CS452/CS552 Data Science with Python / Fall'21 Homework – 2

Car Rollover Prediction

Created by Aylin Aydın - S018183

Abstract

This report provides information about ML algorithms and data set that can predict traffic accidents, developed using the Fatality Analysis Reporting System (FARS) data provided by the National Highway Traffic Safety Administration (NHTSA) at the United States Department of Transportation. It has been noted that the outputs of the algorithms are convenient to use in automated bidding mechanisms placed on the websites of car insurance companies or to provide useful data features that can be asked during the customer registration process. It was chosen to use two data sets from the FARS data collections linked to people and cars involved in accidents. Personal information and vehicle parameters were analyzed, and an integrated data set was generated by examining if they might be accessed before the accident happened. Support Vector Machine (SVM), 3 different Naive Bayes algorithms, Logistic Regression and finally Random Forest algorithms for training the data set. At the end of this entire evaluation process, the Random Forest model was the model that achieved the best score when the scores of the trained models were compared.

Introduction

Data science is being used to learn about behaviors and processes, to design algorithms that process massive volumes of data rapidly and effectively, to improve the security and confidentiality of sensitive data, and to lead data-driven decision-making. In this research, I choose to use the FARS data sets community, which gathers data on fatal motor vehicle accidents that happened in the United States between 1975 and 2019 and makes them publicly available through an API, using data science's potential to guide data-based decision making. With the use of this dataset, it was determined if the car would be substantially damaged or damaged in the event of an accident by comparing the vehicle's rollover value in the event of an accident with unexpected data gathered from insurance company users during registration. As a result, optimum pricing for customers who wish to insure themselves or their vehicles based on data collected at the time of registration may be determined.

Support Vector Machine, Naive-Bayes, Logistic Regression, and Random Forest models are trained in this study to predict car rollovers. "Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well." **[1]**

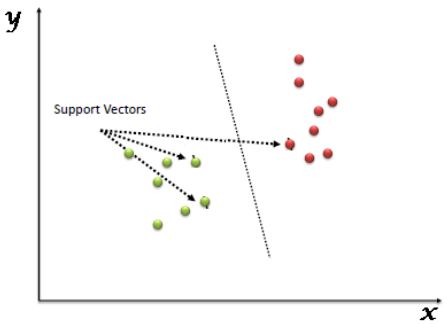


Figure 1. Support Vector Machine

Support Vectors are simply the coordinates of individual observation. The SVM classifier is a frontier that best segregates the two classes (hyper-plane/ line)

"It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a

particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:" **[2]**

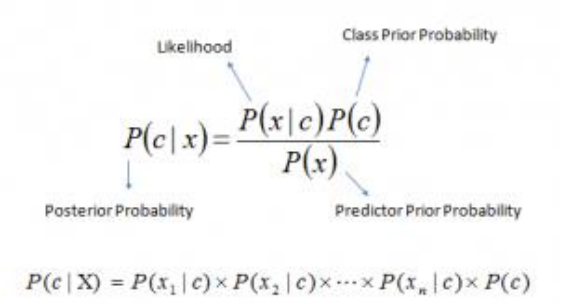


Figure 2. Naïve Bayes Algorithm Description

"Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. We can call a Logistic Regression a Linear Regression model, but the Logistic Regression uses a more complex cost function, this cost function can be defined as the 'Sigmoid function' or also known as the 'logistic function' instead of a linear function. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression." **[3]**

$$0 \leq h_\theta(x) \leq 1$$

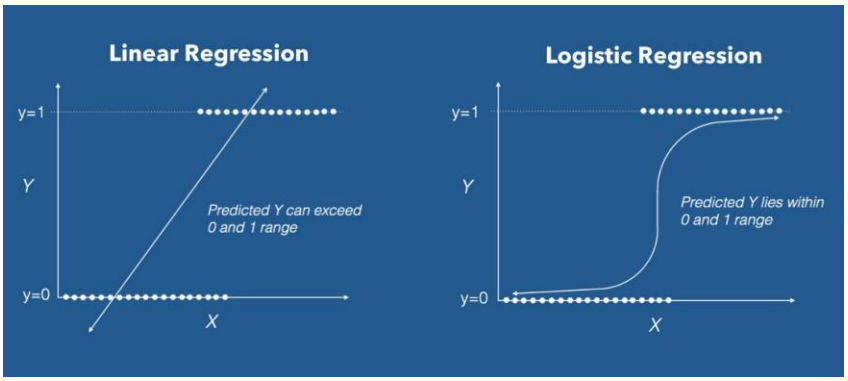


Figure 3. Logistic Regression Description

"Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems." **[4]**

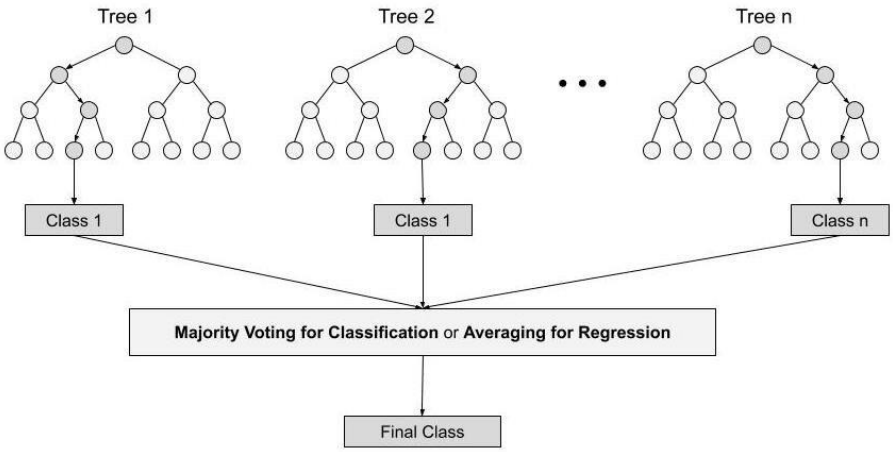


Figure 4. Random Forest Description

Methodology

This study was prepared using numpy, pandas, matplotlib, seaborn and sklearn from python's existing libraries by using FARS's datasets in python software language and performing operations via Jupyter Notebook.

The work was carried out in 7 sections. First, it was necessary to remove the data set from the API and install it and properly combine the 2 data sets that we use. Since we had a large number of columns in both data sets before this merge, the columns were eliminated in accordance with our goal. In addition, this section has also been written to a separate notebook so that the resulting unified data set does not work repeatedly every time the Jupyter Notebook is started, and the resulting data is also written. it is saved to the location as a .csv file. In the second section, Exploratory Data Analysis (EDA) was started to better understand the data and in this section, data recognition with some graphs and statistical values and data processing was started when necessary. In the third part, empty values and unknown values of columns with empty values were dealt with in order to avoid difficulties when training models in the later stages. At the same time, for the same reason, the cardinality of the columns that have too many different values was reduced by grouping the values inside. In the fourth section, the data is encoded and parsed under appropriate conditions as a test and train, since it is now organized in such a way that it can be trained with a model. By the fifth part, important features were identified using a decision tree-based model, and models using ML models mentioned in introduction were created in the sixth part of notebook. Finally, statistical and numerical interpretation of the data obtained in the seventh chapter was carried out.

Implementation

Part 1: Data Loading and Cleaning

First of all, the data sets containing the people and vehicles involved in the accident were pulled from the API year by year, years between 2014 and 2019. The data frames were examined, the necessary arrangements and deductions were made, and then analyzed and prepared to be ready for training. Two different data frames, person, and vehicle were merge.

In [8]: vehicles

Out[8]:

	caseyear	state	statename	st_case	veh_no	nclcmake	vinyear	vehtype	vehtype_t	vinmake_t	vinmodel_t	vintrim_t	vintrim1_t	vintrim2_t	vi
0	2014	1	Alabama	10001	1	TOYT	2011.0	P	Passenger Car	TOYOTA	COROLLA	BASE	S	LE	
1	2014	1	Alabama	10002	1	DODG	1997.0	T	Truck	DODGE	RAM 2500	NaN	NaN	NaN	
2	2014	1	Alabama	10003	1	CHEV	2004.0	P	Passenger Car	CHEVROLET	MALIBU	LT	NaN	NaN	
3	2014	1	Alabama	10003	2	TOYT	1997.0	P	Passenger Car	TOYOTA	CAMRY	CE	LE	XLE	
4	2014	1	Alabama	10004	1	TOYT	1999.0	T	Truck	TOYOTA	TACOMA	NaN	NaN	NaN	

In [9]: persons

Out[9]:

	caseyear	state	statename	st_case	ve_forms	veh_no	per_no	str_veh	str_vehname	county	countyname	day	month	monthname	hour	hournar
0	2014	1	Alabama	10001	1	1	1	0	Occupant of a Motor Vehicle	71	JACKSON (71)	1	1	January	1	1:00a 1:59
1	2014	1	Alabama	10001	1	1	2	0	Occupant of a Motor Vehicle	71	JACKSON (71)	1	1	January	1	1:00a 1:59
2	2014	1	Alabama	10002	1	1	1	0	Occupant of a Motor Vehicle	59	FRANKLIN (59)	1	1	January	13	1:00p 1:59
3	2014	1	Alabama	10003	2	1	1	0	Occupant of a Motor Vehicle	125	TUSCALOOSA (125)	1	1	January	3	3:00a 3:59
4	2014	1	Alabama	10003	2	1	2	0	Occupant of a Motor Vehicle	125	TUSCALOOSA (125)	1	1	January	3	3:00a 3:59

In [30]: accidents

Out[30]:

	caseyear	statename	st_case	veh_no	vinyear	vehtype_t	vinmake_t	vinmodel_t	bodystyl_t	doors	wheels	drivwhls	mfg_t	displci	cylndrs
0	2014	Alabama	10001	1	2011.0	Passenger Car	TOYOTA	COROLLA	Sedan	4.0	4.0	2.0	TOYOTA	110.0	4.0
1	2014	Alabama	10001	1	2011.0	Passenger Car	TOYOTA	COROLLA	Sedan	4.0	4.0	2.0	TOYOTA	110.0	4.0
2	2014	Alabama	10002	1	1997.0	Truck	DODGE	RAM 2500	Pickup	2.0	4.0	2.0	DAIMLER-CHRYSLER	360.0	8.0
3	2014	Alabama	10003	1	2004.0	Passenger Car	CHEVROLET	MALIBU	Sedan	4.0	0.0	0.0	GENERAL MOTORS	214.0	6.0
4	2014	Alabama	10003	1	2004.0	Passenger Car	CHEVROLET	MALIBU	Sedan	4.0	0.0	0.0	GENERAL MOTORS	214.0	6.0

Before merging, the data frames were examined to obtain information about the data contained in the data frames. In order to avoid data loss, copies of data frames were created, and the operations performed proceeded through these copies. The names of the columns in the two data sets were extracted and the FARS User's Manual was used to understand the characteristics of the data they contained. For example, it became clear that the columns *"dispci"* and *"dispcc"* have the same content, but they contain two different metrics, so they increase the number of columns. For this reason, the data were collected in a single column by converting *"displcc"* to *"displci"* and the *"displcc"* column was removed. Columns with a lot of empty data were removed. Then, with the help of FARS User's Manual, we removed the information available during or after the accident, such as the place of the accident, the weather, which direction the car fell, from our data frames to help the insurance company.

Following the most recent merge, the data set was gathered by continually obtaining data from the API to prevent wasting time. The address was stored as a .csv file. The columns in this data set are listed in alphabetical order below. Some of these columns were removed because they were considered to be useless in the later stages or may be related to the moment or after the accident.

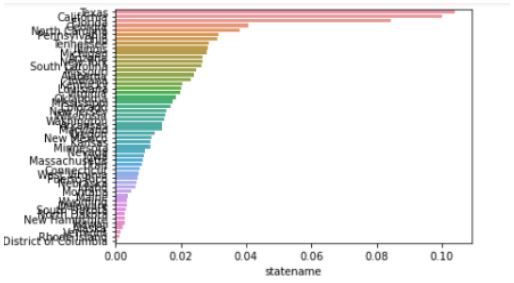
```
['age','body_typname','bodystyl_t','carbtype','caseyear','countyname','cyln
drs','day','displci','doors','drinkingname','drivetyp_t','drivwhls','drl_t'
,'drugsname','engvincd','fuel_t','hispanicname','hourname','incomplt','make
name','mfg_t','mod_yearname','monthname','msrp','origin_t','p_sf1name','p_s
f2name','p_sf3name','per_no','per_typ','plntcity','plntctry_t','rest_usenam
e','rollovername','sch_busname','seat_posname','segmnt_t','sexname','shipwe
ight','spec_usename','st_case','statename','str_veh','str_vehname','tow_veh
name','ve_forms','veh_no','vehtype_t','vinmake_t','vinmodel_t','vinyear','v
lvclndr','vlvtotal','wheels']
```

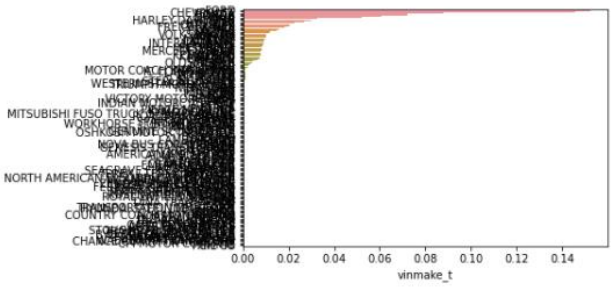
Part 2: Exploratory Data Analysis (EDA)

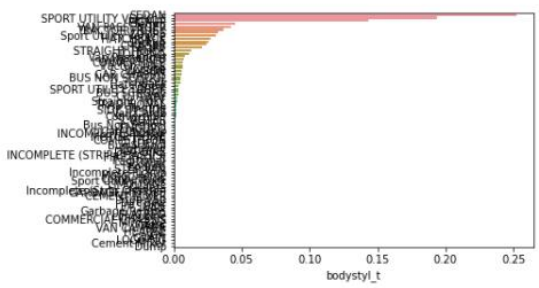
In this section, categorical and numerical characteristics were asked to be investigated. The highlights of these features will also be discussed in detail in this section. An attempt was made to obtain information from the data for the department of property engineering. Firstly, the variable we discussed was examined, and then the analysis phase was started. These reviews were usually conducted in the light of the FARS User's Manual. As a result of these analyzes, duplicate and empty rows and columns were calculated. Histograms, distribution, and bar graphs were created. Unique value counts of categorical characteristics were performed.

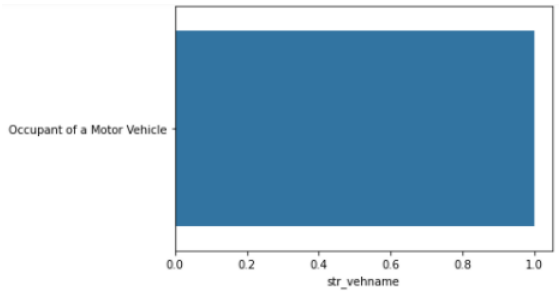
```
In [7]: data.columns.to_series().groupby(data.dtypes).groups

Out[7]: {int64: ['caseyear', 'st_case', 'veh_no', 've_forms', 'per_no', 'str_veh', 'day', 'age', 'per_typ'],
         float64: ['vinyear', 'doors', 'wheels', 'drivwhls', 'displci', 'shipweight', 'msrp', 'vlvclndr', 'vlv
         total'], object: ['statename', 'vehtype_t', 'vinmake_t', 'vinmodel_t', 'bodystyl_t', 'mfg_t', 'cylndr
         s', 'fuel_t', 'carbtype', 'drivetyp_t', 'drl_t', 'segmnt_t', 'plntctry_t', 'plntcity', 'origin_t', 'e
         ngvincd', 'incomplt', 'str_vehname', 'countyname', 'monthname', 'hourname', 'sch_busname', 'makenam
         e', 'body_typname', 'mod_yearname', 'tow_vehname', 'spec_usename', 'rollovername', 'sexname', 'seat_p
         osname', 'rest_usename', 'drinkingname', 'drugsname', 'p_sf1name', 'p_sf2name', 'p_sf3name', 'hispani
         cname']}
```

In order to observe the frequency at which categorical characteristics are determined in which characteristics, a bar plot graph was drawn. As a result of these graphs, it was observed that some properties have too many categories, while others have a single category, and they were stopped being used or noted in order to overcome them in the next section.



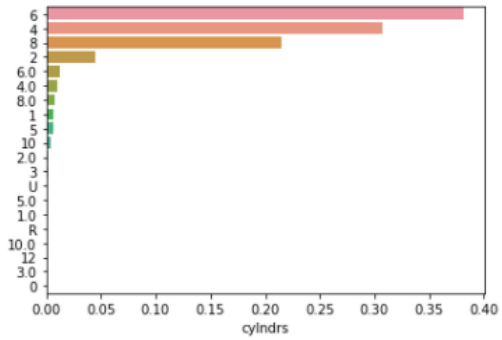


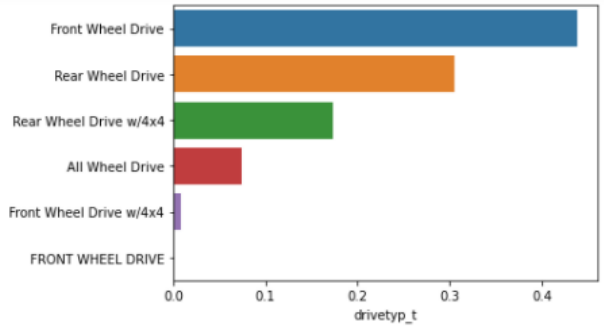


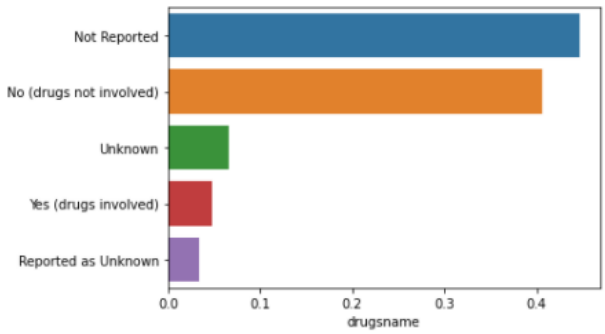


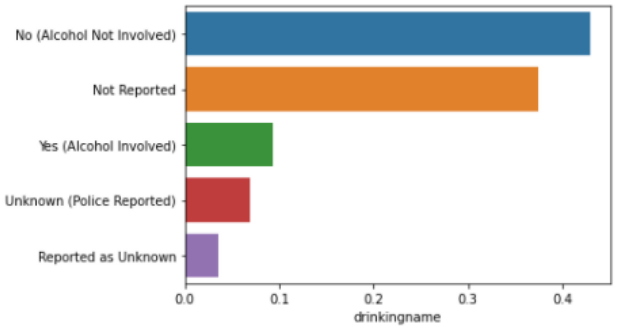
As can be seen from the examples above, some categorical qualities contain just one feature, while others have several features, but their distribution is not uniform. As a result, features with just one feature for the required characteristics will be eliminated from the data collection, but features with too many features will be classified or encoded and utilized. For example, the above-mentioned "statename" illustrates the state where the accident occurred, at first glance, the event of an accident or post-related data looks like, despite the fact that the insurance company will use the car in the applicant's respective provinces, considering that the rollover prediction is actually quite an important feature to consider. The "vinmake t" and "bodystyl t" characteristics, like "statename," have a variety of categories. These sections offer information about car brands and vehicle body styles. A better distribution in the defined categories can be seen by aggregating this information with tiny actions.

Unlike the categories in the other graphs studied, the "str_vehname" category has only one value, and this value is observed in the entire dataset, so it is excluded from the dataset because examining the relationship of this value, which is valid for all data, with the target column will be meaningless.









In comparison to the columns indicated in the preceding paragraph, the distribution of the traits shown above is significantly more regular. However, it still contains terms that make comparisons impossible, such as "Unknown" and "Reported as Unknown." At the same time, even when there are repeated phrases, such as "cylndrs," we meet expressions that are viewed as having distinct meanings due to their similar format. To avoid such circumstances, the expressions included in the categories will be worked on in the next section in order for them to be in a single format and not be ambiguous terms. Despite the fact that, "drugsname" and "drinkingname" are aspects that appear during an accident, it is considered that there is information that may be utilized if insurance firms inquire clients about their usage patterns during the registration stage. However, they will not be used in education considering that they are unethical in the later stages.

The investigations have been continued and the unique values of the columns, the data types and the information they contain have been listed for use in the following sections.

Part 3: Data Cleaning and Feature Engineering

To begin, this part is divided into two stages. To address the missing data and unknowns indicated in the previous chapter, we set out first, and then again, as mentioned in the previous part, to raise the yield of cardinality decrease columns with too many categories. When we printed the number of empty data in the data set according to each column, it was observed that 15705 pieces of data were empty in common with many columns. The work was started by removing these blank data from the data set. After that, when we printed the sum of the empty data in descending order according to the columns again, the operations were continued with the column names listed. Unknowns, unreported data were removed from the dataset without causing too much data loss considering the number, or the most repetitive feature in the dataset was assigned to these data. Subsequently, within the scope of attempts to increase the efficiency of the categories of categorical columns, the categories were grouped in related

ways and their number was reduced. Finally, the unique number of categories of each column in our data set, including numeric columns, has become as follows.

```
caseyear 6                  fuel_t 7                    hourname 24
statename 52                shipweight 3825             sch_busname 2
veh_no 61                   msrp 9625                   tow_vehname 2
vinyear 36                  drivetyp_t 2                spec_usename 7
vehtype_t 2                 drl_t 3                     rollovername 2
bodystyl_t 7                segmnt_t 2                  age 94
doors 5                     plntctry_t 30               sexname 2
wheels 3                    origin_t 4                  rest_usename 4
drivwhls 3                  vlvclndr 9                  drinkingname 2
mfg_t 23                    vlvtotal 16                 drugsname 2
displci 202                 ve_forms 28                 hispanicname 2
cylndrs 9                   monthname 12
```

Part 4: Data Splitting and Transformation

In this section, now that our data set has reached a trainable level, it has been divided into train and test stages. Instead of using data from a specific year of our dataset as test data, it was thought that it was more correct to use thirty percent of the entire dataset as a test dataset. Because the number of vehicles in traffic in a specific year, extraordinary situations that may occur specific to that year were not asked to affect the test and train data set. Then, in order to process categorical data, they must be encoded. At this point, we have the option of using two different encode methods that are appropriate for our situation. One of them is LabelEncoder and the other is OneHotEncoder. The difference between the uses of this decoder is as follows:

We apply One-Hot Encoding when:

1. The categorical feature is **not ordinal** (like the countries above)
2. The number of categorical features is less so one-hot encoding can be effectively applied

We apply Label Encoding when:

1. The categorical feature is **ordinal** (like Jr. kg, Sr. kg, Primary school, high school)
2. The number of categories is quite large as one-hot encoding can lead to high memory consumption **[5]**

But in our case, OneHot complies with the first article of the encoder, while the Label complies with the second article of the encoder. Although it is believed that OneHot encoder is more effective because our categorical columns are ordinal, their number is very large and it takes about 2 hours even with label encoder when training our models, Label encoder was used to run the models and get results. For this reason, it is believed that the results obtained in the results section are curable.

Part 5: Feature Selection

When it comes to picking a feature, we have a lot of categorical data, therefore one of the automated methods was picked with the assumption that it would be difficult to compute their connection and statistical data with the target data. Since it was believed that the best of them would also be a decision-tree based algorithm, it was decided to use RandomForestClassifer. Through RandomForestClassifer, the relationships of the properties found in our dataset with the target value were examined and sorted from the most relevant to the irrelevant.

RandomForestClassifer creates various decision-trees, simply taking all the properties. The importance of this decision-tree-based feature calculates the determination of how well each feature is in determining the final result. For instance, whether a person is overweight, smokes, frequently suffers from chest trouble, and so on. If you are attempting to forecast the chance of having a heart attack based on a variety of factors such as. The most essential aspect will be the ability to tell patients who have experienced a heart attack apart from others. Thus, the method evaluates the Gini impurity or Entropy of the data divisions caused by the characteristics to determine how each of these variables may best separate the data. As a result, if the overweight variable can split the data into two groups, with the highest heart attack rate in one and the lowest in the other, then the characteristic significance of this variable will be the highest. **[6]** When we sort the importance of the properties, we have determined with RandomForestClassifer, we get the following order:

```
'age': 0.088315,                        'sexname': 0.01396821,
've_forms': 0.082471,                   'vlvclndr': 0.0127431785,
'msrp': 0.0799266,                      'hispanicname': 0.010207,
'statename': 0.07703,                   'cylndrs': 0.0099585,
'shipweight': 0.073983,                 'origin_t': 0.0091761931,
'hourname': 0.07314349,                 'drugsname': 0.0090747,
'rest_usename': 0.06714,                'doors': 0.00891862558,
'vinyear': 0.065654704,                 'drivwhls': 0.00784183952,
'monthname': 0.06223856,                'drl_t': 0.007528650306002246,
'displci': 0.047397457,                 'fuel_t': 0.00740170521290,
'caseyear': 0.04689775,                 'drivetyp_t': 0.0065626245189,
'mfg_t': 0.026696833,                   'vehtype_t': 0.005505569656,
'veh_no': 0.02248278192937049,          'segmnt_t': 0.00332642834231,
'drinkingname': 0.0189898,              'wheels': 0.0031267301337496766,
'bodystyl_t': 0.017063655,              'tow_vehname': 0.002138869293,
'plntctry_t': 0.0170240351,             'spec_usename': 0.000568859488,
'vlvtotal': 0.01521295,                 'sch_busname': 0.00026803797
```

Of the above characteristics, 24 were selected by determining the treshold value. But then, considering that it is unethical to use 5 of these 24 data, the model was not used to train.

```
Selected_features=['age','ve_forms','statename','msrp','shipweight','ho
urname','vinyear''monthname','rest_usename','displci','caseyear','mfg_t
','veh_no','plntctry_t','drinkingname','bodystyl_t','vlvtotal','sexname
','cylndrs','doors','vlvclndr','hispanicname','drl_t','origin_t']

ethical= ['age','ve_forms','statename','msrp','shipweight','vinyear','r
est_usename','displci','caseyear','mfg_t','veh_no','plntctry_t','bodyst
yl_t','vlvtotal','cylndrs','doors','vlvclndr','drl_t','origin_t']
```

Part 6: Training and Performance Evaluation

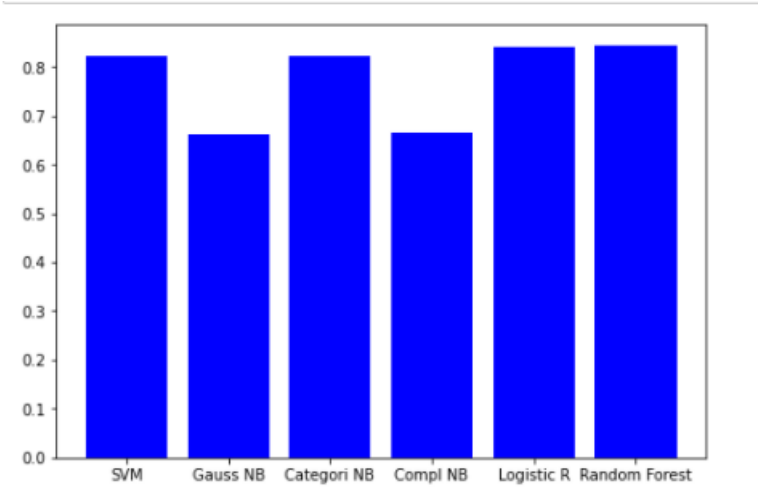
In this part, the data set was trained using four distinct methods in this section. Support Vector Machine (SVM), three distinct Naive Bayes algorithms, Logistic Regression, and lastly Random Forest methods are the algorithms in question. Each of these algorithms was individually trained using a data source. After then, students were supposed to estimate the test data set's y values. Classification reports for both test and train were prepared as a consequence of these estimations. The findings will be addressed further in the results section. When modeling Naive Bayes algorithms, MinMax Scalar was used to keep the values between 0 and 1 instead of using standard scalar unlike others.

The Bayes Theorem is used to create a statistical categorization approach known as Naive Bayes. It's one of the most basic supervised learning algorithms available. The Naive Bayes classifier is a quick, accurate, and trustworthy method. On big datasets, Naive Bayes classifiers have great accuracy and speed. [7]

"GaussianNB implements the Gaussian Naive Bayes algorithm for basic classification.

ComplementNB implements the complement naive Bayes (CNB) algorithm that is particularly suited for imbalanced data sets. Specifically, CNB uses statistics from the complement of each class to compute the model's weights.

CategoricalNB implements the categorical naive Bayes algorithm for categorically distributed data. It assumes that each feature, which is described by the index i , has its own categorical distribution." [8]



The scores of the models are listed as above. By looking at the scores, we can observe that ComplementNB, one of the Naive Bayesian models, is not very useful for our case.

The confusion matrices of the models are listed below:

TEST SVM

	precision	recall	f1-score	support
0	0.82	1.00	0.90	56998
1	0.00	0.00	0.00	12236
accuracy			0.82	69234
macro avg	0.41	0.50	0.45	69234
weighted avg	0.68	0.82	0.74	69234

TEST GaussianNB

	precision	recall	f1-score	support
0	0.91	0.65	0.76	56998
1	0.31	0.72	0.43	12236
accuracy			0.66	69234
macro avg	0.61	0.68	0.59	69234
weighted avg	0.81	0.66	0.70	69234

TEST CategoricalNB

	precision	recall	f1-score	support
0	0.82	1.00	0.90	56998
1	0.53	0.00	0.01	12236
accuracy			0.82	69234
macro avg	0.68	0.50	0.45	69234
weighted avg	0.77	0.82	0.74	69234

TEST ComplementNB

	precision	recall	f1-score	support
0	0.90	0.66	0.76	56998
1	0.30	0.67	0.41	12236
accuracy			0.66	69234
macro avg	0.60	0.67	0.59	69234
weighted avg	0.80	0.66	0.70	69234

TEST LogisticRegression

	precision	recall	f1-score	support
0	0.86	0.97	0.91	56998
1	0.64	0.24	0.35	12236
accuracy			0.84	69234
macro avg	0.75	0.61	0.63	69234
weighted avg	0.82	0.84	0.81	69234

TEST RandomForest

	precision	recall	f1-score	support
0	0.86	0.97	0.91	56998
1	0.65	0.26	0.38	12236
accuracy			0.85	69234
macro avg	0.76	0.62	0.64	69234
weighted avg	0.82	0.85	0.82	69234

Part 7: Interpretation

Finally, in this section, the interpretations of the trained models are given, and the results are discussed under the heading results.

Results

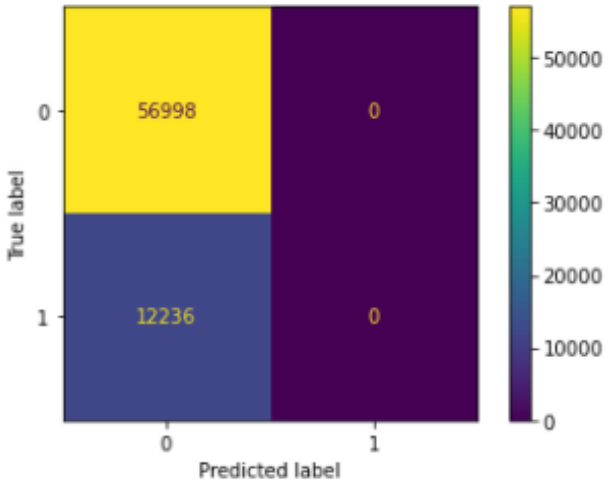
In Conclusion, when the scores of the trained models were compared at the end of the assessment procedure, the Random Forest model was the model that earned the best score. However, when the classification reports are studied, it becomes clear that Random Forest is capable of determining whether or not a rollover has occurred.

TRAIN

	precision	recall	f1-score	support
0	0.86	0.97	0.91	132837
1	0.69	0.27	0.39	28709
accuracy			0.85	161546
macro avg	0.77	0.62	0.65	161546
weighted avg	0.83	0.85	0.82	161546

TEST

	precision	recall	f1-score	support
0	0.86	0.97	0.91	56998
1	0.66	0.26	0.37	12236
accuracy			0.85	69234
macro avg	0.76	0.61	0.64	69234
weighted avg	0.82	0.85	0.82	69234



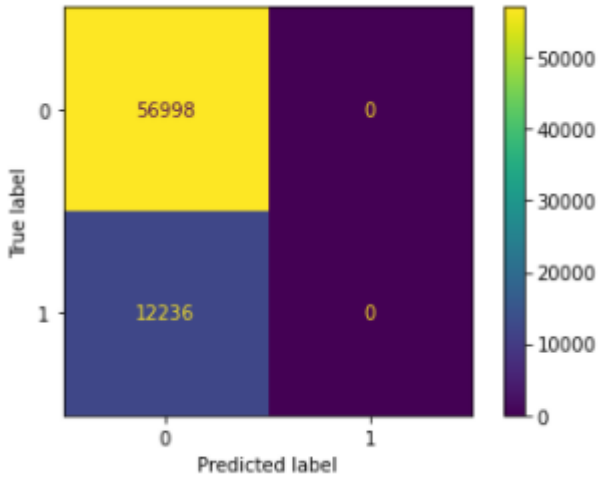
When it came to the interpretation section, it was decided that the best model was Random Forest and an examination was made through it. In these reviews, we investigated what are the most important features of it. As a result of this research, the model was retrained with the five most prominent features and, surprisingly, it was observed that it gave almost the same values as the 19 specifically trained models.

TRAIN

	precision	recall	f1-score	support
0	0.86	0.97	0.91	132837
1	0.64	0.28	0.39	28709
accuracy			0.84	161546
macro avg	0.75	0.62	0.65	161546
weighted avg	0.82	0.84	0.82	161546

TEST

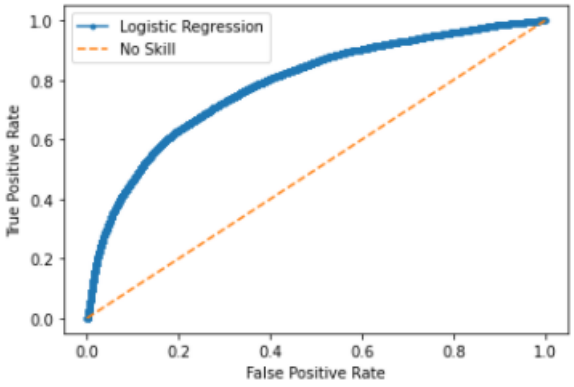
	precision	recall	f1-score	support
0	0.86	0.97	0.91	56998
1	0.63	0.27	0.38	12236
accuracy			0.84	69234
macro avg	0.75	0.62	0.64	69234
weighted avg	0.82	0.84	0.82	69234



These features turned out as follows:

```
['ve_forms', 'rest_usename', 'veh_no', 'vinyear', 'bodystyl_t']
```

In addition, it was observed that the Logistic Regression algorithm works similarly with Ran dom Forest, and the difference between the two is also in the macro average.



According to the roc_auc curve above, the ratio of true positive values is higher in the logisti c regression model. This shows that our model is good at finding the positive class, but the same is not true for the negative class.

The reason why we get these results with selected features may be mainly due to the fact tha t we automatically select the forest of random properties for feature significance in a classification prob lem.

At the end of the discussion, as a result of combining and organizing person and vehicle data sets containing car accidents between 2014 and 2019, it was thought that the data in the rollover colum n may play a role in determining whether the vehicles are damaged. As a result of this process, rollover was collected under two classes and treated as a binary classification problem. As a result of this, data containing unpredictable circumstances were extracted and a predictable model was tried to be prepare d for insurance companies. As a result of this study, it was decided that both Random Forest and Logis tic Regression algorithms can be used because they give similar results. As a result of the study of the i mportance of the characteristics, it was concluded that we can achieve similar results with five differen t data that we can actually obtain, and the compact state of inference was revealed.

Appendix

Data: https://drive.google.com/file/d/1aaL0qzXa3eHz1LNZEH22PvT0lDAJYz-3/view?usp=sharing

References

[1] Support Vector Machine Description: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[2] Naïve Bayes Algorithm Description https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[3] Logistic Regression Description https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

[4] Random Forest Description https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[5] One-Hot Encoding vs. Label Encoding using Scikit-Learn https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/

[6] How does the random forest feature importance work? https://www.quora.com/How-does-the-random-forest-feature-importance-work

[7] Naive Bayes Classification using Scikit-learn https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn

[8] Naïve Bayes Algorithms https://scikit-learn.org/stable/modules/naive_bayes.html#gaussian-naive-bayes