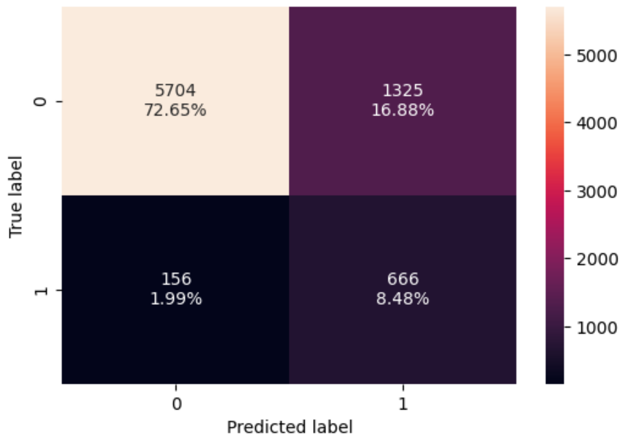

Fig-24

Following is the report of **test set** used and its **confusion matrix**.

	Accuracy	Recall	Precision	F1
0	0.80957	0.81564	0.33679	0.47673

Table 9

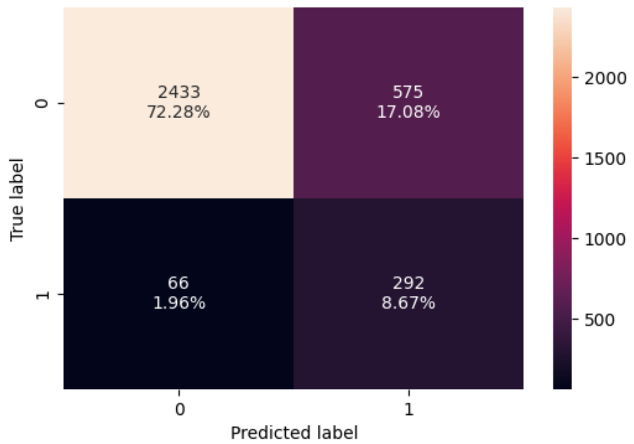

Fig-25

Thus, we can say that in this case study a logistic regression model performs far better than a Naive Baye's classifier model.

5.3. KNN Classifier (K = 3)

We have once again taken the same data sets of train and test to build our model.

- A total accuracy score for train set is 94%, which is greater than logistic regression model. This is a good score for our prediction.
- A total accuracy score for test set is 89%, which is lesser than logistic regression model.

Following is the report of **train set** used and its **confusion matrix**.

	Accuracy	Recall	Precision	F1
0	0.93810	0.59124	0.76415	0.66667

Table 10

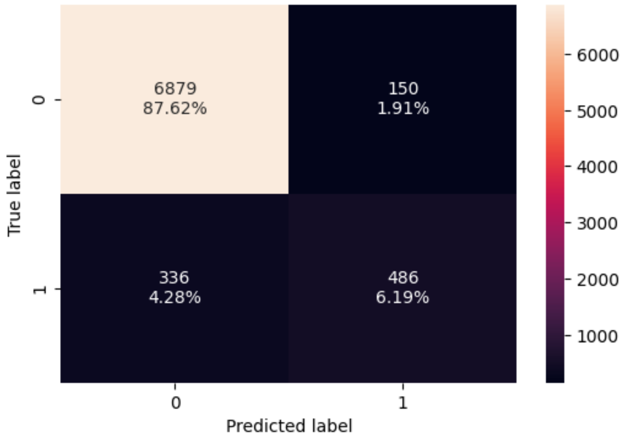

Fig-26

Following is the report of **test set** used and its **confusion matrix**.

	Accuracy	Recall	Precision	F1
0	0.88919	0.33799	0.47082	0.39350

Table 11

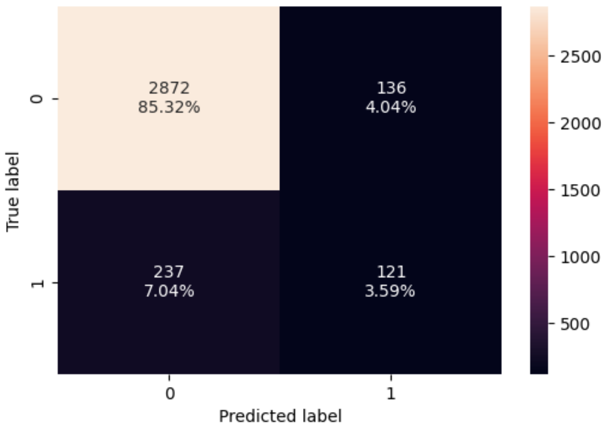

Fig-27

Thus, we can say that in this case study a logistic regression model performs better than a KNN classifier model.

5.4. Decision Tree Classifier

We have once again taken the same data sets of train and test to build our model.

- A total accuracy score for train set is 100%, which is greater than KNN Classifier model. This is a good score for our prediction.
- A total accuracy score for test set is 89%, which is lesser than logistic regression model.

Following is the report of **train set** used and its **confusion matrix**.

	Accuracy	Recall	Precision	F1
0	1.00000	1.00000	1.00000	1.00000

Table 12

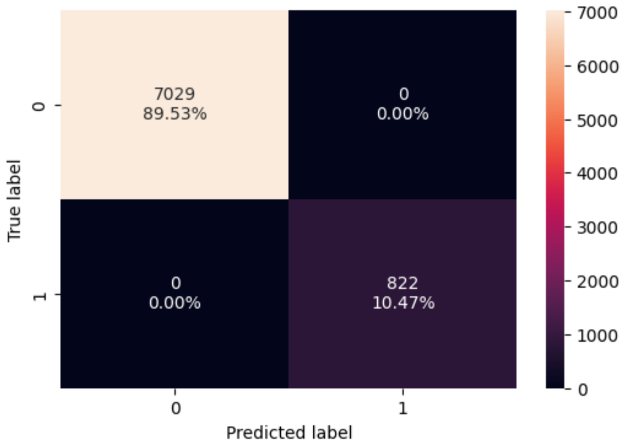

Fig-28

Following is the report of **test set** used and its **confusion matrix**.

	Accuracy	Recall	Precision	F1
0	0.88770	0.42458	0.46914	0.44575

Table 13

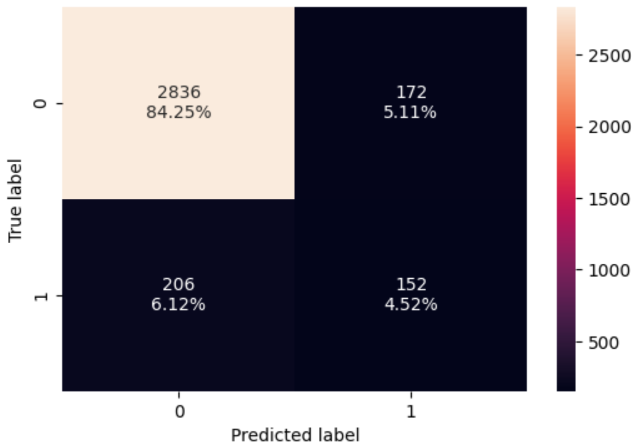

Fig-29

Thus, we can say that in this case study a logistic regression model performs better than a KNN classifier model.

6. MODEL PERFORMANCE IMPROVEMENT

6.1. Logistic Regression (optimal threshold)

- We will deal with high p-value variables and determine optimal threshold using ROC curve.

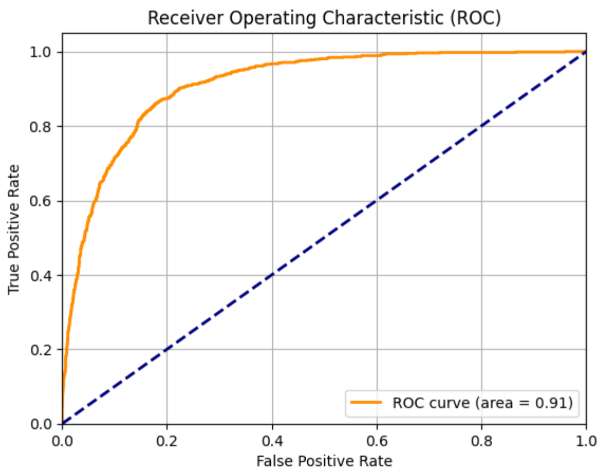

Fig-30: ROC Curve

Optimal Threshold: 0.111

Checking new Logistic Regression model performance on training set:

Following is the report of train set used and its confusion matrix.

	Accuracy	Recall	Precision	F1
0	0.82525	0.86131	0.36012	0.50789

Table 14

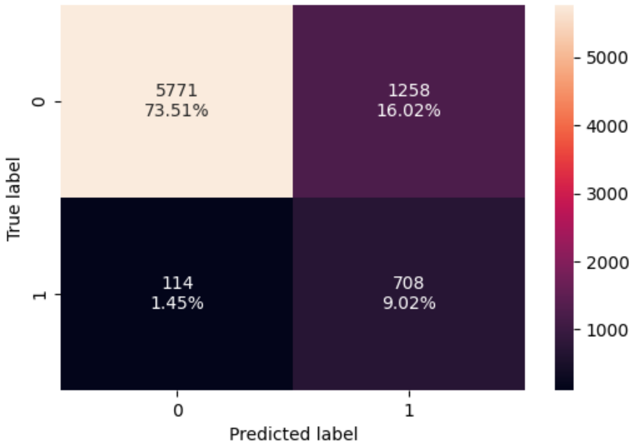

Fig-31

Checking tuned Logistic Regression model performance on test set:

Following is the report of test set used and its confusion matrix.

	Accuracy	Recall	Precision	F1
0	0.82620	0.85475	0.36472	0.51128

Table 15

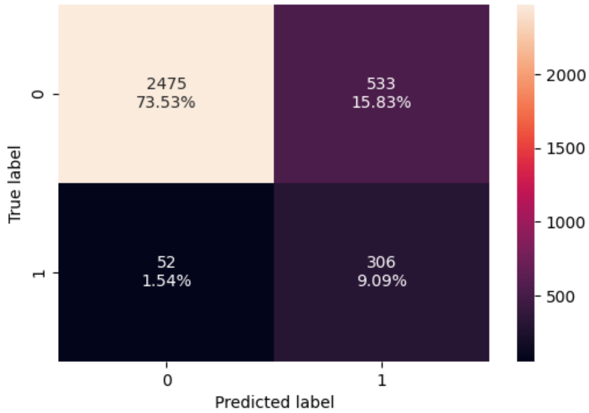

Fig-32

- A total accuracy score for train set is 82%, which is much lesser than KNN Classifier model & logistic regression base model.
- A total accuracy score for test set is 82%, which is lesser than logistic regression base model.

6.2. KNN Classifier (different values of K)

KNN Classifier Performance Improvement is performed by using different k values.

The best value of k is 2 with a recall of: 0.5912408759124088.

Checking tuned KNN model performance on training set:

Following is the report of train set used and its confusion matrix.

	Accuracy	Recall	Precision	F1
0	0.93746	0.40268	1.00000	0.57415

Table 16

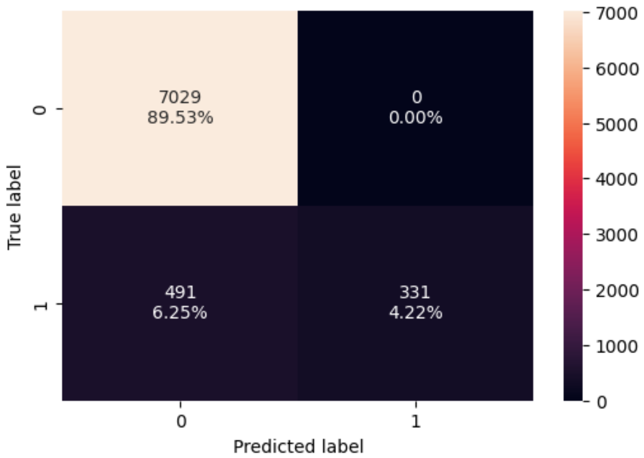

Fig-33

Checking tuned KNN model performance on test set:

Following is the report of test set used and its confusion matrix.

	Accuracy	Recall	Precision	F1
0	0.89275	0.17598	0.48837	0.25873

Table 17

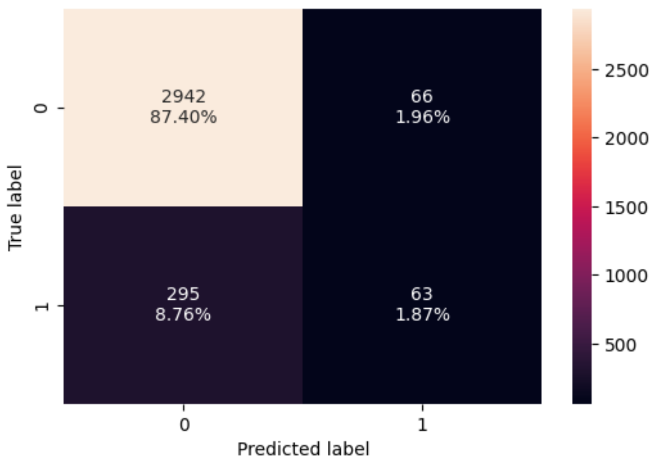

Fig-34

- A total accuracy score for train set is 94%, which is greater than logistic regression base model. This is a good model for prediction.
- A total accuracy score for test set is 89%, which is lesser than logistic regression base model.

6.3. Decision Tree Classifier (pre-pruning)

Checking tuned Decision Tree Classifier performance on training set:

Following is the report of train set used and its confusion matrix.

	Accuracy	Recall	Precision	F1
0	0.64820	0.86375	0.21131	0.33955

Table 18

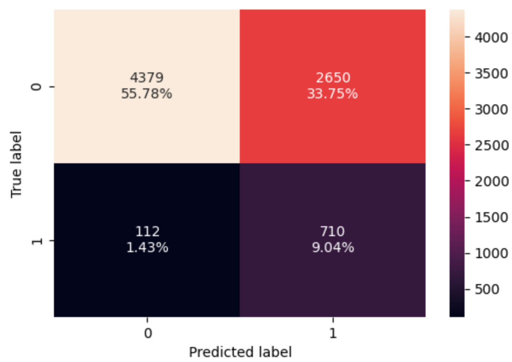

Fig-35

Checking tuned Decision Tree Classifier performance on test set:

Following is the report of **test set** used and its **confusion matrix**.

	Accuracy	Recall	Precision	F1
0	0.65270	0.87989	0.21860	0.35019

Table 19

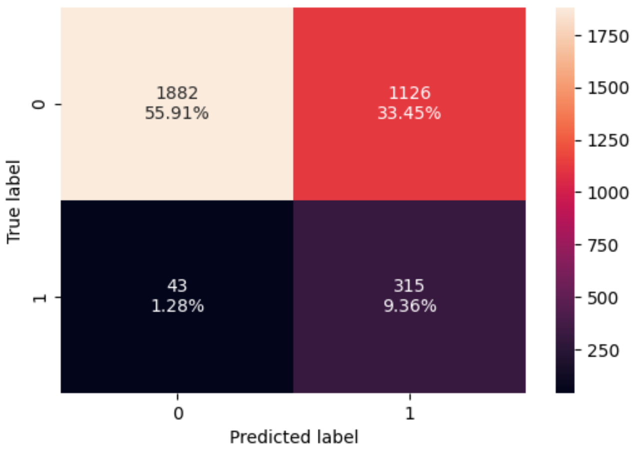

Fig-36

- A total accuracy score for train set is 64%, which is very much lesser than other models.
- A total accuracy score for test set is 65%, which is very much lesser than other models.

Thus, this model is not recommended for model prediction.

Visualizing the Decision Tree

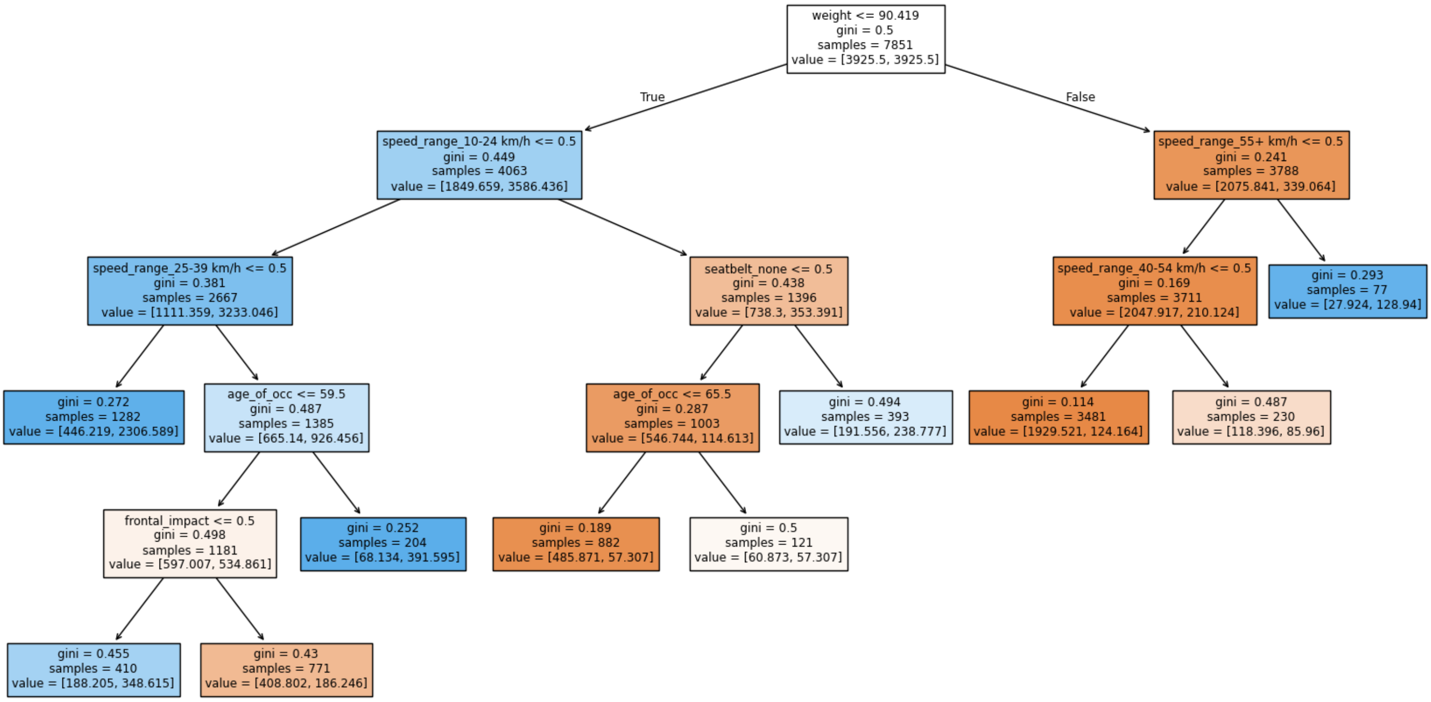

Fig-37

Observations from decision tree:

- **Primary Split**: The most important determinant is **vehicle weight**—specifically, a threshold at ~90.4 units.
- **Lighter vehicles** (weight ≤ ~90.4) have higher overall mortality risk.
- **Heavier vehicles** (weight > 90.4) tend to have better survival outcomes.

On the lighter-weight branch (≤ 90.4)

1. **Second Split: Impact Speed 10–24 km/h?**
   - If **no** (meaning speeds ≥ 25 km/h), survival is even lower.
2. **When speed is moderate (10–24 km/h)**:
   - Next split is **speed 25–39 km/h**:
     - If **no** (so ≤24)—better survival.
     - If **yes**, survival drops again.
     - Further, within 25–39 km/h:
       - **Older occupants (age > 59.5)** are at higher risk.
       - Within younger (≤59.5):
         - A frontal impact worsens outcomes.
3. **If impact speed is slower (<10 or >24 km/h)**:
   - Seatbelt usage becomes critical:
     - **No seatbelt** plus **older age** further reduces survival.
     - **Seatbelt use** improves survival even in slower impacts.

On the heavier-weight branch (> 90.4)

1. **Primary Split: Speed ≥ 55 km/h?**
   - If **yes**, outcomes are poorer.
   - If **no** (speed 40–54 km/h or lower), survival is significantly better.
2. **At moderate-high speeds (40–54 km/h)**:
   - It splits further:
     - One branch shows very low mortality—a relatively safe scenario.
     - The other shows increased risk—likely due to sub-factors not visually labelled (e.g., perhaps lack of safety features or older occupants).

Summary of Key Drivers

1. **Weight**: Heavier vehicles offer noticeably better protection.
2. **Speed**: As expected, higher speeds (especially above 40–55 km/h) dramatically worsen survival chances.
3. **Age**: Older occupants (esp. 60+) are at greater risk, even at moderate speeds.
4. **Seatbelt Usage**: Strongly protective—lack thereof significantly increases mortality risk.
5. **Frontal Impact**: Poses additional danger in younger occupants at moderate speeds.

Bottom Line

- **Heavier vehicles traveling at moderate speeds (40–54 km/h), with occupants wearing seatbelts, especially younger ones, show the best survival rates.**
- **Lighter vehicles, older occupants, high speeds (>55 km/h), no seatbelt, and frontal impacts compound risk significantly.**

These findings underscore the importance of enhancing vehicle structural strength, enforcing seatbelt use, and implementing speed restrictions—especially for older occupants and lighter vehicles.

- Logistic Regression tuned (feature drop, threshold tuning): F1 = 0.51, ROC = 0.91.
- KNN tuned to k=2: F1 = 0.35
- Decision Tree pruned: F1 = 0.35

7. MODEL PERFORMANCE COMPARISON & FINAL MODEL SELECTION

Training performance comparison:

	Logistic Regression Base	Logistic Regression (Optimal threshold)	Naive Bayes Base	KNN Base	KNN Tuned	Decision Tree Base	Decision Tree Tuned
Accuracy	0.91160	0.82525	0.81136	0.93810	0.93746	1.00000	0.64820
Recall	0.35645	0.86131	0.81022	0.59124	0.40268	1.00000	0.86375
Precision	0.63974	0.36012	0.33451	0.76415	1.00000	1.00000	0.21131
F1	0.45781	0.50789	0.47352	0.66667	0.57415	1.00000	0.33955

Table-20

Test set performance comparison:

	Logistic Regression Base	Logistic Regression (Optimal threshold)	Naive Bayes Base	KNN Base	KNN Tuned	Decision Tree Base	Decision Tree Tuned
Accuracy	0.91503	0.82620	0.80957	0.88919	0.89275	0.88770	0.65270
Recall	0.38268	0.85475	0.81564	0.33799	0.17598	0.42458	0.87989
Precision	0.67822	0.36472	0.33679	0.47082	0.48837	0.46914	0.21860
F1	0.48929	0.51128	0.47673	0.39350	0.25873	0.44575	0.35019

Table-21

Observations:

Using logistic regression model, we can say
For {Passengers who did not survive (Label 0)}:
Precision (68%) – 67% of passengers who did not survive are correctly predicted, out of all passengers who did not survive that are predicted.
Recall (38%) – Out of all the passengers who actually did not survive, 38% of passengers who did not survive have been predicted correctly.
For {Passengers who did survive (Label 1)}:
Precision (68%) – 68% of Passengers who did survive are correctly predicted, out of all passengers who had accident that are predicted.
Recall (38%) – Out of all the passengers who actually did survive, 38% of Customers who did Churn have been correctly predicted.
Accuracy, AUC, Precision and Recall for test data is almost in line with training data. This proves no overfitting or underfitting has happened, and overall, the model is a good model for classification.

7.1. Final Model Selection

- Logistic Regression selected for best balance and interpretability.
- Top predictors: seatbelt usage, impact type, speed, age, airbag status.

8. ACTIONABLE INSIGHTS & RECOMMENDATIONS

8.1. Actionable Insights

- Seatbelt use greatly improves survival.
- Crashes above 55 km/h have higher fatalities.
- Frontal impacts are more dangerous.
- Airbag deployment reduces risk.

8.2. Recommendations

- Mandate advanced seatbelt alerts.
- Enforce speed governance in urban areas.
- Require airbag deployment sensors in all cars.
- Educate or restrict elderly drivers based on risk.

8.3. Conclusion

This analysis provides a data-driven foundation for understanding and predicting crash survival outcomes. Implementing the recommendations can significantly improve road safety and guide safer vehicle designs.