CS452/CS552 Assignment 2: Car Rollover Prediction

Eyüb Ensari Topcan S018649

Introduction

Today and in the near future, traffic accidents have endangered our health and property safety with the invention of the car and have led to many material damages, injuries and unfortunately deaths. There are many factors that cause traffic accidents to happen. When we divide these into two as environmental and human factors, inattention, insomnia, alcohol use and non-compliance with rules, etc. human factors, weather conditions, road conditions and vehicle qualities etc. creates environmental factors. Data science is processing and manipulating the data obtained by various methods to make it meaningful, and then insights are extracted. By taking action with this insight, solutions are produced to the problems that the data is related to. By optimizing these solutions, the desired aspects of the data are increased and the undesirable aspects are reduced. Data science techniques are used to optimize this process.

In the case we have, we are an intern at an insurance company and there is a dataset obtained by national Highway Traffic Safety Administration (NHTSA) at United States Department of Transportation provides Fatality Analysis Reporting System (FARS) . Our goal is to make sense of the benefits of our company from the results we obtained using the features in the FARS dataset. The FARS dataset includes data on how traffic accidents that took place in 1975-2019 were affected by vehicle, person, damage, safety, race and other factors. In our research, we will use the data set of vehicle and person related factors of accidents between 2014-2019.

The vehicle data file contains the characteristics of the vehicles involved in the accident. In which state and district the accident occurred, the make and model of the accident vehicle, in which factory it was produced, the type of vehicle, the type of fuel used, the type of engine, the segment of vehicle, information about the sales price, weight, number of doors and wheels, traction system of the vehicle. whether it was a public transportation vehicle, the month, day, hour and It includes many features such as the minute, how the arc was damaged after the accident, and whether there was a rollover.

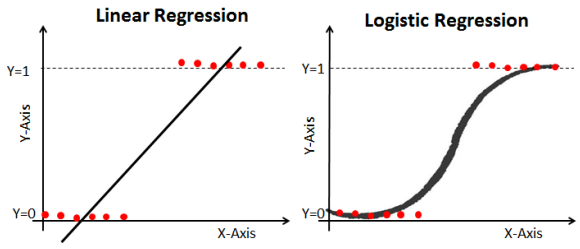
The Person data file contains the features of the persons who caused or were affected by the accident. These features include the age, gender, race of the driver, as well as details in case of injury or death as a result of the accident, what seat he was sitting in at the time of the accident, the type of seat belt and whether it was used, whether there was an airbag. It also includes many features such as whether the people involved in the accident used alcohol, whether alcohol use was detected by police control, alcohol test results, drug use and test results.

As an insurance company, we need to extract the necessary information from the data in this FARS dataset to be able to take action and make an insurance contract. Thus, we do not include the information about the accident moment and aftermath in the dataset into our research. However, knowing the country, neighborhood, time, alcohol/drug use and use of alcohol can improve our perspective on accidents, so I wanted to include them as well. In the conclusion part, I will talk about the precautions that our insurance company can take.

In a nutshell, we use the FARS dataset to research for use by the insurance company we're interning with. Our goal is to find the instrumental and personal factors that cause accidents and add them to our business model, and to raise awareness of our customers, as well as to determine our pricing policy accordingly and maximize our earnings. Our aim is to import the data we will use from the data source to our notebook, det ermine the dates we will use, perform Exploratory data analysis, clean the data and make it suitable for sep arating the data by performing operations on features and training for machine learning algorithms. At the s ame time, it is our goal to find the features that increase the probability of accident (highly correlated), run t he algorithms on our models and visualize the results, evaluate the performance of these results, find the m ost effective machine learning and extract it from the results we have obtained as a result of applying it or o ther machine learning algorithms in the most feasible way.

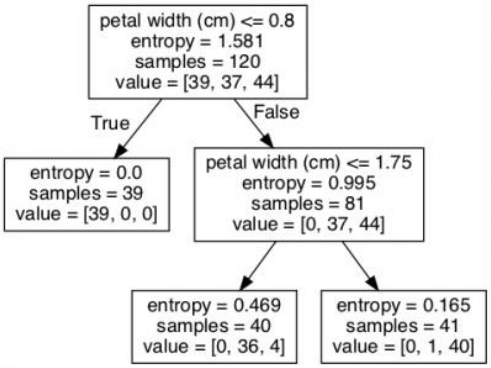
The machine learning algorithms we will use are described below

Logistic Regression is used when the target variable is categorical. It makes binary classification of the feature to be predicted. For example, in the FARS dataset, we calculate the prediction of whether the vehicle rollover as a result of an accident, via Logistic regression. When using logistic regression, our model is likened to a sigmoid function, so the feature is classified as binary (0, 1).

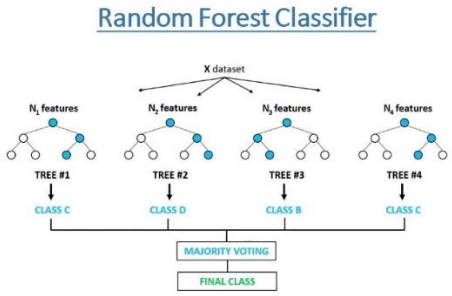


[1]

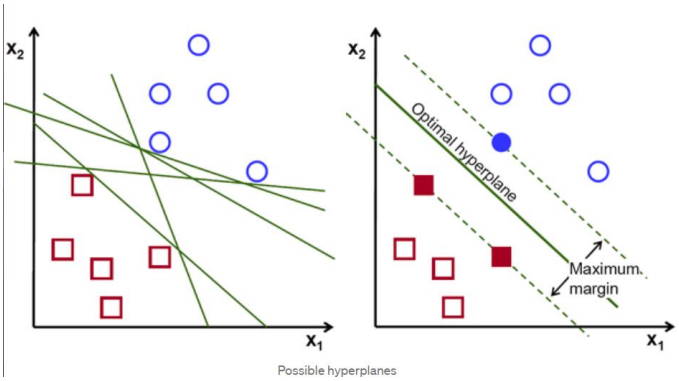
Decision Tree is a non-parametric supervised learning algorithm. It is used for classification of how the target variable will be shaped according to other features and the connection with each other when deciding. The created tree can be visualized.



Random Forests is a supervised learning technique that consists of ensembling decision trees. It allows us to obtain a more accurate final result by combining the conditions provided by different decision trees. For example, while one decision tree finds the most corelated feature of the vehicle data, another decision tree finds the most correlated feature of the person data, and the random forest algorithm combines these two results to obtain the feature that has the most effect on the rollover risk. [2]



Support Vector Machine, supervised learning method is a machine learning algorithm that finds a decision boundary between the two classes that are furthest from any point in the training data. SVM searches for hyperplanes in the n-dimensional space that it extracts from the dataset, and divides the data points into 2 separate classes by distically classifying them. Thus, 2 separate classes with maximum margin are obtained and classification is completed. Support vectors are data points that are closest to each other but in two different classes, where the margin starts.



[4]

Naive Bayes is a supervised learning algorithm that uses the Bayes theorem to reveal the relationship between binary features in conditional probability. It has various methods to predict data in different distributions.

Gaussian Naive Bayes algorithm makes classification by returning the predict method over the likelihood values of the features. Also calculates the posterior Porbability.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Categorical Naive Bayes algorithm is used when applying naive bayes to a dataset with categorical features. It enables continuous features to be discrete and encoded between 0-n and taken into the calculation.

$$P(x_i = t \mid y = c\,;\, \alpha) = \frac{N_{tic} + \alpha}{N_c + \alpha n_i},$$

Complement Naive Bayes is used when the d ataset contains Multinomial features and is imb alanced. For example, it can be used when clas sifying text. It ensures that the efficiency of the code does not decrease while calculating the conditional probabilities of the features.

$$\hat{\theta}_{ci} = \frac{\alpha_i + \sum_{j:y_j \neq c} d_{ij}}{\alpha + \sum_{j:y_j \neq c} \sum_k d_{kj}}$$

$$w_{ci} = \log \hat{\theta}_{ci}$$

$$w_{ci} = \frac{w_{ci}}{\sum_j |w_{cj}|}$$

[5]

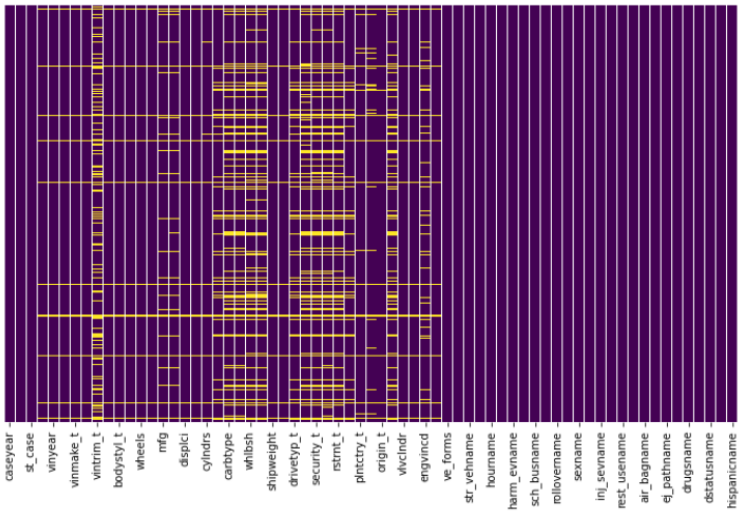
Multinomial Naïve Bayes, It is a supervised lear ning technique that works like categorical naive bayes when applying Bayes theorem to multiple features, but unlike it, it deals with multiple featu res using multinomial distribution calculation.

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Implementation

We import the FARS dataset from the given website so that we can make sense of our research in accordance with its purpose. NHTSA has made this dataset available to us as an API. While importing our dataset, we get the data of the years 2014-2019, because all the information that we can define as up-to-date was realized after 2014. While importing our data, we create 2 separate dataframes as persons and vehicles, then we drop methods that we will not use in our research. In this drop-down process, we also dropped the columns that were already encoded, the data collected during and after the accident, and the columns with empty rows in our research. To visualize this process, I found it appropriate to drop the columns marked in orange in the table below.

Dropped Columns due to Missing Values and its visualization



Vehicles Dataset

caseyear : 0	cylndrs : 12138	drivetyp_t : 39828	tkbedl_t : 209136	turbo : 216060
state : 0	cycles : 251124	salectry : 278310	segmnt : 39474	turbo_t : 216060
statename : 0	fuel : 39504	salectry_t : 278310	segmnt_t : 39474	engvvt : 225258
st_case : 0	fuel_t : 39504	abs : 66480	plant : 11922	mcyusage : 251100
veh_no : 0	fuelinj : 78006	abs_t : 75438	pltctry_t : 15438	mcyusage_t : 251100
ncicmake : 11898	fuelinj_t : 78006	security : 67356	pltcity : 19860	
vinyear : 11898	carbtype : 62226	security_t : 67404	pltctry : 15438	
vehtype : 11898	carbtype_t : 62226	drl : 67932	pltstat : 80358	
vehtype_t : 11898	carbbrls : 277092	drl_t : 67932	pltstat_t : 80358	
vinmake_t : 11898	gvwrange : 143088	rstrnt : 63138	origin : 13428	
vinmodel_t : 11916	gvwrange_t : 143088	rstrnt_t : 63138	origin_t : 13428	
vintrim_t : 74706	whlbsh : 61158	tkcab : 143112	**dispolmt : 49284**	
vintrim1_t : 238614	whlblg : 61164	tkcab_t : 143112	blocktype : 63660	
vintrim2_t : 262014	tiredesc_f : 278574	tkaxlef : 151110	enghead : 126198	
vintrim3_t : 273612	psi_f : 11898	tkaxlef_t : 151110	enghead_t : 126198	
vintrim4_t : 278430	tiresz_f : 117744	tkaxler : 149160	vlvclndr : 11898	
bodystyl : 11898	tiresz_f_t : 117744	tkaxler_t : 149160	vlvtotal : 11898	
bodystyl_t : 11898	tiredesc_r : 278574	tkbrak : 146820	engvincd : 30444	
doors : 11898	psi_r : 11898	tkbrak_t : 146820	incomplt : 11898	
wheels : 11898	rearsize : 232458	engmfg : 175320	battyp : 276924	
drivwhls : 11898	rearsize_t : 232458	engmfg_t : 175320	battyp_t : 259656	
mfg : 18522	tonrating : 203088	engmodel : 194688	batkwrtg : 276924	
mfg_t : 18630	shipweight : 11898	tkduty : 152514	batvolt : 276936	
displci : 11898	msrp : 11898	tkduty_t : 152514	supchrgr : 223848	
displcc : 11898	drivetyp : 39828	tkbedl : 209136	supchrgr_t : 223848	

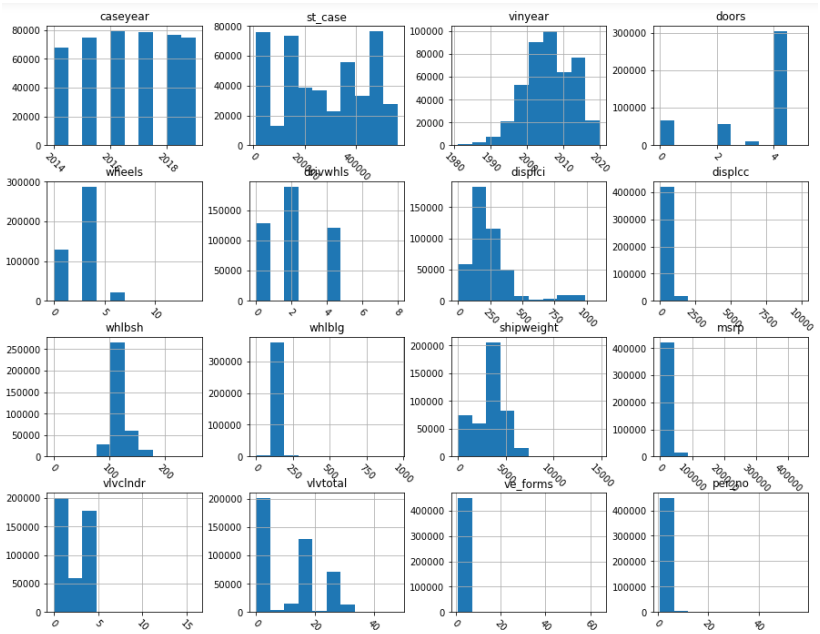
Persons Dataset

caseyear : 0	sch_busname : 0	per_typname : 0	alc_res : 0	death_daname : 0	location : 0
state : 0	make : 45092	inj_sev : 0	alc_resname : 0	death_mo : 0	locationname : 0
statename : 0	makename : 45092	inj_sevname : 0	drugs : 0	death_moname : 0	func_sys : 74348
st_case : 0	mak_mod : 45092	seat_pos : 0	drugsname : 0	death_yr : 0	func_sysname : 74348
ve_forms : 0	body_typ : 45092	seat_posname : 0	drug_det : 0	death_yrname : 0	rur_urb : 74348
veh_no : 0	body_typname : 45092	rest_use : 0	drug_detname : 0	death_hr : 0	rur_urbname : 74348
per_no : 0	mod_year : 45092	rest_usename : 0	dstatus : 0	death_hrname : 0	
str_veh : 0	mod_yearname : 45092	rest_mis : 0	dstatusname : 0	death_mn : 0	
str_vehname : 0	tow_veh : 45092	rest_misname : 0	drugtst1 : 167900	death_mnname : 0	
county : 0	tow_vehname : 45092	air_bag : 0	drugtst1name : 167900	death_tm : 0	
countyname : 0	spec_use : 45092	air_bagname : 0	drugres1 : 167900	lag_hrs : 0	
day : 0	spec_usename : 45092	ejection : 0	drugres1name : 167900	lag_mins : 0	
month : 0	emer_use : 45092	ejectionname : 0	drugtst2 : 167900	p_sf1 : 0	
monthname : 0	emer_usename : 45092	ej_path : 0	drugtst2name : 167900	p_sf1name : 0	
hour : 0	rollover : 45092	ej_pathname : 0	drugres2 : 167900	p_sf2 : 0	
hourname : 0	rollovername : 45092	extricat : 0	drugres2name : 167900	p_sf2name : 0	
minute : 0	impact1 : 45092	extricatname : 0	drugtst3 : 167900	p_sf3 : 0	
minutename : 0	impact1name : 45092	drinking : 0	drugtst3name : 167900	p_sf3name : 0	
road_fnc : 423654	fire_exp : 45092	drinkingname : 0	drugres3 : 167900	cert_no : 423654	
road_fncname : 423654	fire_expname : 45092	alc_det : 0	drugres3name : 167900	work_inj : 0	
harm_ev : 0	age : 0	alc_detname : 0	hospital : 0	work_injname : 0	
harm_evname : 0	agename : 0	alc_status : 0	hospitalname : 0	hispanic : 0	
man_coll : 0	sex : 0	alc_statusname : 0	doa : 0	hispanicname : 0	
man_collname : 0	sexname : 0	atst_typ : 0	doaname : 0	race : 82884	
sch_bus : 0	per_typ : 0	atst_typname : 0	death_da : 0	racename : 82884	

After importing 2 datasets, I merge them as a single DataFrame named peoples and vehicles, accidents. When I call the Merge method, I perform a merge operation over the common columns 'caseyear', 'st_case', 'veh_no', 'statename' in both datasets. When we look at the properties of our Persons dataframe, we see that it has 498002 rows and 131 columns. Our Vehicles dataframe has 313877 rows, 105 columns. I transfer the dataframe formed after dropping the during and after crash data in our dataset to my computer with the to_csv() method. Thus, instead of importing from API every time I run, I can import in csv file and save time.

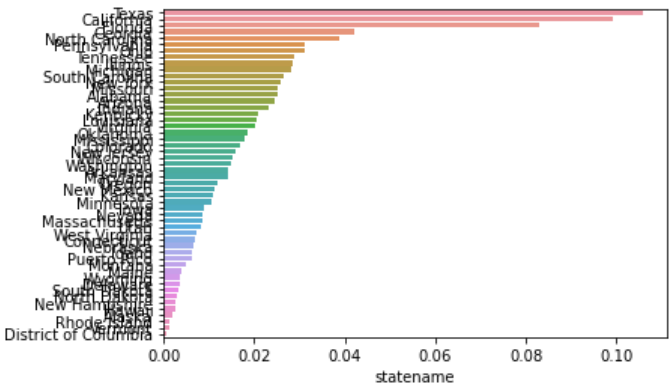
Numeric Features

In the exploratory data analysis part of our research, we apply an initial investigation to our data. Thus, we can have more information about our data. First of all, we extract the categorical and numeric features of our dataset and select our target value. Our target value is the rollovername feature. rollovername gives information about whether the vehicle rolled over as a result of the accident. After we drop the duplicate columns in our dataset, we pour them into the histogram graph. We visualized the value counts of numeric features in the histogram graph.

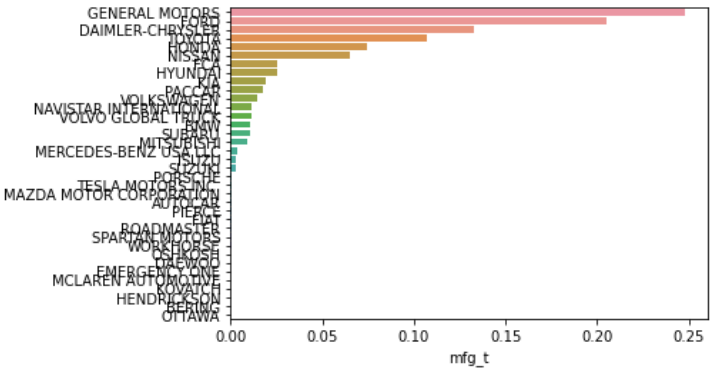


When we look at the histogram graphs of our data, it is calculated whether there is a rollover according to the price of the vehicle in the msrp column. From this we can conclude that cheaper cars (less than $25,000) have more rollovers. It is a feature that shows the production date in the vin year column, and when we look at the histogram graph, we observe that vehicles produced between 2008-2012 are involved in more accidents.

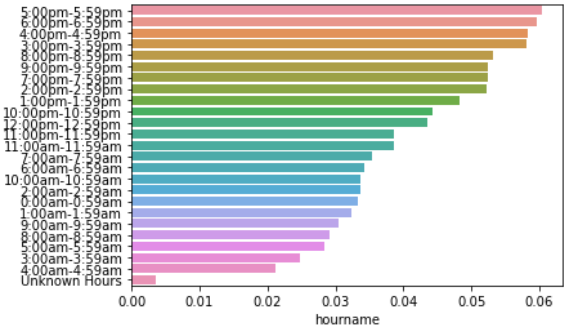
Categorical Features

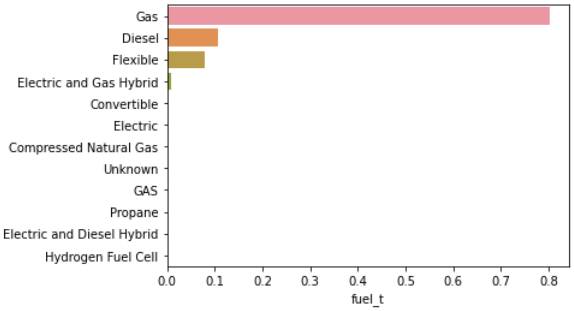


Statename columns visualize the what percentage of accidents occur in which state in USA.



mgf_t column shows that vehicles brands

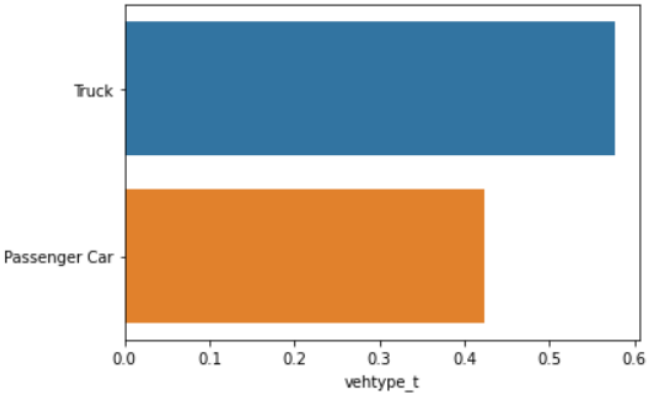


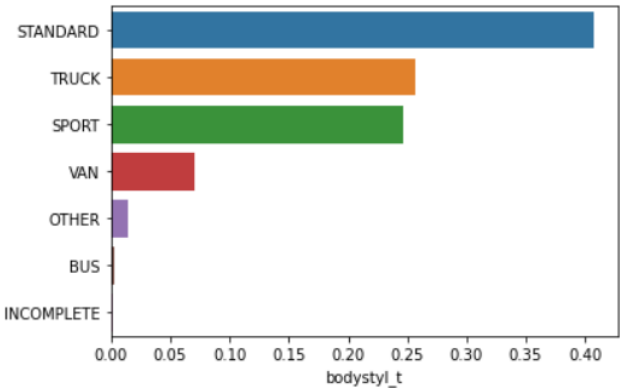


hourname feature shows that interval of time when accident occurs. We can say that after work shifts over is the most accident occurs.

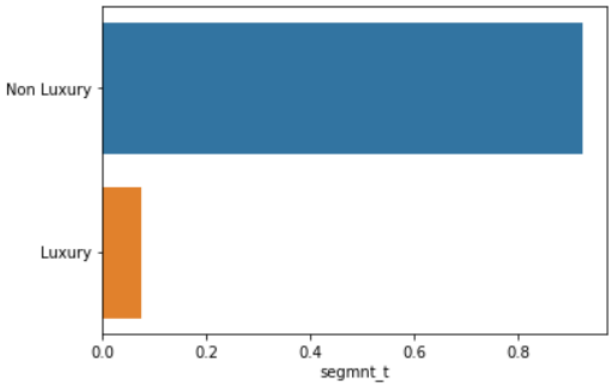
Fuel_t column shows that what percentage of accidents that vehicles involved uses which type of fuel.

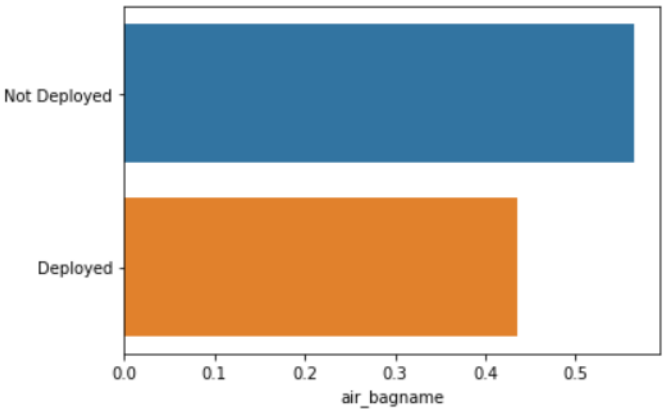
You can find more qualitied graphs on my notebook..



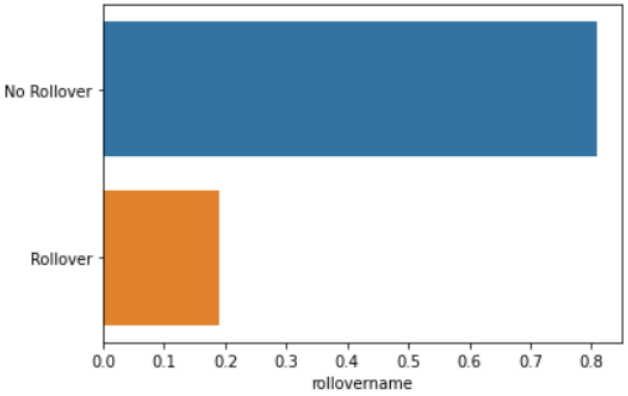


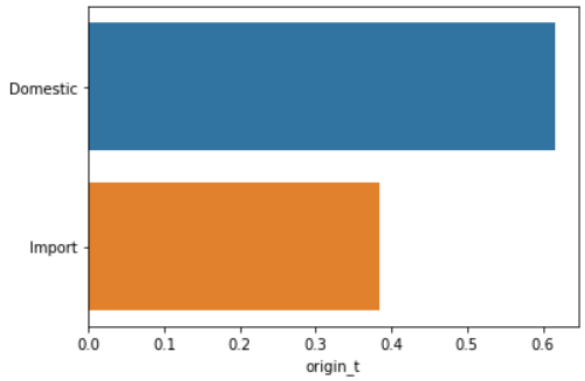
When we examine the barplots of categorical features, we observe that trucks with trucks are more involved in accidents than passenger cars. When we kept our observations more specific and examined the body styles of the vehicles in 7 different categories, vehicles with sedans, coupes and hatchbacks were involved in more accidents. The insurance company may follow a different pricing policy for these vehicles.



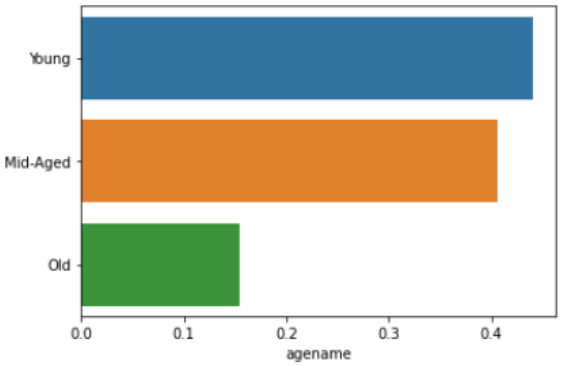


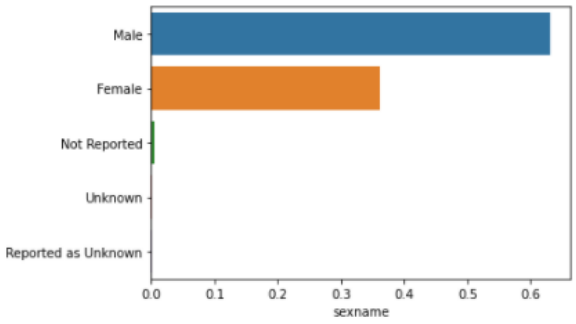
My other conclusion is that vehicles whose segment is described as non-luxury are involved in more accidents than luxury ones. Vehicles with airbags also had more accidents than vehicles without them. From here, we can accept non-luxury vehicles as potential customers, ask them if they have airbags to insure their vehicles, transfer this data and convince them.



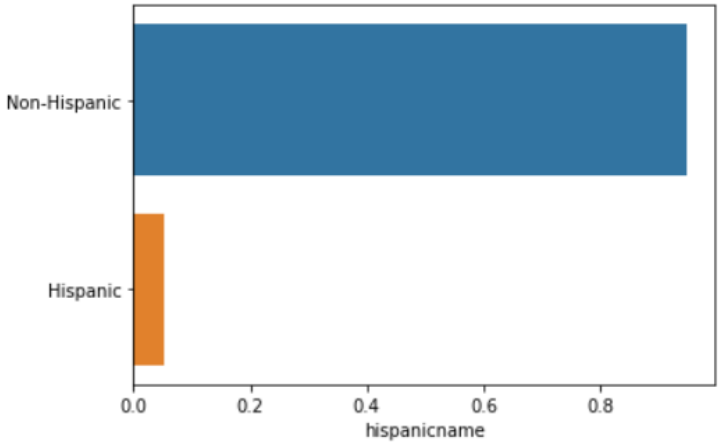


In the barplot of the rollovername feature, the relationship between the rollover status and the shipweight of the vehicles involved in the accident can be examined. In addition, if the origin of the vehicle is Domestic, that is, if the vehicle is produced in America, the accident rate is higher.

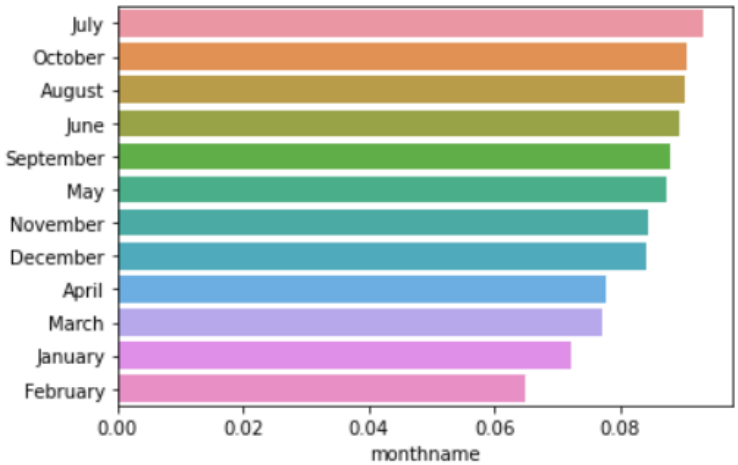




When we examine the ages and genders of the persons involved in the accident, the persons in the young and mid-aged categories were involved in more accidents than the elderly. In addition, the fact that women were involved in fewer accidents than men in our research is an element that our insurance company should pay attention to when pricing and customer analysis.



In our last barplot, it is mentioned whether the person is of Spanish origin or not. This unethical feature of ours has revealed that non-Hispanic people are involved in more accidents and reminded us once again that we should not be racist.

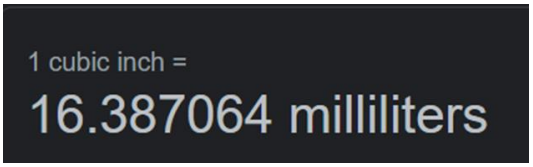


Statistical table

When we extract the statistical table of our dataset, we get the attached image. In this image, we can reach the count, mean, min-max, standard deviation and 25th, 50th, 75th quartiles of each feature. If we look at the statistical scores while making data interpretation, our job becomes easier when we make feature selection and we can make a more accurate classification.

	caseyear	statename	vinyear	vehtype_t	bodystyl_t	doors	wheels	drivwhls	mfg_t	displci	cylndrs	fuel_t	shipweight	msrp
count	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670
mean	2017	23.9	2006.3	0.6	3.1	3.4	3.2	2.1	13	252.8	9.2	7.2	3536.3	23807.6
std	1.7	15.7	6.8	0.5	1.9	1.2	1.8	1.5	9.1	163.8	2.4	1.9	1483	12177.4
min	2014	0	1981	0	0	0	0	0	0	0	0	0	0	0
25%	2015	9	2002	0	1	4	4	2	8	146	6	8	2998	17600
50%	2017	23	2006	1	4	4	4	2	9	214	10	8	3510	23240
75%	2018	38	2012	1	5	4	4	4	19	305	11	8	4411	29845
max	2019	51	2020	1	6	5	14	8	35	1099	14	11	14795	441600
	drivetyp_t	segmnt_t	plntctry_t	origin_t	vlvclndr	vlvtotal	incomplt	ve_forms	per_no	hourname	harm_evname	sch_busname	rollovername	agename
count	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670
mean	3	0.9	22.7	1.4	2.1	11.3	0	2	1.6	11.9	30.4	0	0.2	41.6
std	1.3	0.3	8.9	1.8	1.8	10.2	0.2	2	1.2	6.9	11.2	0.1	0.4	23.4
min	0	0	0	0	0	0	0	1	1	0	0	0	0	0
25%	2	1	16	0	0	0	0	1	1	6	27	0	0	22
50%	2	1	29	0	2	16	0	2	1	12	27	0	0	37
75%	4	1	29	3	4	16	0	2	2	18	40	0	0	58
max	5	1	29	4	16	48	1	64	45	24	59	1	1	108
	per_typname	inj_sevname	seat_posname	rest_usename	rest_misname	air_bagname	ejectionname	ej_pathname	drinkingname	drugsname	dstatusname	work_injname	hispanicname	sexname
count	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670	394670
mean	0.7	3.2	8.6	14.2	0	0.6	2.3	0.6	1	0.9	2.5	0.7	5.7	0.6
std	1	2.1	4.8	4.5	0.1	0.5	1.1	1.7	1.2	1	0.8	0.6	0.8	0.5
min	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25%	0	1	6	9	0	0	2	0	0	0	2	0	5	0
50%	0	3	6	17	0	1	2	0	1	1	3	1	6	1
75%	2	5	9	17	0	1	2	0	1	1	3	1	6	1
max	3	10	29	18	1	1	7	11	4	4	5	3	8	4

While examining the dataset, I came across the displcc and displci columns. These columns are defined as displcc: displacement in cubic centimeters. Displci is defined as displacement in cubic inches. When I searched the internet, I found the attached formula and equated the displci column to the displcc column and dropped the displci.



While handling the missing values in the **Data Cleaning and Feature Engineering** part, which is the third part of my research, I first dropped the columns with more than 70000 missing values, these are "vintrim_t", "abs_t", "drl_t", "security_t", "whlblg", "whlbsh", "blocktype", "rstrnt_t", "carbtype", "carbtype_t", "ncicmake","bodystyl". Then I examined how many unique entries each feature has and dropped the features that I thought would not be useful and that would cost me a lot of computational cost while encoding. These are "st_case", "vinmodel_t", "str_vehname", "plntcity", "drug_detname", 'man_collname'

Then I selected the columns whose data type is object and examined their unique values. I saw that case sensitivity is not taken into account when entering data in some features. I made these features uppercase to reduce Cardinality. The features that I made the contents of uppercase: "mfg_t", "bodystyl_t". Thus, the mfg_t column with 82 unique features decreased to 51 unique features, and the bodystyl_t column with 79 unique features decreased to 45 unique features.

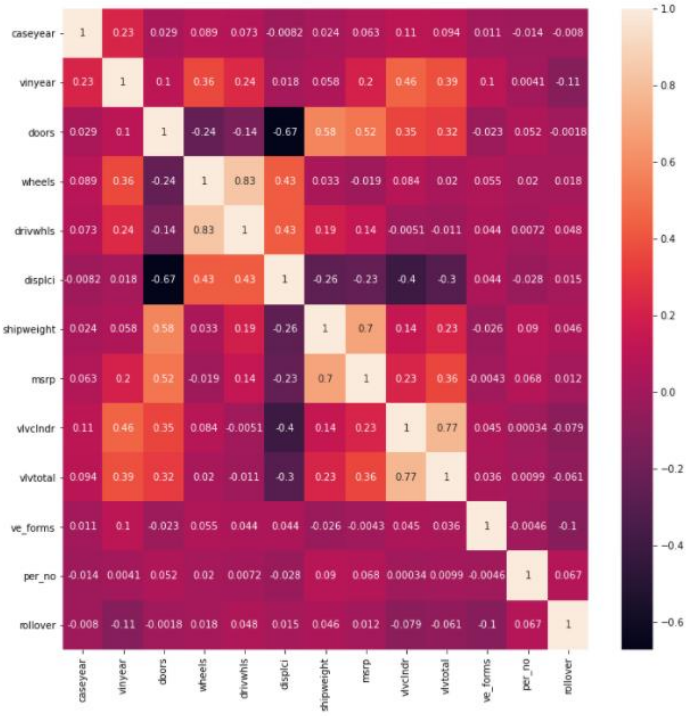
In the **3.2 Reducing cardinality of categorical features** part, I combined the long names of the brands in the mfg_t column with their short names and turned them into 45 unique features by installing 4 more unique features. Then I converted the 4 unique features of our target value, rollovername, into 2 features named rollover and no rollover in order to show whether it is a rollover or not. Likewise, I have reduced the cardinality of the columns written as float and integer in the cylndrs column by converting them to int type. By applying the same application to the segment_t column, I classified the vehicles as luxury, non-luxury, in the air_bagname column the airbag is deployed or not deployed, in the rest_misname column, the vehicle types are 'STANDARD', "TRUCK", 'VAN', 'BUS', 'SPORT', INCOMPLETE. ', 'BIKE', 'OTHER', rest_misnames as YES or NO, I have classified the drivetype_t column as 2X2 and 4X4, which indicates the traction power of the car. I filled the null values of the shipweight column where the vehicle weight is specified with the mean value of the column, 3536.33. Then I classified the hispanicname column, which is an unethical feature of the dataset, whether the driver is of Spanish origin or not. And finally, in the

agename column, I classified 0-30 age range as Young, 31-65 age range as Mid-Aged, and 66 and above as Old

Detailes

3.3. In the Identifying and handling with outliers part, I identified the entries that we can name as outliers in the dataset with boxplot. Then I dropped all the missing values with dropna(). So I got a dataframe with (394670, 43) and no missing values. Then I performed the encode operation to assign numeric values to the categorical features. First of all, I used pandas' get_dummies method and sklearn's OneHotEncoder tool, but they could not encode the categorical features efficiently, even I was getting an out of memory error and it was impossible for me to use them because they were running out of power. When this is the case, I found it appropriate to use the LabelEncoder tool. LabelEncoder categorical features give numeric values from 1 to n and do not increase the number of columns, so calculations do not cost high computational costs. I have fit_transformed my features whose datatype is object, that is categorical, to LabelEncoder. moreover, my dataset of (394670, 43) was preserved.

Part of the statistical analysis method is to calculate the correlation. The correlation examines the dataset as ipair and evaluates the relationship between them between -1 and 0. After encoding our dataset, all our values became numeric. so we can calculate correlation. I have included the results below and we can examine the most related features with the heatmap visualization technique of the seaborn library and take more effective decisions by taking this into account while making the decision.



Positive Corr

rollovername	1.00
rollover	0.62
ejectionname	0.41
ej_pathname	0.26
drinkingname	0.17
drivetyp_t	0.14
vehtype_t	0.11
drugsname	0.11
cylndrs	0.10
per_no	0.10
seat_posname	0.07
per_typname	0.07
shipweight	0.07
drivwhls	0.07
agename	0.06
plntctry_t	0.05

Negative Corr

incomplt	-0.01
sch_busname	-0.01
dstatusname	-0.03
hourname	-0.03
caseyear	-0.04
bodystyl_t	-0.04
mfg_t	-0.05
inj_sevname	-0.06
origin_t	-0.07
hispanicname	-0.09
vlvtotal	-0.09
vlvclndr	-0.12
work_injname	-0.13
ve_forms	-0.15
vinyear	-0.15
rest_usename	-0.24

In Part-4: Data Splitting and Transformation, we will split our data into training and testing datasets and transform data. Since I thought that if we only get the 2019 data from the 2 strategies presented to us in the notebook, I thought that we would not be able to access the older data and experience misclassification, so I found it appropriate to take 30% random data for testing and I specified test_size as 0.3, dropping the rollovername feature for x and only the rollovername feature for y. I found it appropriate to take the . I then placed it in the train_test_split method to generate train and test variables for x and y.

Part-5: I searched the internet for 3 feature selection applications, which are given as automated ways, because it says to use at most 30 useful features in the Feature Selection part. The first result I found was to choose best k variables. For this, I imported the terms SelectKBest and chi2 in the sklearn feature selection library. I determined x and y and selected n_largest 30 in the SelectKBest method and selected 30 features that would maximize the score by affecting the rollovername the most and turned it into an x_topdown list. Then I calculated the importance using Random Forest classifier and created the x_rfc list using the 41 most effective features in this technique. Later, I wanted to make feature selection with RepeatedStratifiedKFold and Decision-Tree based model, but I didn't use them because I couldn't get the result I expected.

Important note: From here on, I went to part-6 and used training and Performance Evaluation techniques, but because I was curious about the score of the features I chose, I put the most effective features that I got from SelectKBest and RandomForestClassifier into the Logistic Regression calculation. I am sharing the results below.

SelectKBest	**RandomForestClassifier**
0.8525012457665053	0.853565425967686

```
x_topdown = ["shipweight","msrp","ej_pathname","agename","ejectionname","rest_usename","vlvtotal",
             "displci","drinkingname","ve_forms","vlvclndr","mfg_t","seat_posname","drugsname","origin_t",
             "drivetyp_t","plntctry_t","per_no","bodystyl_t","work_injname", "cylndrs","per_typname","inj_sevname",
             "vehtype_t","hourname","drivwhls","harm_evname","statename","wheels","vinyear"]
```

```
x_rfc = ['agename','air_bagname','bodystyl_t','caseyear','cylndrs','displci','doors','drinkingname','drivetyp_t','drivwhls',
         'drugsname','dstatusname','ej_pathname','ejectionname','fuel_t','harm_evname','hispanicname','hourname','incomplt',
         'inj_sevname','mfg_t','msrp','origin_t','per_no','per_typname','plntctry_t','rest_misname','rest_usename','sch_busname'
         'seat_posname','segmnt_t','sexname','shipweight','statename','ve_forms','vehtype_t','vinyear','vlvclndr','vlvtotal',
         'wheels','work_injname']
```

Logistic Regression

we first create a Logistic regression model, after we fit_transform and transform our x_train and x_test variables in StandardScaler, 4 sets of dimensions (276269, 41) (118401, 41) (276269, 1) (118401, 1) I created. I created the LogisticRegression model and fit my variables. I used train and test variables to predict the results, while I used the predict_proba method to predict probability, then I printed the model's score, classification error and confusion matrix.

Decision Tree

In order to apply the Decision Tree algorithm, I selected max depth 4 according to the DecisionTreeClassifier import entropy criterion, assigned my x and y train and test variables as parameters, and then fit and predicted. visually adding score and confusion matrix

Random Forests

While applying Random Forest Classifier to our dataset, we first scale the train and test variables of x and y to use a standard range from StandardScaler. Then we create our model in entropy criteria and with 100 estimators, we fit the train data sets to this model. We print the model’s score, classification report and display the results in confusion matrix.

SVM

When I ran my support vector machine calculation on the notebook where I did the homework, it took a lot of time and it didn't work, so I exported the dataset I used for the train test spilt as FarsSVM.csv and ran it on a different notebook. I'm putting the results of my SVC() model, which gives results after about 3 hours, as a screenshot. The reason why we use the kernel linearly when using SVC is that it doesn't do it quadratic while calculating and it costs less computational cost.

Naïve Bayes;

GaussianNB, ComplementNB, CategoricalNB

Naive Bayes is a machine learning term that calculates the probability of a conditional probability occurring using bayes' theorem. In this research, I calculated Gaussian, Complement and categorical naive bayes and compared their results. Unlike other techniques, I used MinMaxScaler instead of StandardScaler when calculating naive bayes, so the model is tuned by obtaining data points with a value between 0-1.

Results

In Part-6: Training and Performance Evaluation

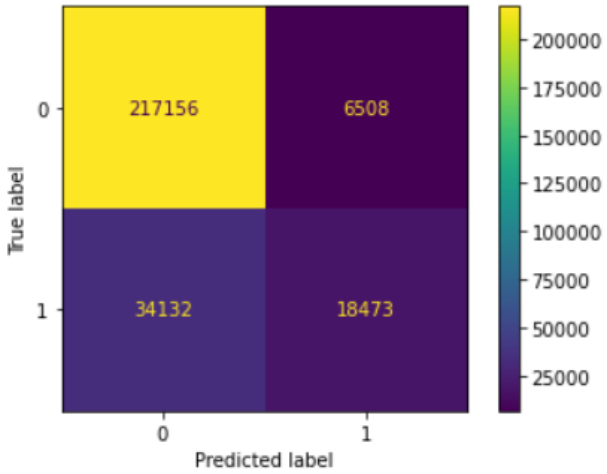
Logistic Regression

```
Logistic Regression score: 0.8515384160606752
```

```
Accuracy 0.8515384160606752
Precision 0.8515384160606752
Recall 0.8515384160606752
F1 0.8515384160606752
```

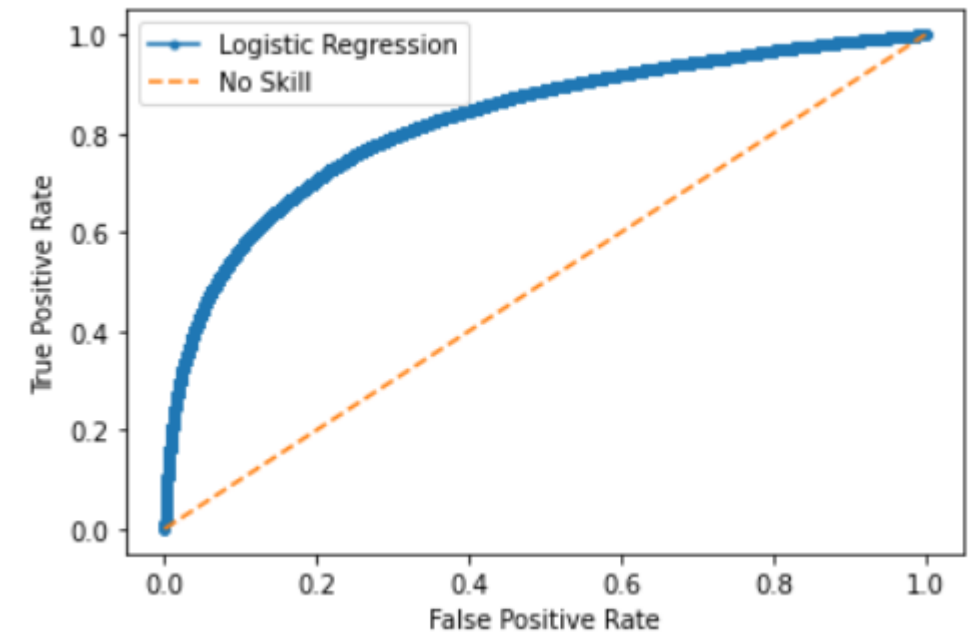
Train Classification Report

	precision	recall	f1-score	support
0	0.86	0.97	0.91	223664
1	0.74	0.35	0.48	52605
accuracy			0.85	276269
macro avg	0.80	0.66	0.70	276269
weighted avg	0.84	0.85	0.83	276269



Test Classification Report

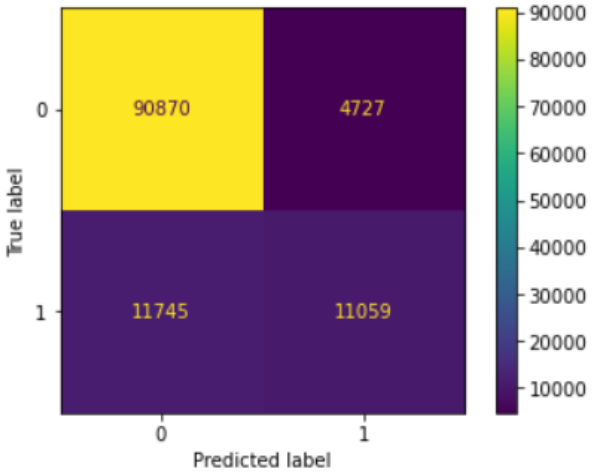
	precision	recall	f1-score	support
0	0.86	0.97	0.91	95597
1	0.74	0.35	0.48	22804
accuracy			0.85	118401
macro avg	0.80	0.66	0.70	118401
weighted avg	0.84	0.85	0.83	118401

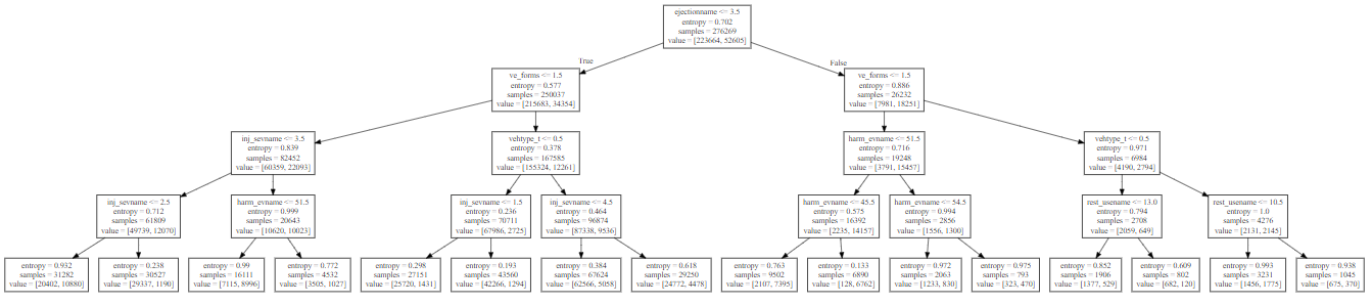


Decision Tree

```
0.8608795533821505
```

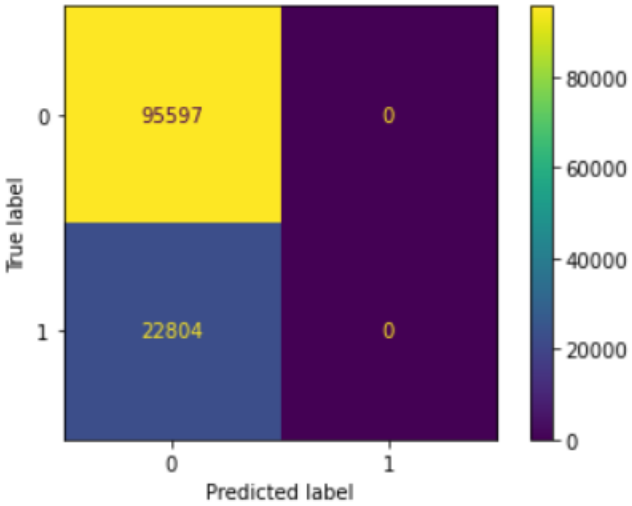
Important Note: I export the .dot file, added with hw zip.





Random Forest

```
Random Forests score: 0.8074002753355124
```



	precision	recall	f1-score	support
0	0.81	1.00	0.89	95597
1	0.00	0.00	0.00	22804
accuracy			0.81	118401
macro avg	0.40	0.50	0.45	118401
weighted avg	0.65	0.81	0.72	118401

Important Note: I export the .dot file, added with hw zip.

Support Vector Machine,

```
data = pd.read_csv("farsSVM.csv")
```

```
data.dropna(inplace=True)
```

```
x = data.drop("rollovername", axis=1, inplace=False) #related features
y = data["rollovername"] #target
```

```
x_train_svm, x_test_svm, y_train_svm, y_test_svm = train_test_split(x, y, test_size=0.3)
```

```
scale_svm = StandardScaler()


x_train_svm_std = scale_svm.fit_transform(x_train_svm)
x_test_svm_std=scale_svm.transform(x_test_svm)


svm_model = SVC(kernel="linear", C=0.1)

svm_model.fit(x_train_svm_std, y_train_svm)
```

```
SVC(C=0.1, kernel='linear')
```

pred_y_train = svm_model.predict(x_train_svm_std) pred_y_test = svm_model.predict(x_test_svm_std)

```
svm_model.score(x_test_svm, y_test_svm)
```

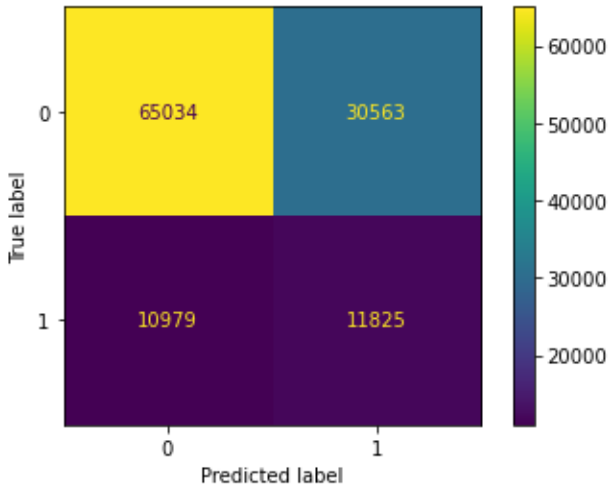
```
0.6491414768456347
```

```
cm=confusion_matrix(y_test_svm, svm_model.predict(x_test_svm))
```

```
disp=ConfusionMatrixDisplay(confusion_matrix=cm)
```

```
disp.plot()
```

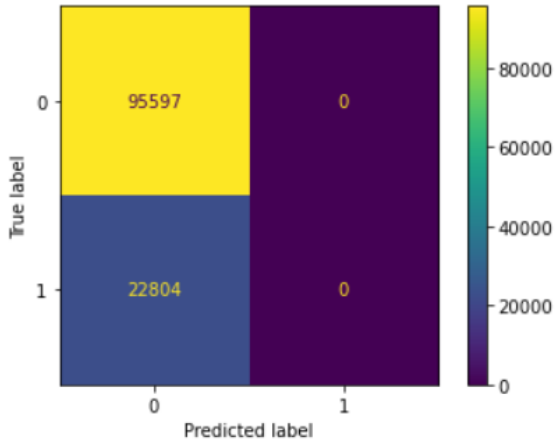
```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x207af539370
>
```



Naïve Bayes;

GaussianNB

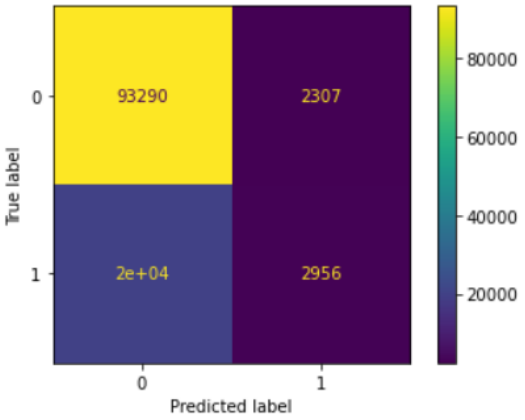
```
GaussianNB score: 0.8074002753355124
```



	precision	recall	f1-score	support
0	0.81	1.00	0.89	223664
1	0.00	0.00	0.00	52605
accuracy			0.81	276269
macro avg	0.40	0.50	0.45	276269
weighted avg	0.66	0.81	0.72	276269

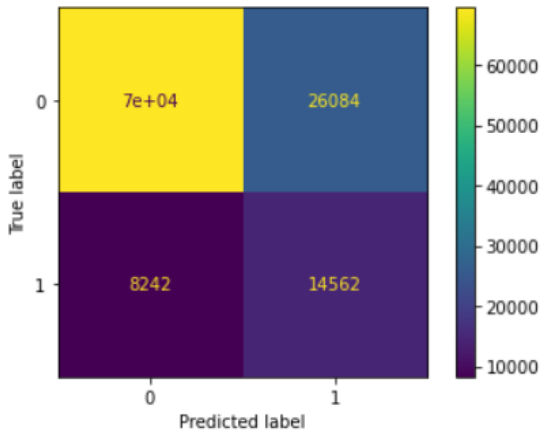
Categorical NB

```
CategoricalNB score: 0.8128816479590544
```



Complement NB

```
ComplementNB score: 0.710086908049763
```



Conclusion

After the preprocessing, manipulation, visualization and training techniques we have done in our research, we will have graphics, numeric values and the insights we will extract. It was our duty to make sense of them and help the decision-making of the insurance company where we did our internship. In summary, I made a few inferences as a result of the techniques we felt for the FARS dataset. First of all, insurance companies ensure that a possible damage is covered financially by selling insurance to vehicles, customers pay a certain fee to insurance companies to insure their vehicles. Among these dynamics, they implement the price policy that insurance companies will apply to their customers by evaluating certain factors. In my research, the meanings we can derive from the numeric and categorical features of the data we have;

- I would like to convey to the board of directors that the number of accidents has been increasing since 2014 and that our insurance company may increase its prices.
- We can observe that vehicles produced between 1980-2004 are involved in fewer accidents than vehicles produced in 2004 and later. For this reason, we can avoid causing financial loss to our company in case of possible accidents by paying more insurance fees than vehicles produced after 2004.
- Since vehicles with 4 doors are more likely to have an accident than vehicles with 0, 2, 3 doors, a higher price may be charged when pricing. Likewise, vehicles with 4 wheels are more likely to be involved in accidents than vehicles with other wheels. A cheaper pricing can be applied to vehicles such as motorcycles.
- Looking at the dispatcher feature, pricing can be increased if their displacement is greater than a certain distance.
- In the shipweight feature, where the vehicle weight is examined, we observe that vehicles weighing more than 3000 kg and lighter than 5000 kg are involved in more accidents. If the weight of the customers' vehicles is measured and it is within this range, our company will gain by increasing the pricing.
- Significantly higher crash rates were found for vehicles less than $1250. For this reason, the market values of the vehicles are researched and the insurance policies of the vehicles in this range are made more expensive, and our company is profitable.
- At the same time, when we look at the missing values, there may be points that we should pay attention to because we drop them even though we drop them in our calculations. For example, the vintrm_t feature specifies the Trim of the vehicle and its type. Likewise, the securtiy_t feature specifies the security systems of the vehicles. Since these data need too many missing values, we should not make inferences in these fields, otherwise it will be misclassification.

Among the categorical features of our research, we can calculate the percentages of all the data and infer from the accident numbers, especially according to the state, in the part where we create the bar plot. For example, we observe that most of the accidents are made in Texas. Therefore, Texas dealers of our insurance company may apply a more expensive pricing compared to other states when insuring vehicles. California and Florida follow this order. Columbia and Rhode island have the lowest accident rate, and pricing can be reduced in these states and customers can be gained. Another point I would like to draw attention to here is that, as a marketing strategy, our profit can be maximized by attracting more customers by making a discount as a campaign in states with average accident numbers from the states.

- I mentioned the pricing policy difference according to the vehicle type and body style from our Categorical Features. If we need to elaborate on this issue, since Truck type vehicles are involved in more accidents, if our customer's vehicle is a truck, higher pricing can be made. In the body style feature, the pricing of the vehicles included in the Standard vehicle group must be increased. This comparison includes pricing for normal vehicles such as sedans or hatches, unlike the comparison between Passenger vehicles and trucks. Trucks were compared to passenger buses. In addition, we can deduce that sports vehicles are also involved in more accidents.

- If we have to make an inference on the brands, General motors, ford and daimler-chrysel group cars were involved in more accidents than other vehicles. For this reason, the problems that may arise in the mechanical accents of these vehicles are more common than in others. The insurance fees of these vehicles should be increased.
- If we need to change the pricing policy according to the number of cylinders, the order 6 > 4 > 8 > 5 can be followed. We must sell the most expensive automobile insurance to 6-cylinder motor vehicles so that our company does not suffer financial loss in case of an accident.
- If we make an inference according to the fuel type, gas-using vehicles have more accidents than diesels. It has been observed that vehicles working with electricity or hybrids have fewer accidents. If our customer's vehicle uses gas fuel, the insurance fee we will charge him should be higher.
- It has been observed that front wheel drive vehicles are involved in more accidents, but rear wheel drive vehicles are also more than 4X4s, so pricing policy should be optimized accordingly.
- Since the research was conducted in the USA, USA origin vehicles were involved in more accidents. If a vehicle is not from the USA, a discount can be applied, especially from origins with a low percentage in plntctry_t feature. For example, German vehicles were involved in fewer accidents than American vehicles.
- We have a very serious comparison, in which we see that incomplete vehicles are involved in many more accidents.
- The number of rollovers in accidents is lower than the number of absences, so the vehicle and personnel are not harmed much.
- In addition, according to agename feature, 0-30 year old are involved in more accidents, followed by 31-65 years old, so younger and mid-aged persons may be priced more.
- Likewise, if our customer is a man, he tends to have more accidents than women, so a different price policy can be prepared for men. but this is unethical
- Vehicles with a high ejected feature are more likely to have an accident, so by checking whether their vehicles are ejected, an increased insurance can be applied to those who are ejected.

Apart from these inferences, I had to drop the during and after crash data when it was necessary, but I did not drop it and included it in my calculations. Because, as an insurance company, we can use this data in our advertisements.

For example, we can advertise that men are involved in more accidents than women, and increase awareness, so that male drivers have more insurance. In the same way, we can make a profit by placing advertisements by stating that the age of 0-30 and 30-65 are involved in too many accidents, raising awareness and even encouraging inexperienced drivers to make compulsory insurance with the support of the state.

17-18, 18-19, 16-17 hours were the 3 hours with the most accidents in the data of the time interval of the accident, which is the most important data for me. These hours coincide with the end of the working hours, and by promoting this data, the number of accidents can be reduced by ensuring that more attention is paid to the traffic rules at those hours. However, as an insurance company, we must make decisions in the interest of our company with these ratios.

Likewise, there have been more accidents in July because the weather is hot, and it can be concluded that there may have been more geese due to the rainy weather in October. From here, accidents happen a lot at these ages, it can be done with an insurance policy. Thanks to the campaigns, we can generate profit by attracting customers.

Confusion Matrix

True negative predictions	**False negative predictions**
False positive prediction	**True positive predictions**

While examining the Classification matrix, the following parameters are taken into account;

Recall is "how many of this class you find over the whole number of element of this class"

The precision is "how many are correctly classified among that class"

The f1-score is the harmonic mean between precision & recall

The support is the number of occurrence of the given class in your dataset

While interpreting the confusion matrix, the 00 point, the upper left, and the 11 point, the lower right, shows the predicted results. The other two squares, site type-1 and type-2, show the number of predicts that failed.

When we examine the results, the Logistic regression score is 85.2% and when the classification matrix is examined according to the features we have chosen, we observe that the average score is 85% again. We observe that the risk of rolling over is 35% and the risk of not being 97%. Moreover, 217 thousand and 18 thousand correct predictions were made in the Confusion matrix.

When the graph is examined, we see that the slope of the roc_auc_curve is taken and the features we have chosen are 85% correct.

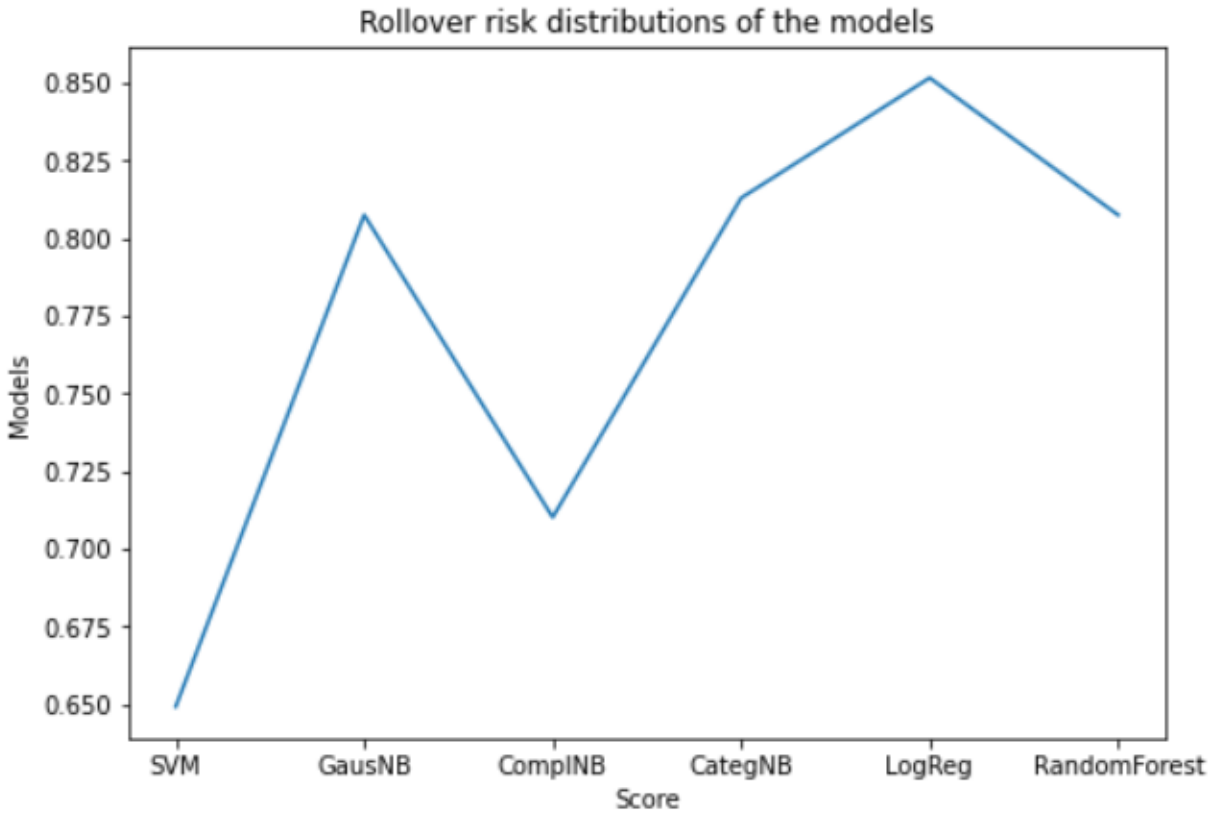
While performing the decision tree classification, we can obtain a score of 86%, and we can examine our graph with the graphviz tool and draw attention to the effective features.

When we use the Random forest Classifier, unlike the decision tree, our score is 80% because we touch more than one feat. However, there are 95 thousand labels with 00 predicted. Unfortunately, 22.8 thousand data were predicted incorrectly.

Since our support vector machine application did not work on my own notebook, I said that I was calculating from a different notebook. You can find this notebook in my homework file. My Confusion matrix score is 64%, and the large number of incorrectly predicted labels indicates that we failed in this method.

Since I have examined the naive bayes method in 3 other sub methods and I know that most of our features are categorical, I predict that CategoricalNB will make a more successful classification with a higher score before I start my calculations. Gaussian NB scored 80%, showing the benefits of numeric features to the rollover predict risk. The point I want to draw attention to here is 95 thousand labels are correct

While predicted, there are 22 thousand errors, but False Positive and True Positive values are 0. Categorical NB is the most successful Naive Bayes technique that calculates the probability of being predicted correctly with a score of 81%, as I predicted.



When we compare the scores in the Chart above, Logistic Regression made the best classification, then Categorical NB and Random Forest made the best classification. It is SVM that makes the worst classification. In this context, Logistic regression can be selected to predict and tune the model and the mode can be tuned until the score is optimized.

As the last part of our research, we will remove these horns and tune our model in Part-7: Interpretation. As I mentioned in the previous sections, the features I selected with the SelectKBest feature method while performing feature selection showed that they were the 30 most useful features with a score of 85.3% in Logistic regression. Moreover, 41 features that I selected with RandomForestClassifier and put into Logistics regression achieved the most successful classification with a score of 85.4%.

x_1 = ["ejectionname","ej_pathname","vinyear","rest_usename"]

x_2 = ["ejectionname","ej_pathname","drinkingname","drivetyp_t"]

x_3 = ["statename","vinyear","vehtype_t","bodystyl_t","mfg_t","cylndrs","fuel_t","shipweight","msrp","drivetyp_t","segmnt_t","plntctry_t","origin_t","vlvclndr","vlvtotal","incomplt","ve_forms","per_no","hourname","harm_evname","agename","sexname","per_typname","rest_usename","rest_misname","air_bagname","ejectionname","ej_pathname"]

After all the calculations, when I decided what a successful classification method logistic regression was, I created the x_1 and x_2 lists purely intuitively (by choosing the features with the highest and what low correlation values). The score of x_1'list is 0.848717493940085..

The score of x_2 list is 0.8479658110995684. Then I created the x_3 list when I only heuristically selected the features that fit my logic, and this models score came in 0.8874671666624437, more than all previous metods.

According to the researches and calculations I have made, the features that cause the accidents have been examined. It has been decided which machine learning method is the most effective while estimating the probability and classification of these accidents. Awareness has been created that we should pay attention to driving carefully, not according to statistical scores. It was one of the most productive internships I've had, thank you to everyone who contributed.

Appendix:

FARS dataset as csv file

https://drive.google.com/file/d/1tShn1Woepxwxku_4Y9Fk42oeZaygh0YQ/view?usp=sharing