1. **A description of the problem and a discussion of the background. (15 marks)**

**Intro**

The scope of the project is to create a model for road traffic accident classification by applying machine learning techniques. The data set contains information about the accidents across the Seattle area. Data includes a rich set of features such as an address, location, latitude, longitude, accident type, the severity of the accident, description of the accidents, weather condition, road condition, visibility condition, number of people and cars involved, and so on.

The chosen dataset contains data from crashes since 1/1/2004 until 9/9/2019. It contains information that might influence the accident occurrences such as weather conditions and post-accident information such as case numbers. The data set comes with a pdf file containing the description of the available features and values.

Such an analysis might present interest to people and organizations wanted to understand the impact of exogenous variables such as environmental ones or road conditions and more or less the endogenous ones such as the speeding or under influence factor. A better understanding of the impact of available variables might provide some guidance on efficient resource allocation or policy development.

In addition, describing the importance and the impact of the variables at hand on accident occurrence might help insurance companies to review their cost and premium allocations.

**Project Outline**

This is the initial version of the final report for the project. The whole analysis of the data set is done using Python language and its libraries. The report, notebooks, code, and data are available on my GitHub profile [here](https://github.com/sarguo/Coursera_Capstone).

The goal of the project is to predict the severity of the crash and which variables influence the accident occurrence. In the data set, the target variable is called SEVERITYDESC and takes two values: “Property Damage Only Collision” and “Injury Collision”.

To achieve this objective, the following steps will be taken:

* Feature Exploration (with data cleaning)
* Dimensionality Reduction
* Model Building
* Optimization and final model selection.

Even though the steps have been defined above, during the exploration of each step in part we might have to go back and forth to adjacent steps to make necessary adjustments into the data structure. The final report will reflect the main points of the process without describing all the intermediate steps and will exclude irrelevant explorations that sometimes are needed to better understand the data and/or the problem. Although, I will mention paths that were identified to solve other interesting problems while working on the data.

An exhaustive exploration of the features will be done to identify valuable insights about the features and values, and how they correlate with one another.

At this point in time, we can state with confidence that our problem is a binary classification one, and will be used an F1 score to evaluate the performance of different models that will be trained.

1. **A description of the data and how it will be used to solve the problem. (15 marks)**

The dataset has an initial total of 194673 observations characterized by 38 different features. The feature that needs to be predicted is the “SEVERITYDESC”. It is a categorical one and takes two distinct values “Property Damage Only Collision” and “Injury Collision”.

Other features are both numerical and categorical. We have the environment, accident, location, type of location and accident, people and cars involved, and other accident describing features. The GitHub repository includes the data set in the csv format and a pdf file with a detailed description of the features and values included in the dataset. Therefore, please check the pdf file whenever is necessary to clarify the nature of the features or values.

The dataset contains many categorical features. Some of them take nominal values, thus no natural order to their values is present, instead are used for encoding purposes only, however other features describe the encodings. “SDOT\_COLDESC” and “SDOT\_COLCODE” are an example of such features. Either feature will have to be hot encoded eventually, thus as of now the preferred feature to keep will be the one that comes as a description, while another one will be dropped.

There are features that can be derived from one or more categorical features or values overlaps. Comparing “WEATHER” and “ROADCONDITION”, the “Rain” value relates to “Wet” condition. Thus, there is a strong correlation between some features, however, some of them will be kept and used to extract the most intelligence during the data exploration.

Eventually, those features that present excessive correlation will be removed as they are redundant and are harmful for the model.

The numerical features presented in the dataset are counters that indicate the number of pedestrians, cyclists, or vehicles involved in an accident.

1. **Exploratory Data Analysis**

I am starting with a data set that has (194673, 38) shape, which means 38 columns and 194673 observations. There are numerous columns that provides redundant information nor presenting any statistical values or due to high amount of missing values will have to be dropped.

I identified one duplicate column that has to be dropped: 'SEVERITYCODE.1'. Although columns 'ST\_COLCODE' and 'ST\_COLDESC' are different from 'SDOT\_COLCODE' and 'SDOT\_COLDESC', these columns contains similar information. Thus, 'ST\_COLCODE' and 'ST\_COLDESC' will be dropped. Even thought, 'ST\_COLCODE' and 'ST\_COLDESC' is more descriptive we will not lose any valuable information since 'SDOT\_COLCODE' and 'SDOT\_COLDESC' in tandem with other columns such as 'COLLISIONTYPE' provides similar information. In addition, SDOT\_COLCODE and SDOT\_COLDESC do not contain any missing values, while 'ST\_COLCODE' and 'ST\_COLDESC' registered 18 and 4905 missing values.

Columns 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'INTKEY', 'SDOTCOLNUM' do not provide any statistically relevant information for my further studies, thus these columns will be dropped.

Columns 'INCDATE' will be dropped since the same information can be found in the columns 'INCDDTM'.

Columns 'PEDROWNOTGRNT', 'INATTENTIONIND', 'EXCEPTRSNDESC', 'EXCEPTRSNCODE' will be dropped due to vast amount of missing data and/or irrelevant statistical information.

Most of the observations for columns 'CROSSWALKKEY' and SEGLANEKEY are registered as '0' which means most of the observations are missing. Thus, this feature will be dropped.

The 'COLLISIONTYPE' already contains the information about the involvement into the accident of the parked cars. Thus, the column 'HITPARKEDCAR' will be dropped.

SDOTCOLNUM represents the number given to the collision by SDOT. However, it has 79737 missing values. Thus, it can be dropped since these particular feature does not provide any valuable information for further studies.

We will keep only the 'SEVERITYDESC' column since the 'SEVERITYCODE' provides the same information and it will be dropped.

'X' and 'Y' represents the log and lat, thus names of the columns are changed accordingly.

Columns 'PEDCOUNT' and 'PEDCYLCOUNT' are merged under one non-vehicle feature 'NONVEH'. Combining these features and exploring its values we can observe that only 6% of the accidents shows involvement into the accident of pedestrians or cyclists.

We have reduced the number of columns of our working dataset from 38 features to 17.

**Values exploration: NaN, missing values, wring coding, and so on.**

The ‘UNDERINFL’ feature contains binary values however these values are registered in two formats as 1/0 and Yes/No. Thus, 1 and Yes, respectively 0 and No will be merged under Y/N format. We have NaN values, as well, however in case of some accident it is virtually impossible to ignore the under-influence state of the driver. Also, due to legal implication, by mistake no police officer will report a non-under influence as an under-influence. Thus, the NaN values can be safely converted to 'N'. This column shows that only 4.6% of the accidents involved under-influence drivers.

The 'SPEEDING' columns registers binary data. However, all the 'N' is registered as NaN. As with 'UNDERINFL' column, it is assumed that wrongfully registering non-speeding event as a speeding one are very unlikely. Vice-versa should hold true as well. Thus, all the missing values 'NaN' are converted to 'N'. About 4.8% of accidents involves speeding violation.

Features 'MONTH', 'DATE', 'TIME', 'WKDAY' are added to cross check the validity of variables from other columns, such as 'Daylight' from column 'LIGHTCOND' and we will try to extract some useful information based on new detailed data.

Our data set contains missing values.

|  |  |
| --- | --- |
| **LOG** | 5334 |
| **LAT** | 5334 |
| **ADDRTYPE** | 1926 |
| **LOCATION** | 2677 |
| **COLLISIONTYPE** | 4904 |
| **JUNCTIONTYPE** | 6329 |
| **WEATHER** | 5081 |
| **ROADCOND** | 5012 |
| **LIGHTCOND** | 5170 |

These 9 columns, identified above, present observations with missing values.

To some degree these values safely can be manipulated. As an example, missing values of the 'LIGHTCOND' feature can be replaced with value 'Daylight', since we have needed information in other ‘TIME’ column, which shows the hour when the accident occurred. In addition, values 'Dark - Unknown Lighting', 'Dark - Street Lights Off', 'Dark - No Street Lights' will be merged since it presents virtually the same information under different wording.

Similarly, it was identified 180 observations that were missing values for ‘LIGHCOND’; however we can extract the time of the day from ‘TIME’ column. Thus the NaN has been replaced with 'Daylight' attribute.

The ‘Unknown’ value will go through a similar process, by comparing with ‘TIME’ column, and values will be adjusted accordingly.

The 'ROADCOND' values are cross-checked against 'WEATHER' values. Thus, Unknown, NaN, or Other values under ROADCOND should be replaced with Wet if the WEATHER shows Raining/Snowing/Sleet/Hail/Freezing Rain.

On another hand, when the road is 'Wet' it is more likely under WEATHER should be ‘Raining’. Although, it could be Sleet/Hail/Freezing Rain or Snowing, considering the proportions/probability of occurrence the 'Raining' value is more likely to be appropriate for substitution.

Changes has been made on 'LIGHTCOND' based on daytime hours, and to features 'WEATHER' and 'ROADCOND' by cross-checking the available information. Similar approach can be taken while making needed adjustments to columns 'JUNCTIONTYPE' and 'COLLISIONTYPE'.

Checking values of both columns, it can be observed few correlated aspects that can be used to extract needed information to fill the NaN values. So, an accident happened to a 'Parked Car' is more likely to happen while a car is parked in either 'Mid-Block (not related to intersection)' or 'At Intersection (but not related to intersection)' or 'Driveway Junction'. But, since first option dominates in terms of frequency, it will be fairly safe to assume that if a 'NaN' or 'Unknown' values from 'JUNCTIONTYPE' matches the value of 'Parked Car' value from 'COLLISIONTYPE' than those values must be changed to 'Mid-Block (not related to intersection)'.

If 'JUNCTIONTYPE' takes value 'Mid-Block (but intersection related)' or 'At Intersection (intersection related)', it is more likely that the accident happened while trying to turn left. Again, the frequency for the left turn dominates, thus we will change the missing values for 'COLLISIONTYPE' to 'Left Turn' if it matches with the intersection related values of the 'JUNCTIONTYPE' column.

After above data manipulation, the column 'COLLISIONTYPE' still registers 330 NaN values, that will be replaced with 'Other' value. Similarly, the column 'JUNCTIONTYPE' contains 1692 NaN values that will be converted into 'Unknown' values.

The ‘LOCATION’ and ‘ADDRTYPE’ will be cross checked as well. 751 observation having values of NaN under the 'LOCATION', still shows under 'ADDRTYPE' that the accident happened on alley. Thus, 751 NaN observation, besides exact location, still offers some analytical information. So, these NaN observations will take 'Unknown' value, instead of dropping it. Other 1926 NaN values for LOCATION column are NaN under ADDRTYPE column as well.

At this point, there are still 1926 NaN values for both 'ADDRTYPE' and 'LOCATION' columns. The 'LOCATION' columns represent the exact address of the accident that occurred. Thus, to extract the exact address by exploring the data is not possible. However, we can try to address some NaN values for 'ADDRTYPE' columns by crosscheck it with 'JUNCTIONTYPE' and 'COLLISIONTYPE' columns. As an example, the 'Mid-Block' accidents in 'JUNCTIONTYPE' relates to 'Block' in 'ADDRTYPE' columns. Or, 'Intersection (but not related to intersection)' accidents in 'JUNCTIONTYPE' relates to 'Block' in 'ADDRTYPE' columns. Therefore, we can cross-check the NaN values between these columns and extract as much useful info as possible.

It can be observed that crosschecking information between above mentioned columns permitted to identify additional values for existing missing data. From 1926 of missing data we got left 574 NaN.

Even thought, NaN do not provide any information for the 'ADDRTYPE' and ‘LOCATION’ columns, the respective observations contain other descriptive and useful information for other features. Thus, we will not discharge completely these NaN observations, but rather will keep them under the 'Unknown' value.

|  |  |  |
| --- | --- | --- |
|  | **Pre-** | **Post-** |
| **LOG** | 5334 | 5334 |
| **LAT** | 5334 | 5334 |
| **ADDRTYPE** | 1926 | 0 |
| **LOCATION** | 2677 | 0 |
| **SEVERITYDESC** | 0 | 0 |
| **COLLISIONTYPE** | 4904 | 0 |
| **PERSONCOUNT** | 0 | 0 |
| **VEHCOUNT** | 0 | 0 |
| **INCDTTM** | 0 | 0 |
| **JUNCTIONTYPE** | 6329 | 0 |
| **SDOT\_COLDESC** | 0 | 0 |
| **UNDERINFL** | 4884 | 0 |
| **WEATHER** | 5081 | 0 |
| **ROADCOND** | 5012 | 0 |
| **LIGHTCOND** | 7170 | 0 |
| **SPEEDING** | 185340 | 0 |
| **NONVEH** | 0 | 0 |
| **DATE** | 0 | 0 |
| **MONTH** | 0 | 0 |
| **TIME** | 0 | 0 |
| **WKDAY** | 0 | 0 |

**NaN values at the beginning:** 1100024

**NaN values now:** 10668

We have been able to drop the number of missing values from 1.100.024 to

10.668 without losing any observation.

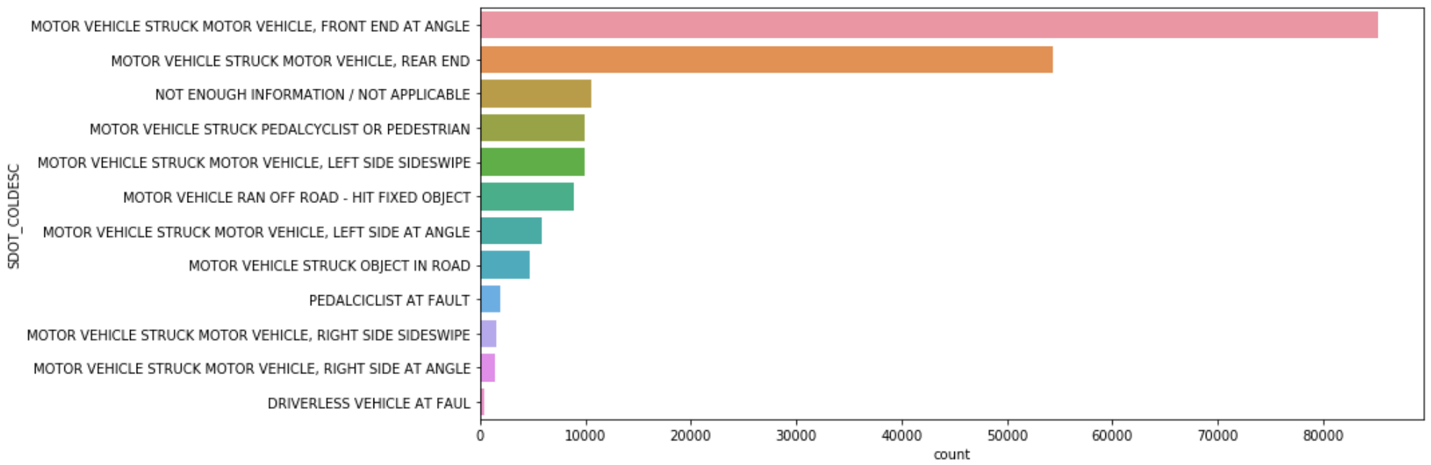
**Visual exploration of the features.**

In order to better understand the value of our data set, features require to be visually manipulated.

The ‘SDOT\_DESC’ variable presents a very descriptive classification of accidents. However, numerous classes can be merged under a single class. As an example would be, the driverless vehicles classes that can be merged under a single class. Same logic could be applied and for pedal cyclists or for vehicles that structed pedal cyclists from different angles.

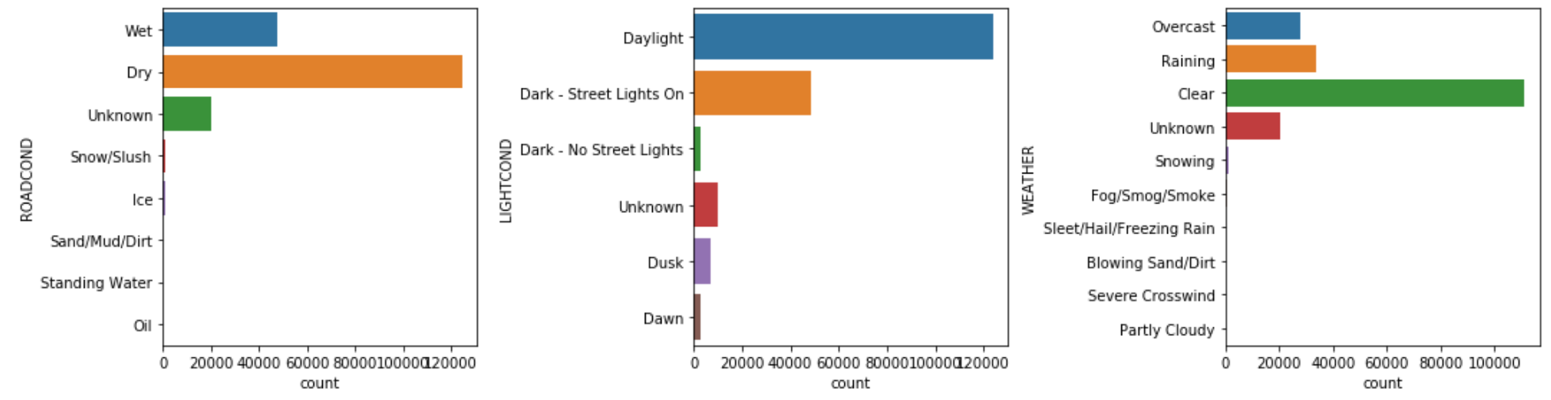


From the graph above, can be observed that "MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END AT ANGLE" and "MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END" group of accidents dominates the kind of accidents. Grouping together "MOTOR VEHICLE STRUCK AT PEDALCYCLIST" and "MOTOR VEHCILE STRUCK PEDESTRIAN" under a single group it would be the 3rd most often accidents registered. Since, PEDALCYCLIST and PEDESTRIANS are going to be treated as a single group, these categories will be grouped under a single category. "MOTOR VEHICLE STRUCK TRAIN", "MOTOR VEHICLE RAN OFF ROAD - NO COLLISION", and "MOTOR VEHICLE OVERTURNED IN ROAD" will be grouped under the "NOT ENOUGH INFORMATION / NOT APPLICABLE" since there are too few observations for these categories to extract some reliable intelligence.



After the regrouping our list of variables dropped from 39 to 12 without loosing any information but bringing improvements to letter models by having less groups to deal with.

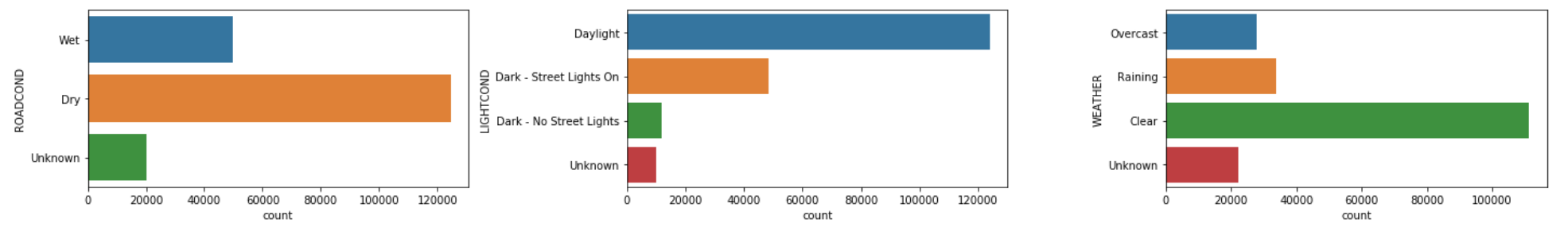
A visual investigation is done for ROADCOND, LIGHTCOND and WEATHER.



Under the ROADCOND options 'Snow/Slush', 'Ice', 'Standing Water', and 'Oil' will go under 'Wet' category, 'Sand/Mud/Dirt' goes under 'Unknown'.

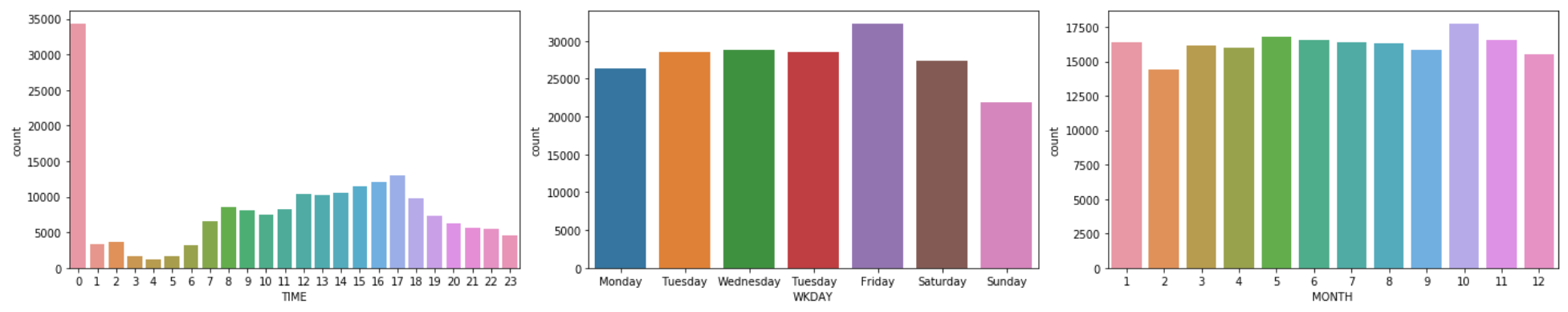
Under the LIGHTCOND options 'Dawn' and 'Dusk' will be merged under 'Dark - No Street Lights'.

Under the WEATHER options 'Snowing', 'Sleet/Hail/Freezing Rain', 'Fog/Smog/Smoke', 'Blowing Sand/Dirt', 'Severe Crosswind' and 'Partly Cloudy' classified under 'Unknown' category since the frequency of a such records have a very insignificant impact to be considered.



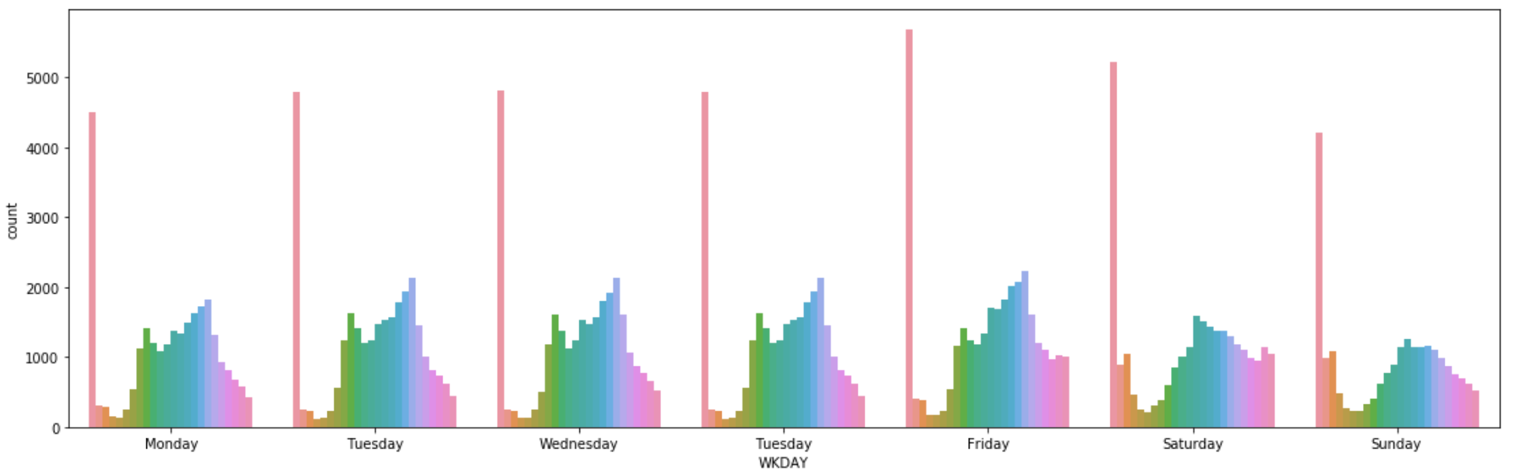
It can be observed that most accidents are happens under 'Dry' road conditions, during the 'Daylight', in a 'Clear' weather. This might be due to the fact that drivers are overconfident and paying less attention to the road. On another hand, during the bad weather/light condition/road condition there are less participants on the traffic.

Time related insights should be extracted out of 'TIME', 'WKDAY', 'MONTH' columns.

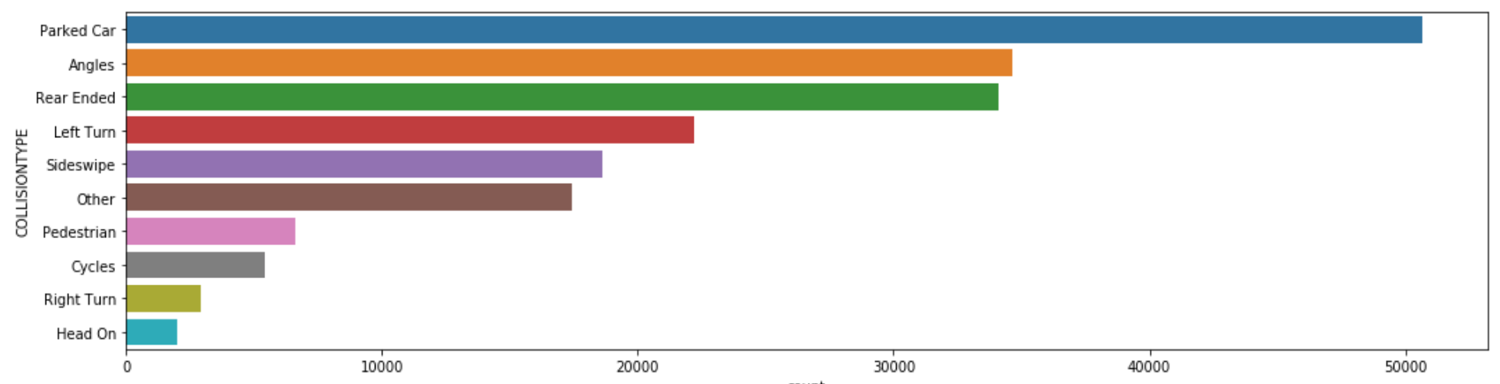


Most accidents, as expected, are registered during the peak hours (between 15 and 17). Overall, during the week the number of accidents is relatively equally distributed, although on Friday there are the most accidents that happens and on Sunday the least amount of accidents happens. As of 'MONTH' of the year there are not much of differences in the number of accidents. It can be observed that the most amount of accidents, although not by much, happens in Oct. A higher number for October probably is due to the fact that weather/road conditions is changing and requires a certain adaptability of the drivers. For February, the least amount of accidents, which is probably because Feb is a short month.

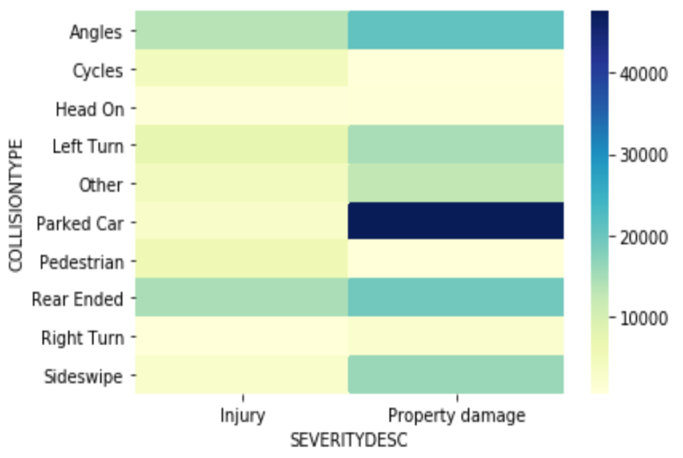
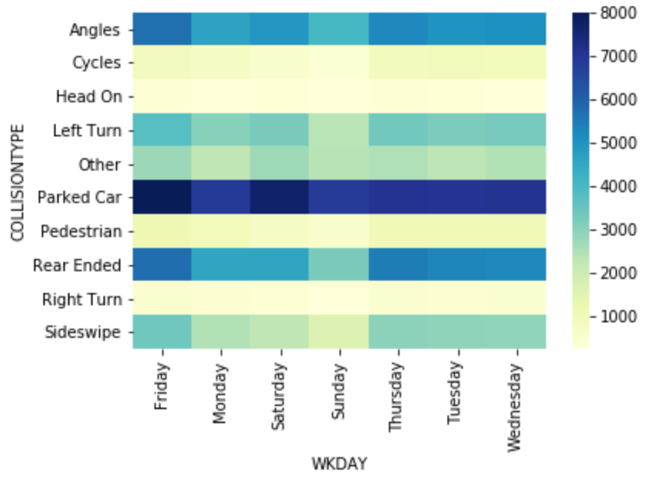
In fact, accidents in the evening and until early in the morning constantly decreasing in the number. Thus, around the midnight, considering the trend, should be around 4 thousand cases not over 34 thousand as the data shows us. One of the explanations is the fact that missing data takes midnight hour (0:00:00).



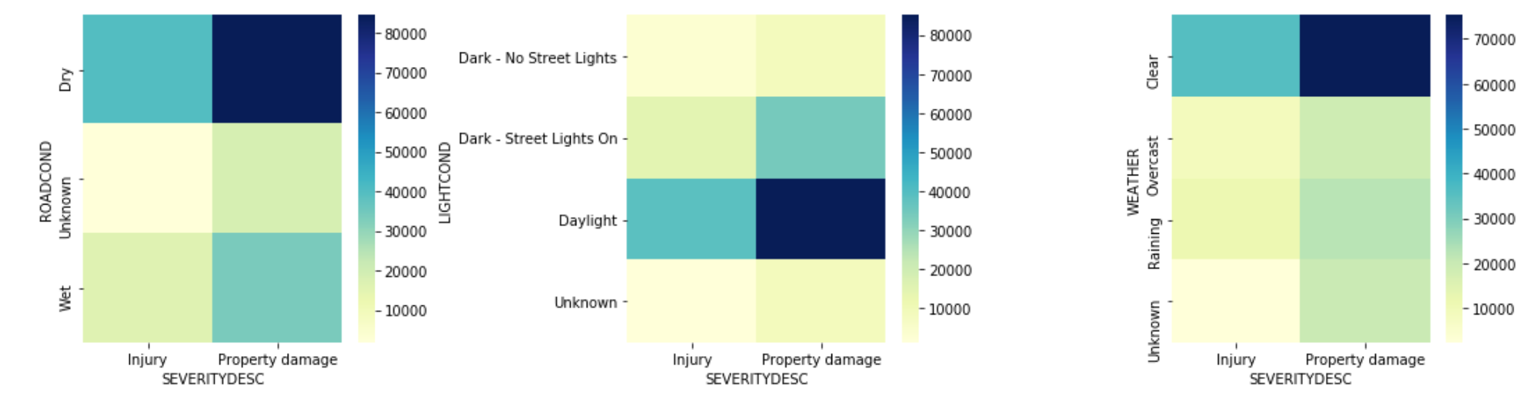
While during the week the most accidents happens during the peak hours, on weekends the most accidents happens around noon.



'Parked Car' related accidents (26%) dominates by a significant amount other type of accidents. Next most frequent types of accidents being 'Angles' (18%) and 'Rear Ended' (18%), which are having relatively about the same frequency of occurrences. However, 'Parked Car' related accidents are among the least categories of accidents that results in 'Injury' (1.6%), from accident severity perspective, while other two types of accidents mentioned above are leading the 'Injury' category (about 7% each). Another observation to mention is the fact that the frequency of occurrences of all three types of accidents mentioned here registers the highest amount on Friday and Saturday.

The severity of the accidents also can be classified under the environment condition.



Surprisingly, the most amount of accidents resulting in injuries (13%) happens in a dry road condition, under good visibility and clear weather. The same conclusion regarding the dry road condition, good visibility, and clear weather, holds true and for the accidents resulting in property damages only (26%). I would assume that overconfidence of the drivers in their driving skills affects negatively the safety on public roads.

1. **Results Section**

Several machine learning models will be tested on the existing data set: Decision Tree, Random Forest, Logistic Regression and KNN. To identify the most suitable algorithm for our objectives the entry data will go through pre-processing and normalization stages, of course when such steps are advisable. Results will be evaluated using similar evaluating tools and compared against each other based on accuracy score, F1-score, classification report, confusion matrix, and so on (Jackard Index and LogLoss when/where the use of such metrics are suitable).

**Decision Tree Classifier**

For Decision Tree Classifier categorical features will be converted to cardinal numbers, and in fact the new obtained data set will be the one used for other algorithms further.

Data set will be split into training and testing (0.7/0.3) data for model evaluation purposes. SEVERITYDESC serves as a target variable.

Decision Tree's Training data set Accuracy: 0.750915455232588

Decision Tree's Test data set Accuracy: 0.7483990274305674

Decision Tree Training Set Evaluation F1-Score=> 0.31070405946022783

Decision Tree Test Set Evaluation F1-Score=> 0.3079966092116417

precision recall f1-score support

0 0.74 0.99 0.85 40847

1 0.89 0.19 0.31 17555

accuracy 0.75 58402

macro avg 0.81 0.59 0.58 58402

weighted avg 0.78 0.75 0.68 58402

The recall for 'Property Damage' is pretty high (0.99) which implies that the model picks very well this kind of values of SEVERITYDESC. However, the social cost is not equal among these two values and is much higher for the 'Injury' value. A value of 0.19 for the recall for 'Injury' implies that our model predicts 'Injuries' only 19% of the time, even thought when it does it predicts with a probability of 89%. Since the 'False Negative' must be taken in consideration the F1-score is the metrics of choice. Unfortunately, the F1-score is low and other classification models must be evaluated.

**Random Forest Classifier**

In order to maintain the consistency on data usage, the Random Forest Classifier model uses same training and testing data as for Decision Tree Classifier.

**Random Forest model results**

Random Forest Accuracy score on train dataset: 0.9493656023658739

Random Forest Accuracy score on test dataset: 0.7290846203897127

Random Forest Training Set Evaluation F1-Score=> 0.912706846819493

Random Forest Testing Set Evaluation F1-Score=> 0.4896458293013353

precision recall f1-score support

0 0.78 0.86 0.82 40847

1 0.56 0.43 0.49 17555

accuracy 0.73 58402

macro avg 0.67 0.64 0.65 58402

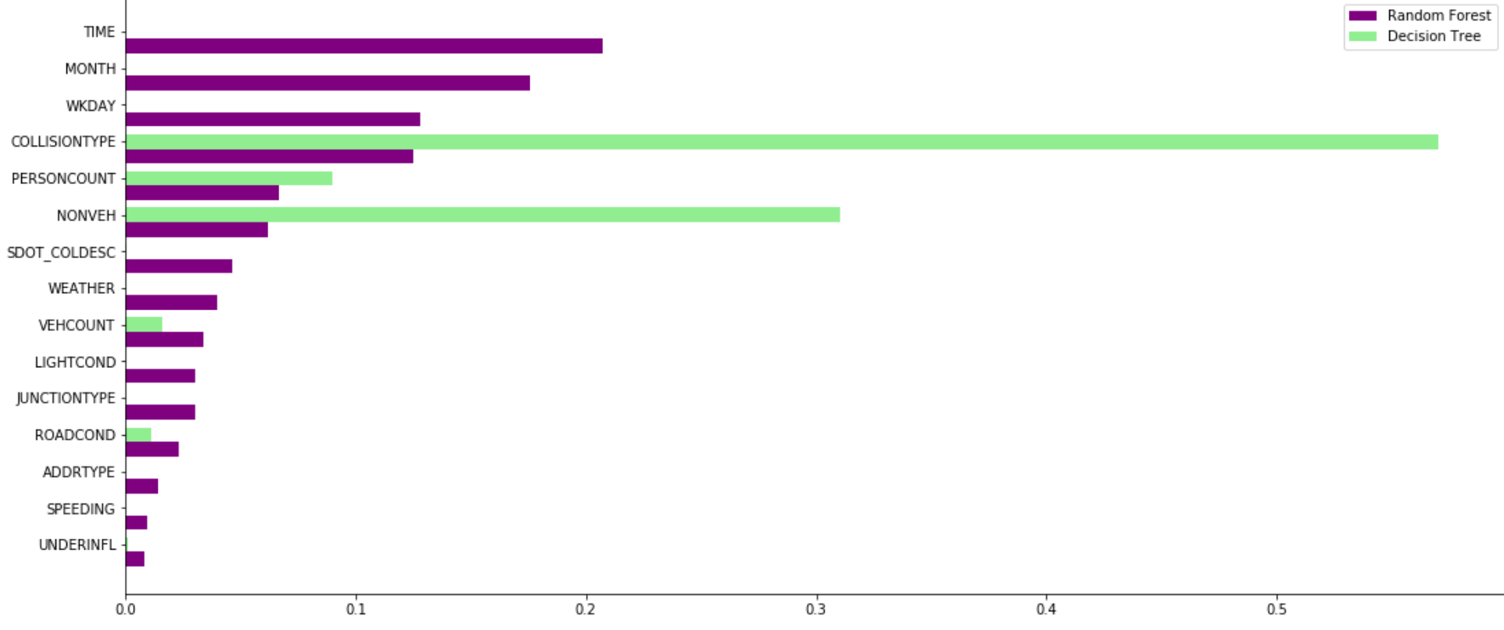
weighted avg 0.71 0.73 0.72 58402

By using the Random Forest model, we get an increase in F1 score from 0.30 to 0.48 in the out-of-sample evaluation without changing the accuracy score.

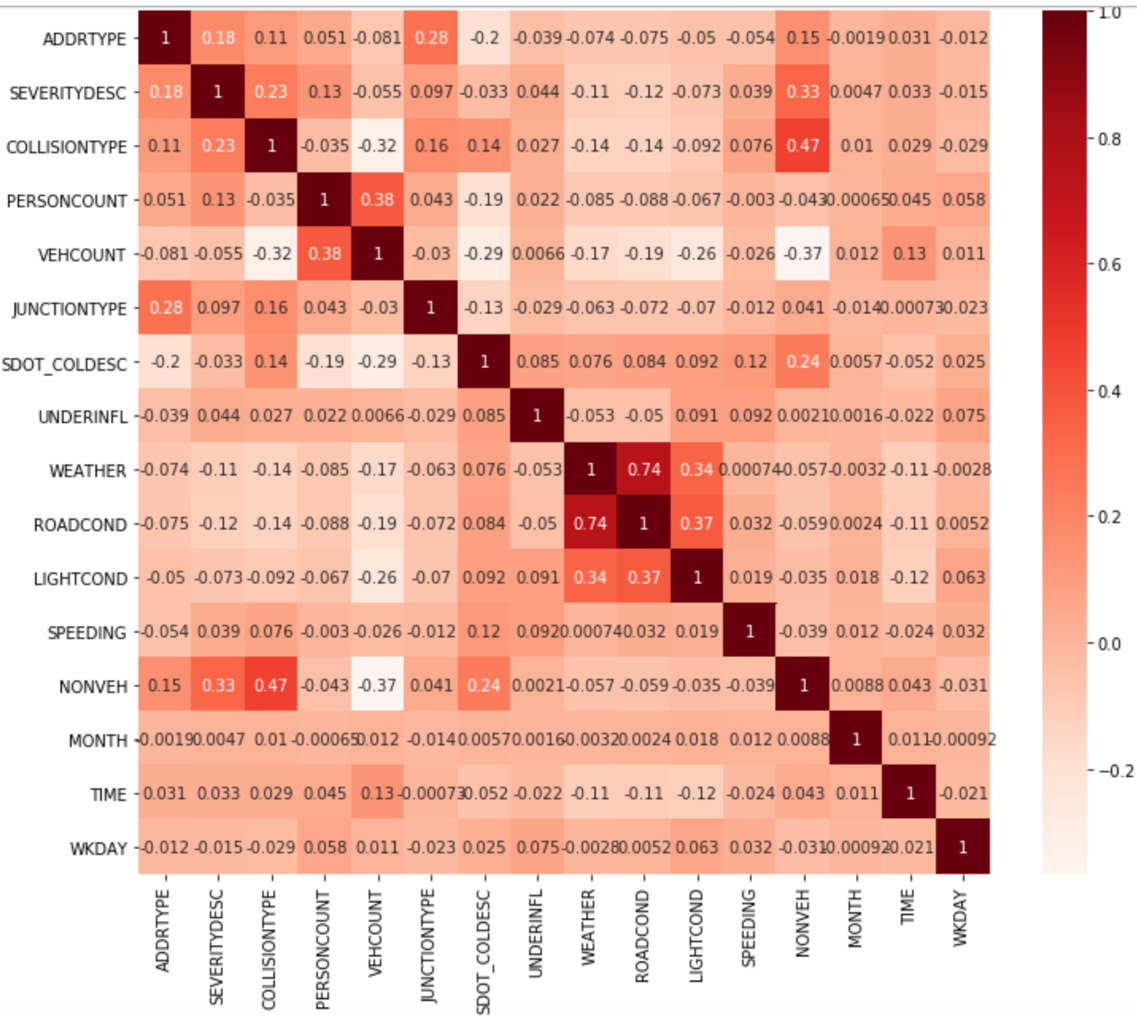
**Feature Importance**

In an attempt to add improvements or to reveal some more insights regarding the applicability of the models on our data set I will try to identify the importance of variables used in above models. This way, I might exclude the least important ones in hope to eliminate the 'noise' from the data set and hopefully it will help to show a better performance of the models.

The bar graph bellow shows the weights given for features for Decision Tree and Random Forest algorithms.



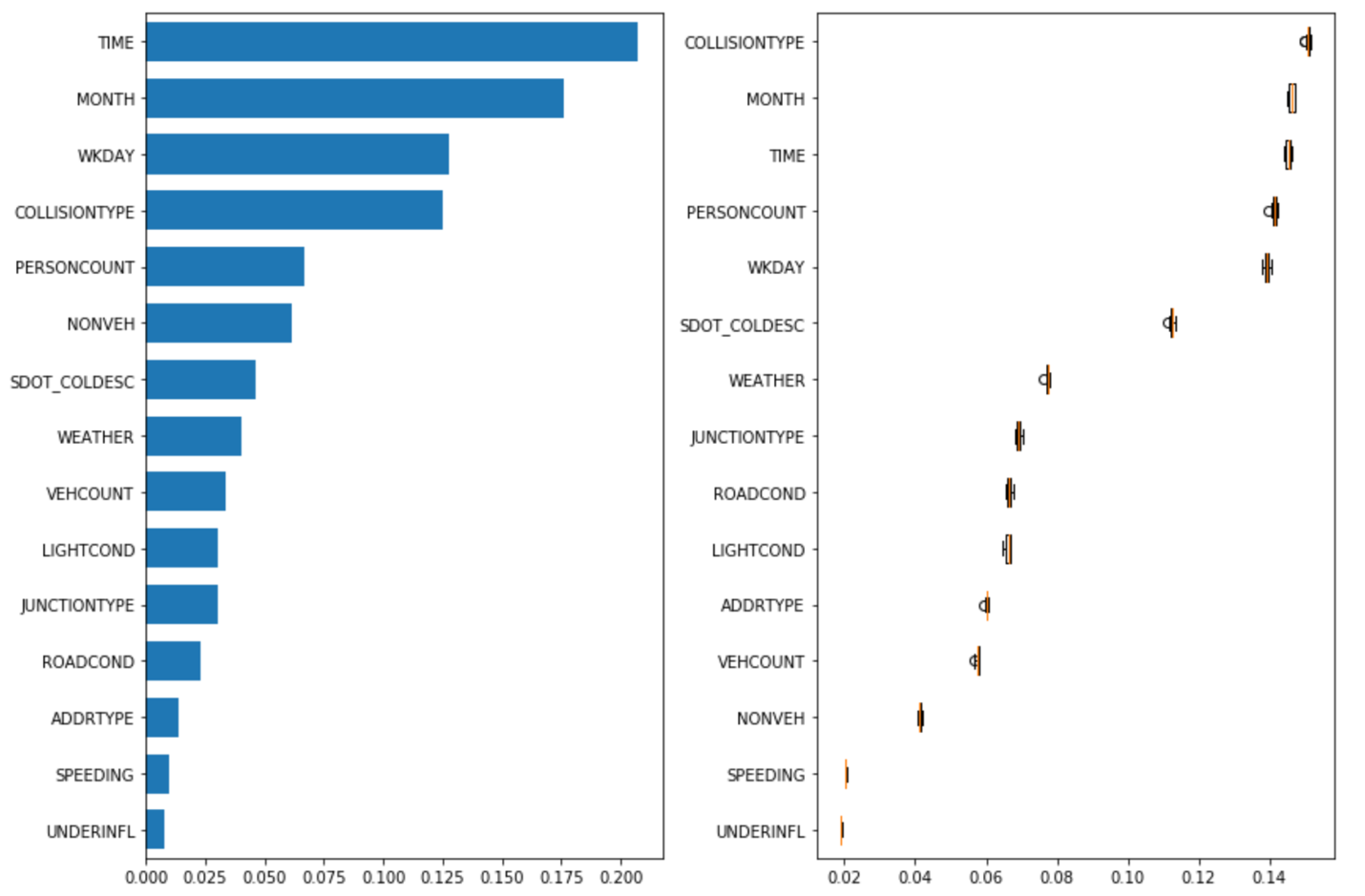
It can be noticed, that decision tree model puts a higher importance to a particular set of features, while random forest focuses on different set of features. Thus, I will evaluate the feature correlation, the permutation importance and eventually the "least" important ones will be dropped.



The correlation matrix implies that there is a relatively strong correlation between ROADCOND, LIGHTCOND and WEATHER, which is expected, considering the fact that we have done a lot of cross checking between these variables. Although the correlation matrix does not reveal which variables are worthy of dropping/keeping, the permutation importance algorithm might suggest some insights on this matter.

### Permutation Importance with Multicollinear or Correlated Features

According to the graph on the right, UNDERINFL and SPEEDING do not have a sufficient impact on predicted variable, thus it might be dropped. Although, the literature is suggesting to keep variables that has an accuracy impact of at least 0.05, I will drop NONVEH, VEHCOUNT as well, since these variables represent only the 'after the accident' status not something that can influence the occurrence of the accident itself.



**Reviewed Random Forest model results**

Reviewed Random Forest's Accuracy score on train dataset: 0.939451534075482

Reviewed Random Forest's Accuracy score on test dataset: 0.7225951166055957

Reviewed Random Forest's Training Set Evaluation F1-Score=> 0.8951681553102012

Reviewed Random Forest's Testing Set Evaluation F1-Score=> 0.48062065206937454

precision recall f1-score support

0 0.78 0.85 0.81 40847

1 0.55 0.43 0.48 17555

accuracy 0.72 58402

macro avg 0.66 0.64 0.65 58402

weighted avg 0.71 0.72 0.71 58402

The Reviewed Random Forest algorithm, with less features, provides a similar performance of predicting the severity of accidents by using less features. That translates into less time and resource usage to predict the severity of the accidents.

**Logistic Regression model**

Since the target feature takes binary values, it will be appropriate to use logistic regression algorithm and to check its performance against tree algorithms. While Decision Tree and Random Forest algorithms do not require normalization of the data, the Logistic Regression requires it. Thus, needed modules and steps will be taken.

The same variables as for 'Reviewed Random Forest' algorithm will be used since the logic behind dropping the un-needed features remains the same and for the logistic regression algorithm.

The Jackard Index obtained is a modest one: 0.7188, and a log-loss of 0.55.

**Logistic Regression model results**

precision recall f1-score support

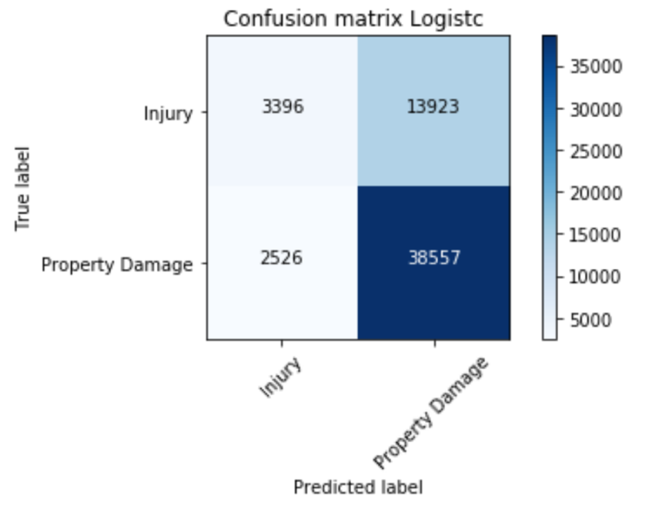
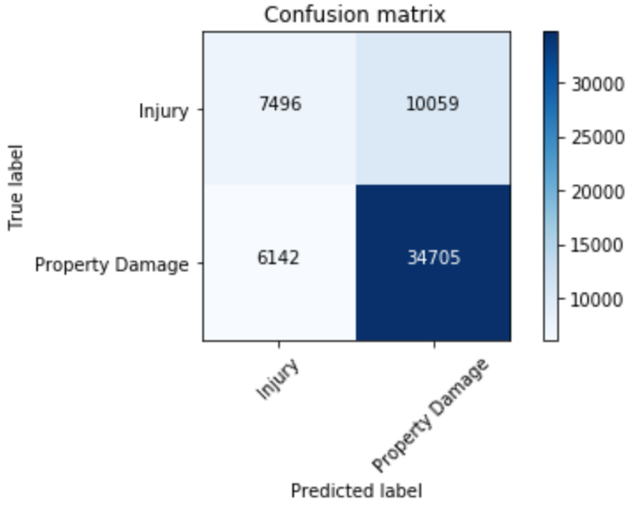
0 0.73 0.94 0.82 41083

1 0.57 0.20 0.29 17319

accuracy 0.72 58402

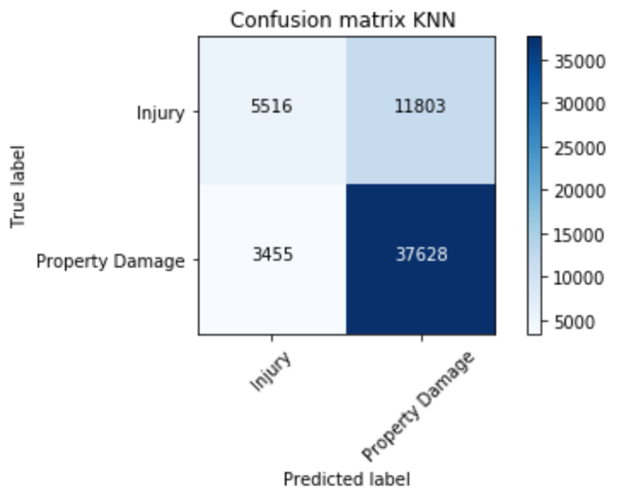
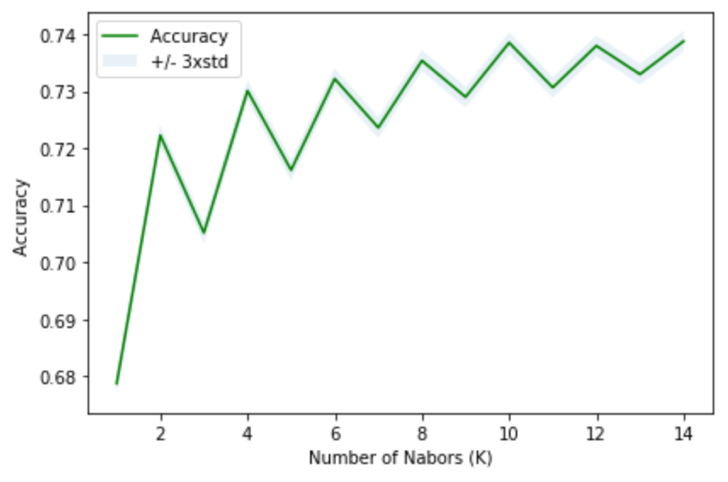
macro avg 0.65 0.57 0.56 58402

weighted avg 0.69 0.72 0.67 58402



The Random Forest algorithm (left) still predicts better the accident severity then Logistic Regression model. That holds true especially if we place higher social cost on Injuries than on Property Damage.

**K-Nearest Neighbors**



**K-Nearest Neighbors model results**

The best accuracy was with 0.7387418239101401 with k= 14

precision recall f1-score support

0 0.76 0.92 0.83 41083

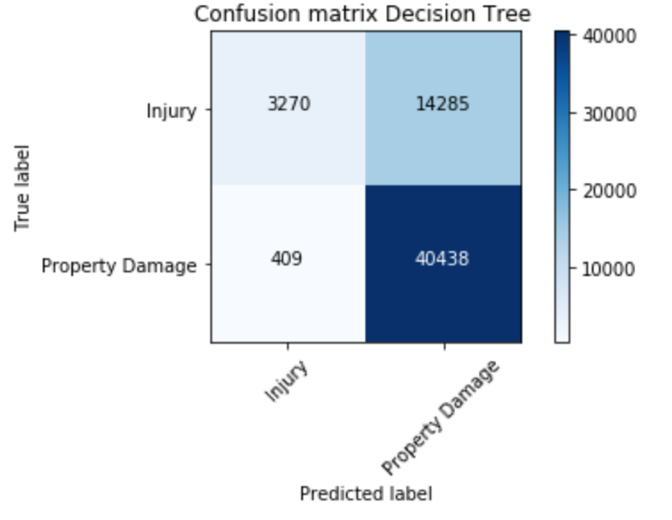
1 0.61 0.32 0.42 17319

accuracy 0.74 58402

macro avg 0.69 0.62 0.63 58402

weighted avg 0.72 0.74 0.71 58402

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Decision Tree** | **Random Forest** | **Logistic Regression** | **KNN** |
| **Accuracy** | 75% | 72% | 72% | 74% |
| **Precision** | 89% | 55% | 57% | 61% |
| **Recall (sensitivity)** | 19% | 43% | 20% | 32% |
| **F1 - Score** | 31% | 48% | 29% | 42% |
| **Specificity** | 99% | 85% | 94% | 92% |



For our purposes the Random Forest algorithm reveals the best prediction for the Injury, although results are modest. The table above summarizes the results. While the Decision Tree predicts 99% of the time of accidents resulting in Property Damage (fails), the Random Forest predicts the best (among presented algorithms) of Recalls (Sensitivity). High F1 score for Random Forest emphasis the importance of recalls from algorithm/technical standpoints as well as from social cost perception.

1. **Conclusion**

Present paper revealed some interesting points to consider when evaluating/predicting the accident severity. While accidents due to speeding or under influence are less than 5% of total, very few resulted in Injuries and statistically do not present any interest for our models. On another hand, accidents involving parked cars or rear ended, which in fact are the leading categories, implies that attention distraction of the drivers is at the culprit. Another supporting argument that most accidents are due to driver’s state of mind is the fact that most injuries are registered during a day in a clear and dry road condition, which implies that the overconfidence of the drivers might be the cause. But, these hypothesis must be statistically checked may be even considering additional features.