1 Business Understanding

It is a matter of fact that death tolls on our roads are increased when the drivers are confronted with non-optimal circumstances regarding wheather and road conditions. The goal of this project is to provide a guidline on how to reduce the risk of having a trafic accident. In this regard, it will be attempted to predict the risk of having accident under given circumstances. The calculated risk can be used as a metrics to decide whether the trip should be delayed or an alternative route or means of transportation should be used.

2 Data Understanding

The available dataset provides a collection of all reported collisions in Seattle from 2004 to present with many details (=features) on the occurrence. In total, 194673 rows and 38 features in the raw dataset. It could be seen that about 30% of collisions are categorie 2 (injury) and the rest categorie 1 (prop damage). No data for other categories is available.

It is the aim ofthis project to identify certain correlations of these features with the probability and severity of accidents.

The extensive number of features is collected, with sometimes very high degree of detail. E.g. the type of involved party as well as their direction of movement when the accident occurred, are separated. Attention needs to be paid for missing information. Due to non-continuous (categorical) features, it would make sense to exclude samples (rows) with incomplete set of features.

2.1 Selection of most important features

'INCDTTM': a time-dependency of probability/severity of accidents is expected

'INATTENTIONIND': the usage of cell phones and also being well rested when driving are important factors governing probability/severity of accidents

'UNDERINFL': consumption of alcohol is one of the main reasons for accidents

'WEATHER': impaired view, e.g. due to havy rain

'ROADCOND': the control over the vehicle can be lost due to bad road conditions

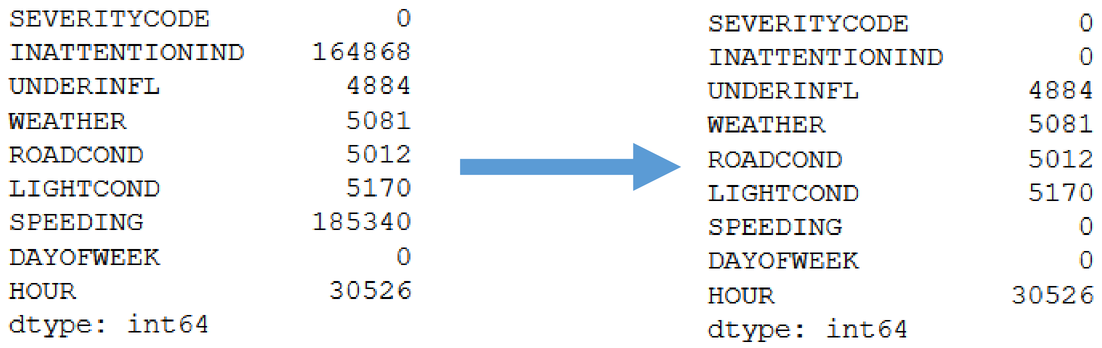
'LIGHTCOND': unfortunate incidence of light could impair sight as well --> probably only matters during daylight

'SPEEDING': by laws of physics, the a collision will be more severe at high velocity

3 Data Preparation

3.1 Homogenizing features

The features “INATTENTIONIND” and “SPEEDING” are considered important for our analysis. Since both features contains a huge number of empty rows (see left of arrows), it was assumed that an empty cell means a “No”. By filling the missing fields, the loss of a large chunk of the dataset could be prevented .



```
SEVERITYCODE           0
INATTENTIONIND    164868
UNDERINFL           4884
WEATHER             5081
ROADCOND            5012
LIGHTCOND           5170
SPEEDING          185340
DAYOFWEEK              0
HOUR               30526
dtype: int64
```

```
SEVERITYCODE           0
INATTENTIONIND         0
UNDERINFL           4884
WEATHER             5081
ROADCOND            5012
LIGHTCOND           5170
SPEEDING               0
DAYOFWEEK              0
HOUR               30526
dtype: int64
```

3.2 Deleting rows

In order to use a comprehensive dataset for modeling, the data must be further cleaned. In specific, rows will be deleted that have missing values for any of the features. The following shows the number of unique elements for each feature after cleaning of the dataset is finished:

```
SEVERITYCODE       2
INATTENTIONIND     2
UNDERINFL          2
WEATHER           11
ROADCOND           9
LIGHTCOND          9
SPEEDING           2
DAYOFWEEK          7
HOUR              24
```

The final dataset consists of 9 columns and 163704 rows.

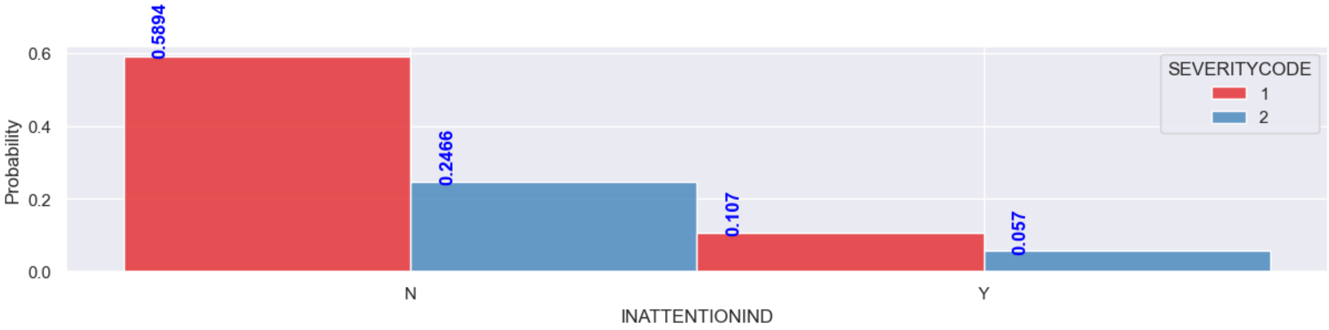
3.3 Encoding

Representing categorical features by numbers is a prerequisite to use the data in modeling. This is achieved by using the LabelEncoder of the sklearn.preprocessing module. The following shows the first 5 rows of the final feature matrix:

```
[[0 0 4 8 5 0 2 14.0]
 [0 0 6 8 2 0 2 18.0]
 [0 0 4 0 5 0 3 10.0]
 [0 0 1 0 5 0 4 9.0]
 [0 0 6 8 5 0 2 8.0]]
```

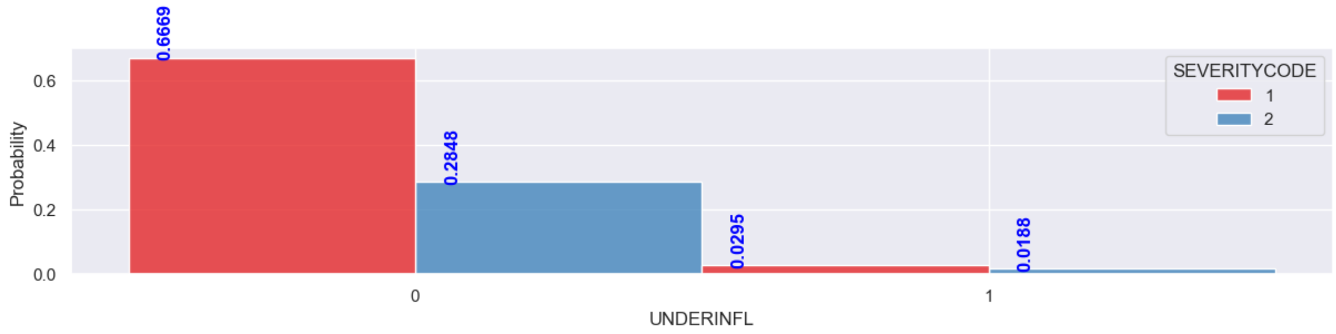
4 Data Assessment

4.1 Inattention



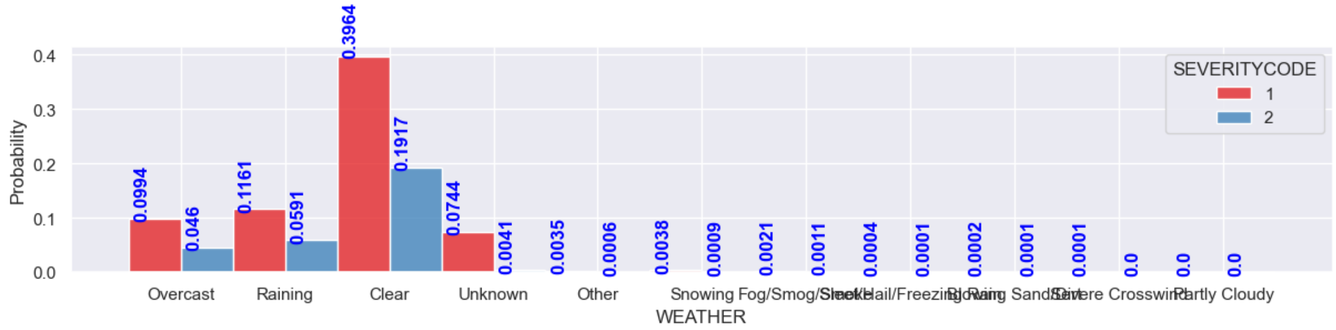
It is clear that not paying attention during driving increases the ratio of Cat2 collisions.

4.2 Driving under influence of alcohol/drugs/etc.



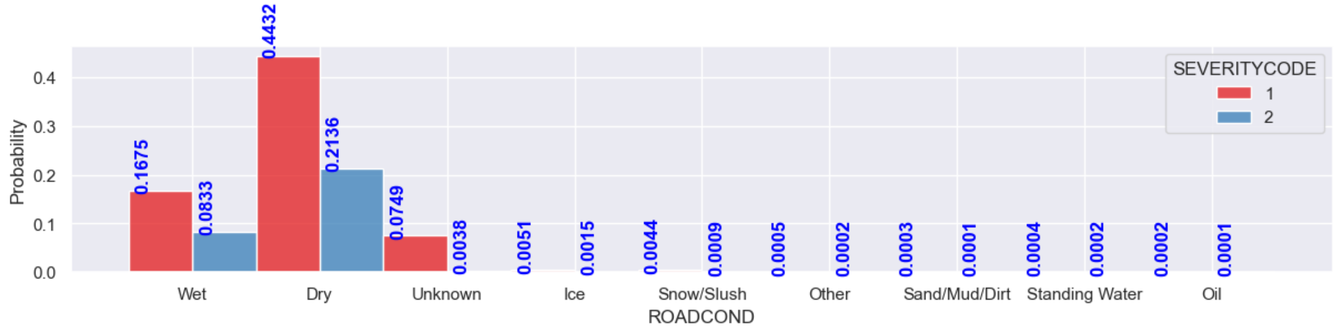
As before, driving under influence of alcohol/drugs/medication/etc. increases the probability of encountering more severe collision.

4.3 Weather



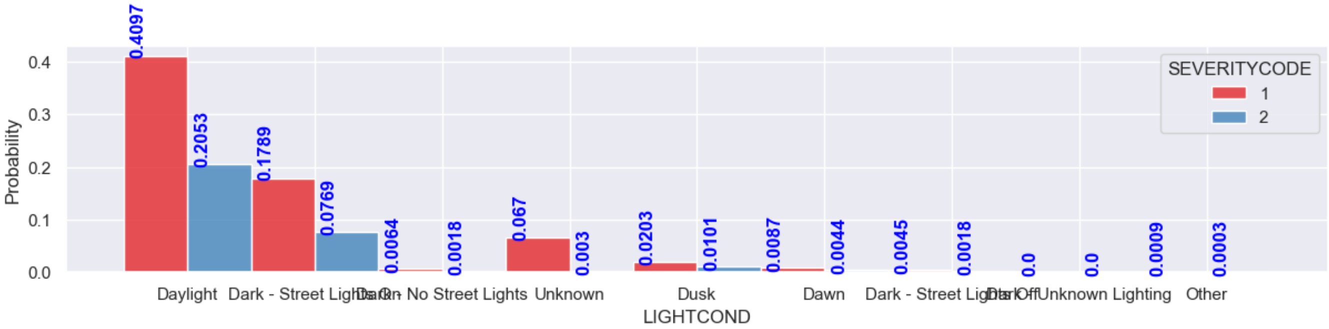
No clear dependence of occurrence and/or severity on weather can be identified,

4.4 Road conditions



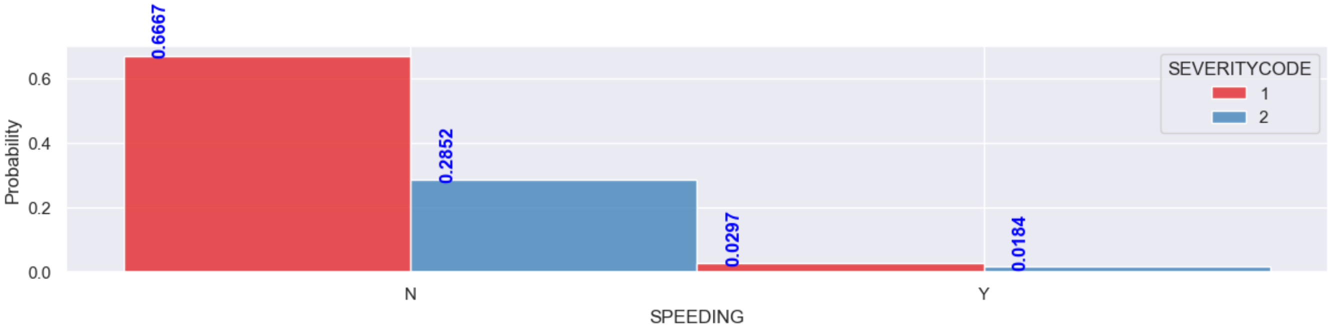
As with the weather, no clear dependence of occurrence/severity on road conditions.

4.5 Light conditions



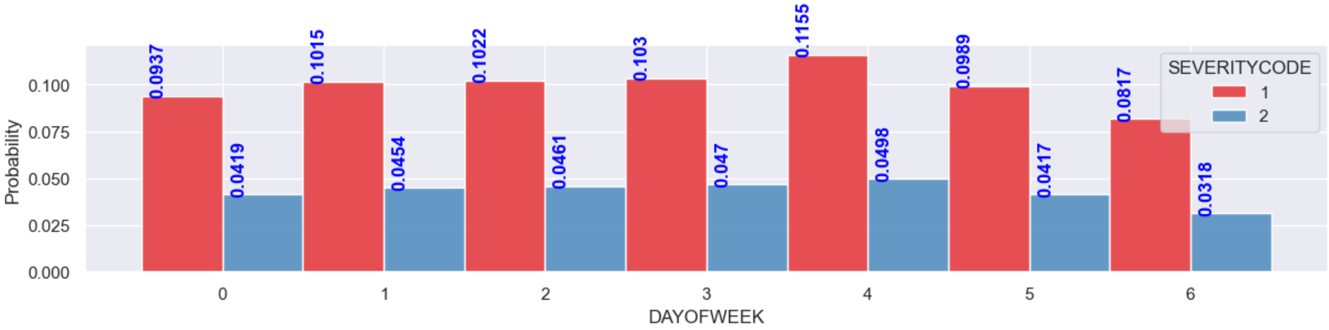
Light conditions also seem to be unimportant when it comes to predicting occurrence/severity of a collision.

4.6 Speeding



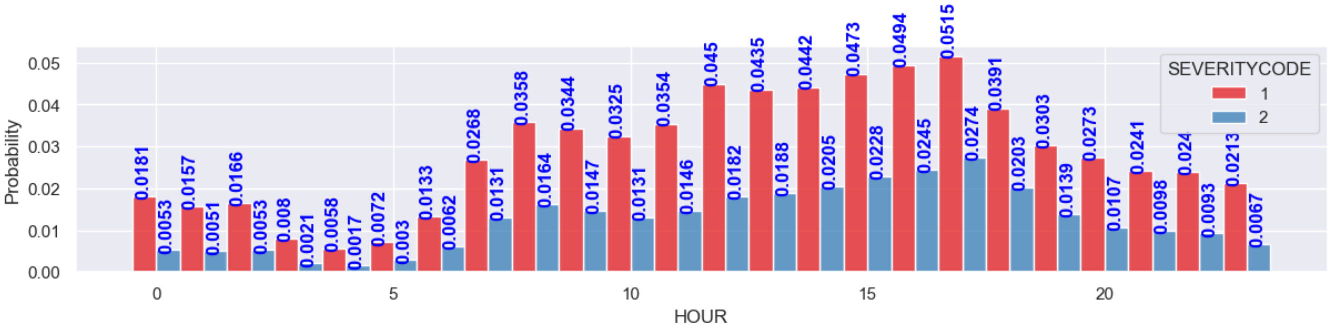
Clearly, speeding increases the probability of getting into a Cat2 collision.

4.7 Day of week



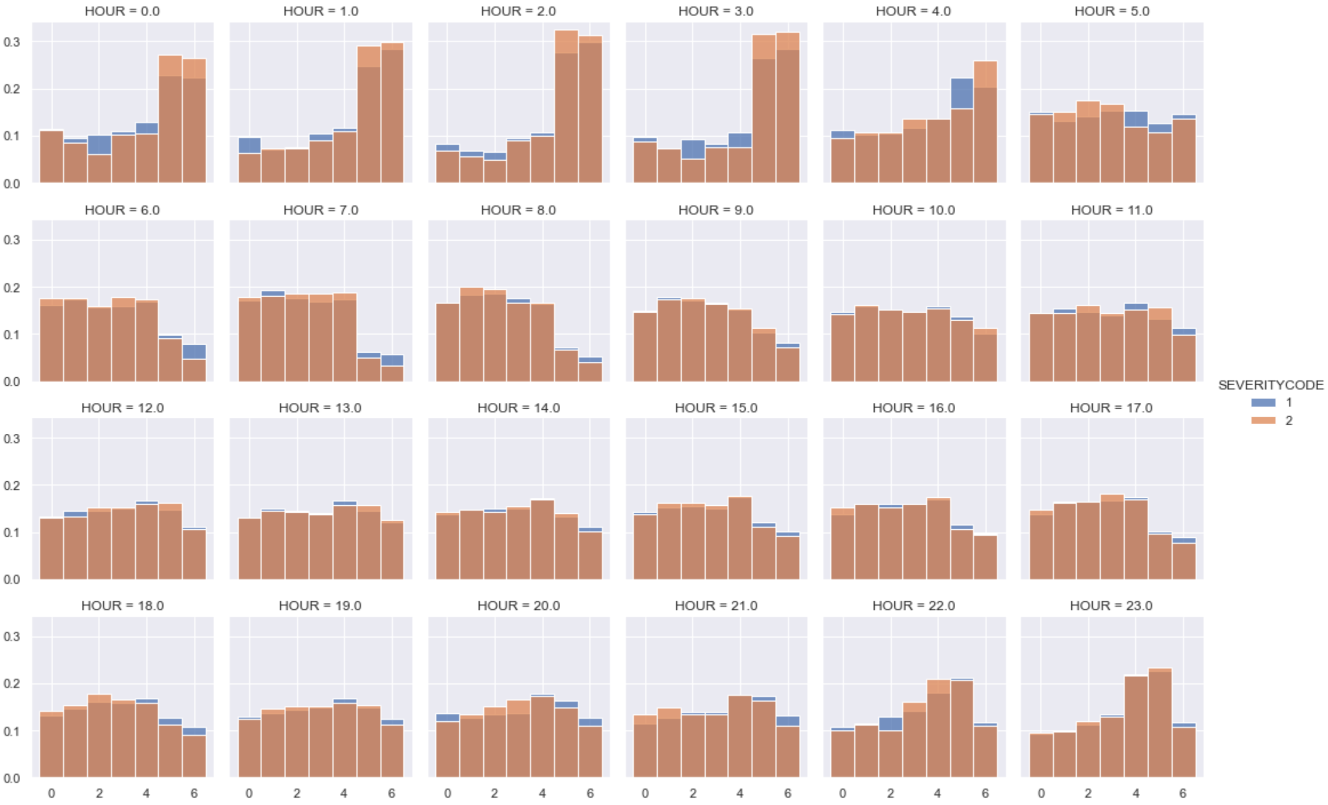
While no significant dependence of severity on day of week, the occurrence is lowest on Sundays.

4.8 Hour



Although no significant dependence of severity on hour, the occurrence is lowest at 4am.

4.9 Interdependence of day of week with hour



This plot shows the interdependence of hour and day of week. Surprisingly, the outcome is slightly different from previous plots that focused on a single feature. It can be seen that the occurrence is lowest on Sundays between 6 to 9am.

5 Modeling

5.1 Decision Tree (criterion="entropy", max_depth=4)

DecisionTrees accuracy:
f1 score: 0.5718137644087863
jaccard score: 0.6964356617085611

Confusion matrix:

```
[[91208     0]
 [39756     0]]
              precision    recall  f1-score   support

           1       0.70      1.00      0.82     91208
           2       0.00      0.00      0.00     39756

    accuracy                           0.70    130964
   macro avg       0.35      0.50      0.41    130964
weighted avg       0.49      0.70      0.57    130964
```

5.2 SVM (random_state=0, tol=1e-05)

SVM accuracy
f1 score: 0.5721603164855855
jaccard score: 0.6963952955552162

Confusion matrix:

```
[[91186    22]
 [39732    24]]
              precision    recall  f1-score   support

           1       0.70      1.00      0.82     91208
           2       0.52      0.00      0.00     39756

    accuracy                           0.70    130964
   macro avg       0.61      0.50      0.41    130964
weighted avg       0.64      0.70      0.57    130964
```

5.3 Logistic Regression (C=0.01, solver='newton-cg')

LogisticRegression accuracy
f1 score: 0.5728811935583709
jaccard score: 0.6958800696842813
Log Loss: 0.6029738428415058

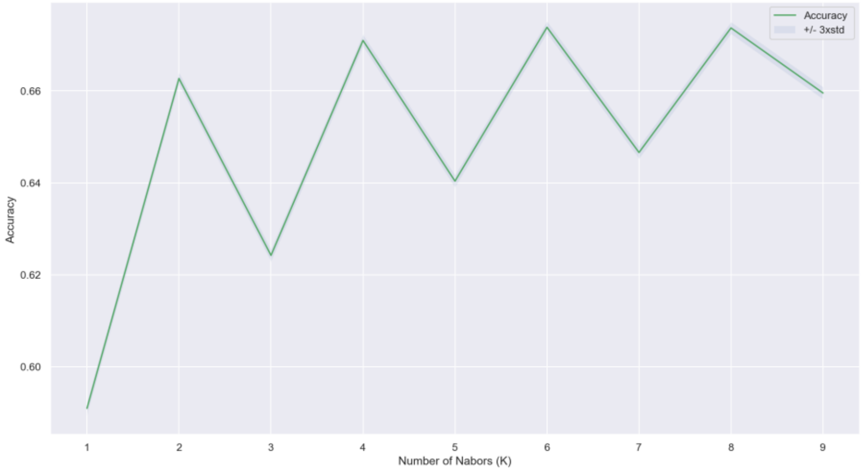
Confusion matrix:

```
[[91074   134]
 [39668    88]]
              precision    recall  f1-score   support
```

```
           1       0.70      1.00      0.82     91208
           2       0.40      0.00      0.00     39756

    accuracy                           0.70    130964
   macro avg       0.55      0.50      0.41    130964
weighted avg       0.61      0.70      0.57    130964
```

5.4 KNN



KNN (k= 6) accuracy
f1 score: 0.5995373342548063
jaccard score: 0.66082184025539

Confusion matrix:

```
[[84042  7166]
 [35970  3786]]
              precision    recall  f1-score   support

           1       0.70      0.92      0.80     91208
           2       0.35      0.10      0.15     39756

    accuracy                           0.67    130964
   macro avg       0.52      0.51      0.47    130964
weighted avg       0.59      0.67      0.60    130964
```

5.5 Modeling summary

Classifier	f1 score	jaccard score	log loss
KNN (k=8)	0.6025756214094656	0.659787832536549	N/A
Decision Tree	0.5722303145913911	0.6967086937399208	N/A
SVM	0.5734033971092973	0.6965685016075598	N/A
Logistic Regression	0.6022789278407159	0.5742063298880061	0.6960382597208802

6 Conclusion

It is very hard to use 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'SPEEDING' to predict occurrence since it is not known how many trips there were without any accident under the given circumstances. However, these features are still valuable to assess the severity of potential accidents.

In this regard, the features 'INATTENTIONIND', 'UNDERINFL', 'SPEEDING' show that the ratio of level 2 collisions are significantly above 30%, which means a higher risk of more severe accident when the driver does not pay attention, is drunk or is speeding.

The most intriguing finding is that the lowest number of accidents are occurring between 6 to 9 o'clock in the morning. However, no significant impact of incident time on the severity can be observed.

In conclusion, it can be stated that non-urgent trips should be carried out considering following requirements:

- pay attention to the traffic, i.e. do no use your cell phone, etc.
- do not be under influence of alcohol, medicaments or drugs while driving
- adhere to the speed limits
- drive between 6 to 9am

The best way to predict the occurrence and severity of a collision is to use a KNN classifier with k=8. All other investigated classifiers (SVM, Decision Tree, Logistic Regression) show extremely bad recall performance for categorie 2. This means that categorie 2 was hardly ever predicted using latter three classifiers.

7 Deployment

In case of non-urgent trips, the above presented model should be considered to decide when to undertake the trip.