Final Report
Group 8

Jia-Ern Pai

Ethan Litman

Jichong Wu

1 Introduction

1.1 Background

Thousands of vehicles involved rollovers in fatal traffic crashes every year. The rollover is one of the significant safety problems in fatal crashes. The National Highway Traffic Safety Administration (NHTSA) indicated that 18.9 percent fatal crashes in 2014 (7,592 of 40,164) involved rollovers[1]. Occupants in a rollover crash have greater likelihood of experiencing fatal injuries than occupants in non-rollover crash. Reducing rollovers in traffic crashes will decrease fatalities.

1.2 Previous Studies

NHTSA has studied rollovers in traffic crashes over the years, and NHTSA had the following conclusions.

1. The vehicle geometric properties, such as the height of the mass center and the track width are significantly related to the likelihood of rollover events[2].

2. The vehicle safety equipment, Electronic Stability Control (ESC) can significantly reduce the likelihood of rollover events[3]. NHTSA required automobile manufacturers to install ESC on their products. The following is the required ESC installation rate.

MY	Required ESC installation rate
2009	55% of vehicles
2010	75% of vehicles
2011	95% of vehicles
2012	100% of vehicles

1.3 Data Source

This project used the Fatality Analysis Reporting System (FARS) provided by NHTSA. FARS is a census of fatal traffic crashes that includes the 50 States, the District of Columbia and Puerto Rico since 1975. The crashes in FARS must result in the death of at least one person within 30 days of the crash. NHTSA has a cooperative agreement with an agency in each State government to provide information in a standard format on fatal crashes occurring in the State.

[1] Traffic Safety Facts 2014: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. (Report No. DOT HS 812 261). Washington, DC: National Highway Traffic Safety Administration.
[2] Pai J. (2017, August) Trends and Rollover-Reduction Effectiveness of Static Stability Factor in Passenger Vehicles NHTSA Evaluation (Report No. DOT HS 812 444), Washington, DC: National Highway Traffic Safety Administration.
[3] Sivinski R. (2011, June) Crash Prevention Effectiveness of Light-Vehicle Electronic Stability Control: An Update of the 2007 NHTSA Evaluation (Report No. DOT HS 811 486). Washington, DC: National Highway Traffic Safety Administration.

The data observations in FARS came from police crash reports in the States, death certificates, State coroners and medical examiners, State driver and the vehicle registration records and emergency medical service records. NHTSA's FARS datasets can be downloaded at https://www.nhtsa.gov/content/nhtsa-ftp/251.

FARS database includes the following datasets each year:

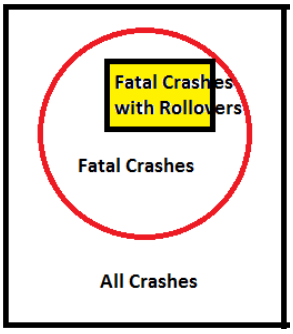
- Accident dataset: The variables in the accident dataset describe the circumstances of a fatal crash. Variables, such as the weather condition, roadway condition and local speed limit are belonged to the accident dataset.
- Person dataset: The variables in the person dataset describe the characteristics of occupants that experienced a fatal crash. Variables, such as the number of occupants in a crash event, driver's gender and driver's age are belonged to the person dataset.
- Vehicle dataset: The variables in the vehicle dataset describe the characteristics of vehicles in a fatal crash. Variables, such as the MY and first impact direction are belonged to the vehicle dataset.

1.4 Limitations of Data Source

There are two main limitations when using FARS.

- Crash information: FARS has a limited number of variables, but fatal crashes in the reality are complicated. FARS cannot describe each fatal crash in details. For example, in a multi-vehicle crash, researchers might not be able to tell the reason that cause a rollover, since the vehicle that involved a rollover might strike or be struck by other vehicle.

- Inference Application: The following Venn diagram shows the traffic crashes in FARS.



It is not appropriate to apply the analysis results based on FARS to all traffic crashes, since FARS only includes the traffic crashes where at least one occupant was fatally injured.

1.5 Target Population

From 2004 to 2016, NHTSA indicated that the average age of automobiles in operation in the United States is about 10-year[4], and this project used the FARS ranged from 2014 to 2018 with the restriction of MY ranged from 1989 to 2019. This project also applied the following conditions to filter FARS.

- Number of vehicles in a fatal crash: This project only considered the single-vehicle fatal crashes, since single-vehicle fatal crashes are easier to be interpreted by the FARS variables than multi-vehicle fatal crashes.
- Vehicle type: This project only included passenger vehicles, such as sedans, SUVs/CUVs, pickup trucks, and vans with the weight less than or equal to 10,000 pounds. Vehicles, such as buses and cargo cabs with the weight greater than 10,000 pounds follow different traffic rules.

1.6 Purpose of Studies

The purpose of this project is to predict rollovers in fatal single-passenger-vehicle crashes.

2 Exploratory Data Analysis

2.1 Target Variable

The target variable is the rollover status in a fatal single-passenger-vehicle crash. The target variable is a binary variable with two levels: rollover and non-rollover.

Base on the FARS ranged from 2014 to 2018 with the restriction of MY ranged from 1989 to 2019, the following table shows the frequency and rate of rollovers in single-passenger-vehicle fatal crashes.

	Frequency	Rate
Rollover	21820	35.28%
Non-Rollover	40021	64.71%

There are 35.28 percent of single-passenger-vehicle fatal crashes involved rollovers. The rollover is a significant safety problem in single-passenger-vehicle fatal crashes. To examine the trend of rollovers, the following table shows the frequency and rate of rollovers in the target population by year.

Year	Rollover	Non-Rollover
2014	4548 (37.34%)	7631 (62.66%)
2015	4685 (36.75%)	8060 (63.25%)
2016	4550 (35.42%)	8294 (64.58%)
2017	4238 (34.45%)	8065 (65.55%)

[4] https://www.bts.gov/content/average-age-automobiles-and-trucks-operation-united-states#:~:text=as%20of%20Sep.-,17%2C%202019.,%2Dmarkit%2D%20as%20of%20Sep.

2018	3799 (32.27%)	7971 (67.73%)

There are 3,799 single-passenger-vehicle crashes that involved rollovers in 2018. The rollover rate in 2018 is 32.27 percent while the rollover rate in 2017 is 34.45 percent. The rollover rate decreases when the year increases. One possible reason is that proportion of vehicles without ESC installation was reduced over the years.

2.2 Feature Variables

Based on the NHTSA previous studies (see Section 1.2) and researchers' knowledge, the feature variables were selected from the Accident, Person and Vehicle datasets in FARS.

2.2.1 Accident Dataset: Weather Condition

The weather condition in the accident dataset was selected, since the weather condition might affect the driver's vision. The weather condition is a categorical variable that included the categories of clear/normal, fog/cloudy, rain/sleet, snow, and windy. The following contingency table shows the weather condition by rollover status.

Table 1: Contingency Table of Weather Condition and Rollover Status

	Rollover	Non-Rollover
Clear/Normal	15723 (35.54%)	28517 (64.46%)
Fog/Cloudy	4042 (35.96%)	7197 (64.04%)
Rain/Sleet	1638 (29.44%)	3926 (70.56%)
Snow	353 (51.23%)	336 (48.77%)
Windy	64 (58.71%)	45 (41.29%)

There are 15,723 single-passenger-vehicle crashes that involved rollovers when the weather is clear/normal. The rollover rate in the clear/normal weather is 35.54 percent while the rollover rate in the windy weather is 58.71 percent. The rollover rate in the windy weather is greater than the rollover rates in the other weather conditions.

The chi-square test was used to examine the association between the weather condition and rollover status. The level of significance was set at 0.05 through this report. The following hypothesis statements were used.

H_0: The weather conditions and rollovers are independent
H_1: The weather conditions and rollovers are not independent

The chi-square test was based on the data in Table 1. The following table shows the chi-square statistic and its p-value.

Chi-square statistic	P-value
189.7558	5.980968829657794e-40

The weather conditions and rollovers are not independent, since the p-value of the chi-square test is less than 0.05. The weather condition will be used as a feature variable when predicting the rollover status.

2.2.2 Accident Dataset: Light Condition

The light condition in the accident dataset was selected, since the light condition might affect the driver's vision. The light condition is a categorical variable that included the categories of dark, dawn/dusk, and light. The following contingency table shows the light condition by rollover status.

Table 2: Contingency Table of Light Condition and Rollover Status

	Rollover	Non-Rollover
Dark	9277 (40.72%)	13507 (59.28%)
Dawn/Dusk	965 (37.43%)	1613 (62.57%)
Light	11578 (31.73%)	24901 (68.27%)

There are 9,277 single-passenger-vehicle crashes that involved rollovers in the dark. The rollover rate in the dark is 40.72 percent while the rollover rate in the light is 31.73 percent. The rollover rate in the dark is greater than the rollover rates in the other light conditions.

The chi-square test was used to examine the significant association between the light condition and rollover status. The following hypothesis statements were used.

H_0: The light condition and rollover status are independent
H_1: The light condition and rollover status are not independent

The chi-square test was based on the data in Table 2. The following table shows the chi-square statistic and its p-value.

Chi-square statistic	P-value
500.5370	2.0406178824711528e-109

The light condition and rollover status are not independent, since the p-value of the chi-square test is less than 0.05. The light condition will be used as a feature variable when predicting the rollover status.

2.2.3 Vehicle Dataset: Roadway Surface

The roadway surface condition in the vehicle dataset was selected, since the roadway surface condition might affect the effectiveness of vehicle break system. The roadway surface condition is a categorical variable that included the categories of dry, oil, and wet roadway surfaces. The following contingency table shows the roadway surface by rollover status.

Table 3: Contingency Table of Roadway Surface and Rollover Status

	Rollover	Non-Rollover
Dry	18309 (35.65%)	33046 (64.35%)
Oil	202 (57.39%)	150 (42.61%)
Wet	3309 (32.65%)	6825 (67.35%)

There are 18,309 single-passenger-vehicle crashes that involved rollovers on the dry roadway. The rollover rate on the dry roadway is 35.65 percent while the rollover rate on the oil roadway is 57.39 percent. The rollover rate on the oil roadway is greater than the rollover rates on the other roadway conditions.

The chi-square test was used to examine the significant association between the roadway surface condition and rollover status. The following hypothesis statements were used.

H_0: The roadway surface and rollover status are independent
H_1: The roadway surface and rollover status are not independent

The chi-square test was based on the data in Table 3. The following table shows the chi-square statistic and its p-value.

Chi-square statistic	P-value
500.5370	2.0406178824711528e-109

The roadway surface and rollover status are not independent, since the p-value of the chi-square test is less than 0.05. The roadway surface will be used as a feature variable when predicting the rollover status.

2.2.4 Vehicle Dataset: Roadway Grade

The roadway grade in the vehicle dataset was selected, since the roadway grade might affect the effectiveness of vehicle break system. The roadway grade is a categorical variable that included the categories of grade and level roadways. The following contingency table shows the roadway grade by rollover status.

Table 4: Contingency Table of Roadway Grade and Rollover Status

	Rollover	Non-Rollover
Grade	7208 (44.91%)	8843 (55.09%)
Level	14612 (31.91%)	31178 (68.09%)

There are 7,208 single-passenger-vehicle crashes that involved rollovers on the grade roadway. The rollover rate on the grade roadway is 44.91 percent while the rollover rate on the level roadway is 31.91 percent. The rollover rate on the grade roadway is the greater than the rollover rate on the level roadway.

The chi-square test was used to examine the significant association between the roadway grade and rollover status. The following hypothesis statements were used.

H_0: The roadway grade and rollover status are independent
H_1: The roadway grade and rollover status are not independent

The chi-square test was based on the data in Table 4. The following table shows the chi-square statistic and its p-value.

Chi-square statistic	P-value
878.5009	4.629546644838622e-193

The roadway grade and rollover status are not independent, since the p-value of the chi-square test is less than 0.05. The roadway grade will be used as a feature variable when predicting the rollover status.

2.2.5 Vehicle Dataset: Roadway Alignment

The roadway alignment in the vehicle dataset was selected, since the roadway alignment might be related to the rollovers. The roadway alignment is a categorical variable that included the categories of curve and straight roadways. The following contingency table shows the roadway alignment by rollover status.

Table 5: Contingency Table of Roadway Alignment and Rollover Status

	Rollover	Non-Rollover
Curve	8997 (50.98%)	8650 (49.02%)
Straight	12823 (29.02%)	31371 (70.98%)

There are 8,997 single-passenger-vehicle crashes that involved rollovers on the curve roadway. The rollover rate on the curve roadway is 50.98 percent while the rollover rate on the straight roadway is 29.02 percent. The rollover rate on the curve roadway is greater than the rollover rate on the straight roadway.

The chi-square test was used to examine the significant association between the roadway alignment and rollover status. The following hypothesis statements were used.

H_0: The roadway alignment and rollover status are independent
H_1: The roadway alignment and rollover status are not independent

The chi-square test was based on the data in Table 5. The following table shows the chi-square statistic and its p-value.

Chi-square statistic	P-value
2664.3365	0.0

The roadway alignment and rollover status are not independent, since the p-value of the chi-square test is less than 0.05. The roadway alignment will be used as a feature variable when predicting the rollover status.

2.2.6 Vehicle Dataset: Vehicle Type

The vehicle type in the vehicle dataset was selected, since NHTSA indicated that the vehicle geometric properties are significantly related to the likelihood of rollover (see Section 1.2). The vehicle type is a categorical variable that included the categories of car, pickup truck, SUV/CUV, and van. The following contingency table shows the roadway alignment by rollover status.

Table 6: Contingency Table of Vehicle Type and Rollover Status

	Rollover	Non-Rollover
Car	8704 (28.86%)	21448 (71.14%)
Pickup Truck	6097 (41.56%)	8574 (58.44%)
SUV/CUV	6212 (44.27%)	7819 (55.73%)
Van	807 (27.02%)	2180 (72.98%)

There are 8,704 cars that involved rollovers in single-vehicle crashes. The rate of rollovers experienced by cars is 28.86 percent while the rate of rollovers experienced by SUV/CUV is 44.27 percent. The rate of rollovers experienced by SUV/CUV is greater than the rate of rollovers experienced by the other types of vehicles.

The chi-square test was used to examine the significant association between the vehicle type and rollover status. The following hypothesis statements were used.

H_0: The vehicle type and rollover status are independent
H_1: The vehicle type and rollover status are not independent

The chi-square test was based on the data in Table 6. The following table shows the chi-square statistic and its p-value.

Chi-square statistic	P-value
1382.5924	1.763942124164862e-299

The vehicle type and rollover status are not independent, since the p-value of the chi-square test is less than 0.05. The vehicle type will be used as a feature variable when predicting the rollover status.

2.2.7 Vehicle Dataset: Model Year

The MY in the vehicle dataset was selected, since NHTSA began to require the ESC installation starting in MY 2014 (see Section 1.2). The MY was grouped into three categories: MY 1989-

2007, MY 2008-2010, and MY 2011-2019. The following contingency table shows the MY group by rollover status.

Table 7: Contingency Table of Model Year Group and Rollover Status

	Rollover	Non-Rollover
MY 1989-2007	17548 (40.73%)	25539 (59.27%)
MY 2008-2010	1859 (28.15%)	4746 (71.85%)
MY 2011-2019	2413 (19.86%)	9736 (80.14%)

17,548 passenger vehicles manufactured between 1989 and 2007 involved rollovers in single-vehicle crashes. The rate of rollovers experienced by the passenger vehicles manufactured between 1989 and 2007 is 40.73 percent while the rate of rollovers experienced by the passenger vehicles manufactured between 2011 and 2019 is 19.86 percent. The rate of rollovers decreases when the MY increases.

The chi-square test was used to examine the significant association between the MY group and rollover status. The following hypothesis statements were used.

H_0: The MY group and rollover status are independent
H_1: The MY group and rollover status are not independent

The chi-square test was based on the data in Table 7. The following table shows the chi-square statistic and its p-value.

Chi-square statistic	P-value
1971.8730	0.0

The MY group and rollover status are not independent, since the p-value of the chi-square test is less than 0.05. The MY group will be used as a feature variable when predicting the rollover status.

2.2.8 Person Dataset: Driver's Gender

The driver's gender in the person dataset was selected, since the driving behaviors of males and females are different. The driver's gender is a categorical variable with two categories, male and female drivers. The following contingency table shows the driver's gender by rollover status.

Table 8: Contingency Table of Driver's Gender and Rollover Status

	Rollover	Non-Rollover
Female	5731 (33.53%)	11363 (66.47%)
Male	16089 (35.96%)	28658 (64.04%)

There are 5,731 female drivers that experienced rollovers in single-passenger-vehicle crashes. The rate of rollovers experienced by the female drivers is 33.53 percent while the rate of

rollovers experienced by the male drivers is 35.96 percent. The male drivers are more likely to experience rollovers than the female drivers.

The chi-square test was used to examine the significant association between the driver's gender and rollover status. The following hypothesis statements were used.

H_0: The driver's gender and rollover status are independent
H_1: The driver's gender and rollover status are not independent

The chi-square test was based on the data in Table 8. The following table shows the chi-square statistic and its p-value.

Chi-square statistic	P-value
31.8556	1.6607267637296446e-08

The driver's gender and rollover status are not independent, since the p-value of the chi-square test is less than 0.05. The driver's gender will be used as a feature variable when predicting the rollover status.

2.2.9 Person Dataset: Driver's Age

The driver's age in the person dataset was selected, since the driving behaviors are related to the driver's age. The driver's age is a numerical variable. The following table shows the mean and standard deviation of the driver's age in rollover and non-rollover crashes.

	Rollover	Non-Rollover
Mean of Driver's Age (year)	38.46	41.69
Standard Deviation of Driver's Age (year)	17.44	18.70

The average age of drivers that experienced rollovers in single-passenger-vehicle crashes is 38.46 years while the average age of drivers that experienced non-rollovers in single-passenger-vehicle crashes is 41.69 years. The likelihood of rollovers decreases when the driver's age increases.

The one-way ANOVA test was used to examine the significant difference between the average age of drivers that experienced rollovers and the average age of drivers that experienced non-rollovers.

H_0: The means of driver's age in rollovers and non-rollovers are the same
H_1: The means of driver's age in rollovers and non-rollovers are different

The following table shows the result of one-way ANOVA test.

	Sum of Square	Degree of Freedom	F-Statistic	P-value
ROLL	1.466400e+05	1	439.6108	2.873407e-97

Residual	2.062749e+07	61839		

The means of driver's age in rollovers and non-rollovers are significantly different, since the p-value of the one-way ANOVA test is less than 0.05. The driver's age will be used as a feature variable when predicting the rollover status.

3 Missing Data

Missing values exist in the target population. The following table shows the missing rate of each feature variable.

	Missing Rate
Weather Condition	4.8%
Light Condition	0.7%
Roadway Surface	1.2%
Roadway Grade	7.6%
Roadway Alignment	2.3%
Vehicle Type	2.3%
Model Year	3.9%
Driver's Gender	3.7%
Driver's Age	4.7%

This project assumed that the missing values are missing completely at random and did not apply any imputations. The algorithms in the later sections will not be seriously affected by the missing values because of the following reasons.

- The missing rates are around 5 percent, and such missing rates are tolerable.
- The target population is large. There are still 61,841 data observations in the analytical dataset after removing the missing values.

4 Algorithms

This project applied four algorithms to predict the rollover status in single-passenger-vehicle crashes. The algorithms included the decision tree, random forest, logistic regression, and KNN. In each algorithm, 70 percent of the analytical dataset was used as the training dataset, and 30 percent of the analytical dataset was used as the testing dataset.

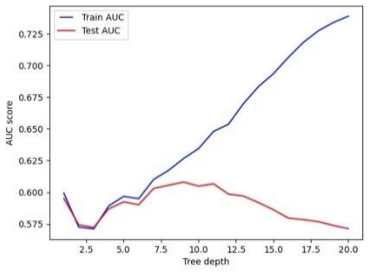
This project also used 10-fold cross validation to generate a sequence of accuracy scores in each algorithm. The statistical test will be used to compare the prediction performances of different algorithms.

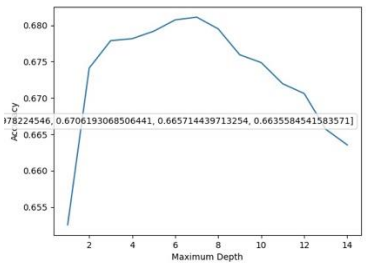
4.1 Decision Tree

We utilized the Decision Tree algorithm as one of our training models. The advantages of this model is that it treats all features as independent events, is a classification model, and is easy to

comprehend. We determined the optimal depth of our decision tree (max_depth = 3) after we saw a significant over-fitting error for depths > 5.

The most important features in our decision tree was the “My_Group”, “Align”, and “Road” condition variables.





	feature	importance
5	ALIGN_GROUP	0.476
8	MY_GROUP	0.325
4	VEHICLE_TYPE	0.188
1	LIGHT_GROUP	0.011
0	WEATHER_GROUP	0.000
2	SURFACE_GROUP	0.000
3	GRADE_GROUP	0.000
6	SEX	0.000
7	AGE	0.000

4.2 Random Forest

We also utilized Random Forest modeling to build from the success of our Decision Tree modeling. Our random forest yielded an accuracy of 66%, F1 score of 65%.

Given that we had a minority class (Roll-Over), we used SMOTE algorithm to randomly over sample the minority class. This resulted in an accuracy of 60%, noticeable less than our training dataset. However, the accuracy hovers around 67% for our other models because we have approximately 67% non-rollover events. Therefore, our model that has 60% accuracy with 50:50 class representation is a marked improvement. In future work, we hope to explore alternative over-sampling techniques to improve our model predictions.

4.3 Logistic Regression

This project used the logit link function to build the logistic regression, since the target variable is binary (rollover and non-rollover). This project applied the Maximum Likelihood Estimation (MLE) to estimate parameters in the logistic regression.

The categorical variables in the analytical dataset were presented by using dummy variables. For example, this project used the following dummy variable to present the driver’s gender.

$$\text{GENDER\_FEMALE} = \begin{cases} 1, & \text{if a female driver} \\ 0, & \text{otherwise} \end{cases}$$

The following table shows the output of the logistic regression.

Dep. Variable:	y	No. Observations:	43288
Model:	Logit	Df Residuals:	43271
Method:	MLE	Df Model:	16

```
Date:                Thu, 29 Apr 2021   Pseudo R-squ.:                 0.08417
Time:                        11:35:19   Log-Likelihood:                -25751.
converged:                       True   LL-Null:                       -28118.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
                                 coef    std err          z      P>|z|
--------------------------------------------------------------------------------
AGE                           -0.0089      0.001    -14.666      0.000
WEATHER_CLEAR/NORMAL           1.3326      0.131     10.167      0.000
WEATHER_FOG/CLOUDY             1.3180      0.132     10.000      0.000
WEATHER_RAIN/SLEET             1.1686      0.139      8.432      0.000
WEATHER_SNOW                   2.0391      0.160     12.763      0.000
LIGHT_DAWN/DUSK               -0.1078      0.054     -1.999      0.046
LIGHT_LIGHT                   -0.2771      0.022    -12.332      0.000
SURFACE_DRY                    0.0421      0.127      0.330      0.741
SURFACE_WET                   -0.1414      0.131     -1.081      0.280
GRADE_LEVEL                   -0.3467      0.024    -14.325      0.000
ALIGN_STRAIGHT                -0.8576      0.023    -36.611      0.000
VEHICLE_CAR                   -0.6851      0.027    -25.467      0.000
VEHICLE_PICKUP                -0.1415      0.030     -4.641      0.000
VEHICLE_VAN                   -0.7106      0.056    -12.692      0.000
MY_2008-2010                  -0.4310      0.036    -11.858      0.000
MY_2011-2019                  -0.8774      0.031    -28.756      0.000
GENDER_FEMALE                  0.0271      0.025      1.103      0.270
```

The overall logistic regression is significant, since the LLR p-value is less than 0.05. However, the driver's gender is not significant, since the p-value of GENDER_FEMALE (0.270) is greater than 0.05. The driver's gender should be removed from the logistic regression.

The p-values of SURFACE_DRY (0.741) and SURFACE_WET (0.280) are also greater than 0.05. These two p-values might be affected by the reference group, since the size of the reference group (SURFACE_OIL) is relatively smaller than the SURFACE_DRY and SURFACE_WET. The oil roadway surface should be combined with the wet roadway surface.

The following table shows the output of the modified logistic regression.

```
Dep. Variable:                      y   No. Observations:                43288
Model:                          Logit   Df Residuals:                    43273
Method:                           MLE   Df Model:                           14
Date:                Mon, 03 May 2021   Pseudo R-squ.:                0.08397
Time:                        12:06:56   Log-Likelihood:               -25757.
converged:                       True   LL-Null:                      -28118.
Covariance Type:            nonrobust   LLR p-value:                    0.000
================================================================================
                               coef    std err          z      P>|z|
--------------------------------------------------------------------------------
AGE                           -0.0089      0.001    -14.645      0.000
WEATHER_CLEAR/NORMAL           1.3785      0.041     33.303      0.000
WEATHER_FOG/CLOUDY             1.3486      0.047     28.550      0.000
WEATHER_RAIN/SLEET             1.1421      0.070     16.408      0.000
WEATHER_SNOW                   2.0214      0.113     17.823      0.000
LIGHT_DAWN/DUSK               -0.1082      0.054     -2.008      0.045
LIGHT_LIGHT                   -0.2751      0.022    -12.266      0.000
SURFACE_WET/OIL               -0.1079      0.044     -2.448      0.014
GRADE_LEVEL                   -0.3466      0.024    -14.324      0.000
ALIGN_STRAIGHT                -0.8557      0.023    -36.560      0.000
VEHICLE_CAR                   -0.6857      0.027    -25.516      0.000
VEHICLE_PICKUP                -0.1478      0.030     -4.934      0.000
VEHICLE_VAN                   -0.7134      0.056    -12.752      0.000
MY_2008-2019                  -0.4298      0.036    -11.831      0.000
MY_2011-2019                  -0.8760      0.031    -28.722      0.000
```

The overall logistic regression is significant, since the LLR p-value is less than 0.05. The p-values of the feature variables are less than 0.05, and this is a valid logistic regression model.

Based on the feature variables in the above logistic regression, this project applied 10-fold cross validation and yield the following ten accuracy score.

1st fold	2nd fold	3rd fold	4th fold	5th fold	6th fold	7th fold	8th fold	9th fold	10th fold
0.6750	0.6768	0.6791	0.6787	0.6863	0.6791	0.6824	0.6808	0.6733	0.6973

The average accuracy score of the logistic regression is 0.6809.

4.4 KNN

This project applied the KNN algorithm to predict the rollover status. Based on the numerical trials and accuracy score comparisons, the size of the neighborhood is set at 40. After size 40, the accuracy score was not improved by increasing size of the neighborhood.

The KNN algorithm is a distance-based algorithm, and the categorical variables in the analytical dataset were labeled by artificial distance. For example, the weather condition was presented by the following artificial distance.

$$\text{Weather} = \begin{cases} 0, & \text{clear/normal} \\ 1, & \text{FOG/CLOUDY} \\ 2, & \text{rain/sleet} \\ 3, & \text{snow} \\ 4, & \text{windy} \end{cases}$$

Based on the KNN algorithm, this project applied 10-fold cross validation and yield the following ten accuracy score.

1st fold	2nd fold	3rd fold	4th fold	5th fold	6th fold	7th fold	8th fold	9th fold	10th fold
0.6738	0.6713	0.6787	0.6734	0.6863	0.6699	0.6741	0.6715	0.6668	0.6874

The average accuracy score of the logistic regression is 0.6753.

5 Algorithm Comparison

The 10-fold cross validation yielded ten accuracy scores in each algorithm. This project could not examine whether the accuracy score follows a normal distribution, since the sample size of the accuracy score (10) is less than 30. As the result, this project used a non-parametric statistic test, Wilcoxon signed-rank test to examine the significant difference among the algorithm prediction performance.

The Wilcoxon signed-rank test can only examine two algorithms at one time. The following sections show the test results.

5.1 Decision Tree vs. Random Forest

The following table summarizes the accuracy scores of the decision tree (see Section 4.1) and random forest (see Section 4.2).

Decision Tree									
1st fold	2nd fold	3rd fold	4th fold	5th fold	6th fold	7th fold	8th fold	9th fold	10th fold
0.6417	0.6299	0.6392	0.6156	0.6560	0.6396	0.6415	0.6475	0.6428	0.6456
Mean: 0.6399									
Random Forest									
1st fold	2nd fold	3rd fold	4th fold	5th fold	6th fold	7th fold	8th fold	9th fold	10th fold
0.6711	0.6655	0.6713	0.6713	0.6801	0.6764	0.6752	0.6738	0.6703	0.6874
Mean: 0.6742									

The mean of accuracy scores yield by the random forest (0.6742) is greater than the mean of accuracy scores yield by the decision tree (0.6399). The following hypothesis statements were used.

H_0: Two sets of scores follow the same distribution
H_1: Two sets of scores do not follow the same distribution

The following table shows the analysis result.

Test Statistic	P-value
0.0	0.001953125

The prediction performance of random forest is significantly better than the prediction performance of decision tree, since the p-value is less than 0.05.

5.2 Logistic Regression vs. Random Forest

The following table summarizes the accuracy scores of the random forest (see Section 4.2) and logistic regression (see Section 4.3).

Random Forest									
1st fold	2nd fold	3rd fold	4th fold	5th fold	6th fold	7th fold	8th fold	9th fold	10th fold
0.6711	0.6655	0.6713	0.6713	0.6801	0.6764	0.6752	0.6738	0.6703	0.6874
Mean: 0.6742									
Logistic Regression									
1st fold	2nd fold	3rd fold	4th fold	5th fold	6th fold	7th fold	8th fold	9th fold	10th fold
0.6750	0.6768	0.6791	0.6787	0.6863	0.6791	0.6824	0.6808	0.6733	0.6973
Mean: 0.6809									

The mean of accuracy scores yield by the random forest (0.6742) is less than the mean of accuracy scores yield by the logistic regression (0.6809). The following hypothesis statements were used.

H_0: Two sets of scores follow the same distribution
H_1: Two sets of scores do not follow the same distribution

The following table shows the analysis result.

Test Statistic	P-value
0.0	0.001953125

The prediction performance of logistic regression is significantly better than the prediction performance of random forest, since the p-value is less than 0.05.

5.3 Logistic Regression vs. KNN

The following table summarizes the accuracy scores of the logistic regression (see Section 4.3) and KNN (see Section 4.4).

Logistic Regression									
1st fold	2nd fold	3rd fold	4th fold	5th fold	6th fold	7th fold	8th fold	9th fold	10th fold
0.6750	0.6768	0.6791	0.6787	0.6863	0.6791	0.6824	0.6808	0.6733	0.6973
Mean: 0.6809									
KNN									
1st fold	2nd fold	3rd fold	4th fold	5th fold	6th fold	7th fold	8th fold	9th fold	10th fold
0.6738	0.6713	0.6787	0.6734	0.6863	0.6699	0.6741	0.6715	0.6668	0.6874
Mean: 0.6753									

The mean of accuracy scores yield by the logistic regression (0.6809) is greater than the mean of accuracy scores yield by the KNN (0.6753). The following hypothesis statements were used.

H_0: Two sets of scores follow the same distribution

H_1: Two sets of scores do not follow the same distribution

The following table shows the analysis result.

Test Statistic	P-value
20.0	0.4921875

There is no significant difference between the prediction performance of logistic regression and the prediction performance of KNN, since the p-value is greater than 0.05. However, the logistic regression used eight feature variables while the KNN used nine feature variables.

As the conclusion, the logistic regression provided the best prediction performance than the decision tree and random forest. Comparing with KNN, the logistic regression achieved the same prediction performance with fewer feature variables.

6. GUI Design

We used PyQt5 for this GUI design and code, and Qt Designer wasn't used.

6.1 Overall Layout

The GUI is comprised of a main window and four dropdown menus:

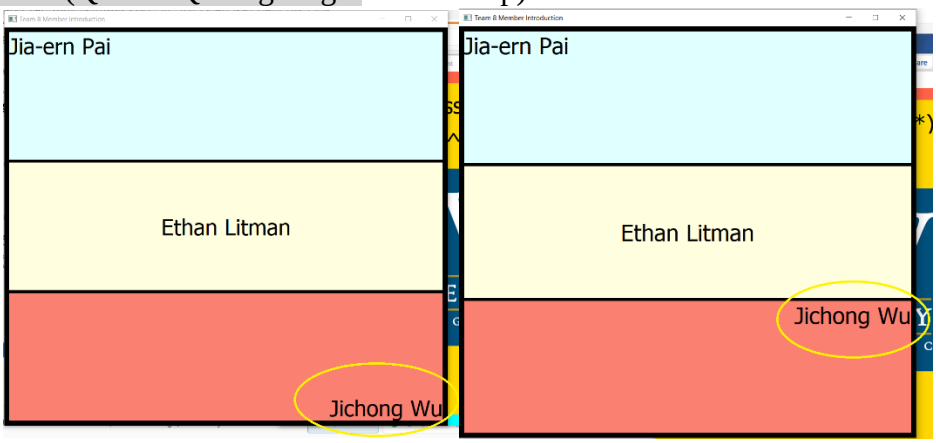
- **About** – introduce the team, the professor of this class, and the Exit button
- **EDA** – display basic analysis and steps during data preprocessing and cleaning
- **Models** – display model metrics, results with charts by four different models we built, including Decision Tree, Random Forest, Logistic Regression and KNN
- **Conclusion** – display the accuracy scores of each model and note that we didn't make conclusion only based on the accuracy score, this is more for a good way of GUI to present virtually.

6.2 The main window design

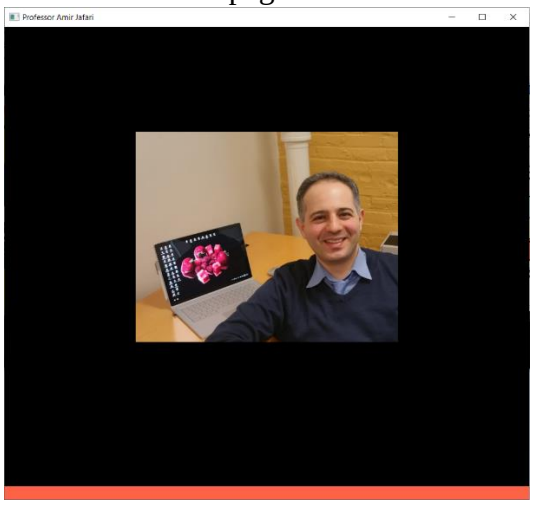
1. **Background**: comprised of some opening text messages (by QLabel), customized by setStyleSheet() function for text size and color. Also used the QPixmap() function of loading a photo to the window which was the main challenge for this part. The issue was how to load a photo from the internet instead of local machine. The urllib.request.urlopen().read() did the job.
2. **Menu bar**: background color was consumed.
3. **Status tip bar**: background color and messages displayed in this section was customized.

6.3 The "About" menu

1. **The "Team 9 Member Introduction" tab**: used QLabel to display team member names, adjusted the label size and background color; adjust the text of team member names to make it bold and big, also did some research on how to align the text within each label text box to make them line up nicely (top left, center, bottom right) by using setAlignment(QtCore.Qt.AlignRight/Center/Top)



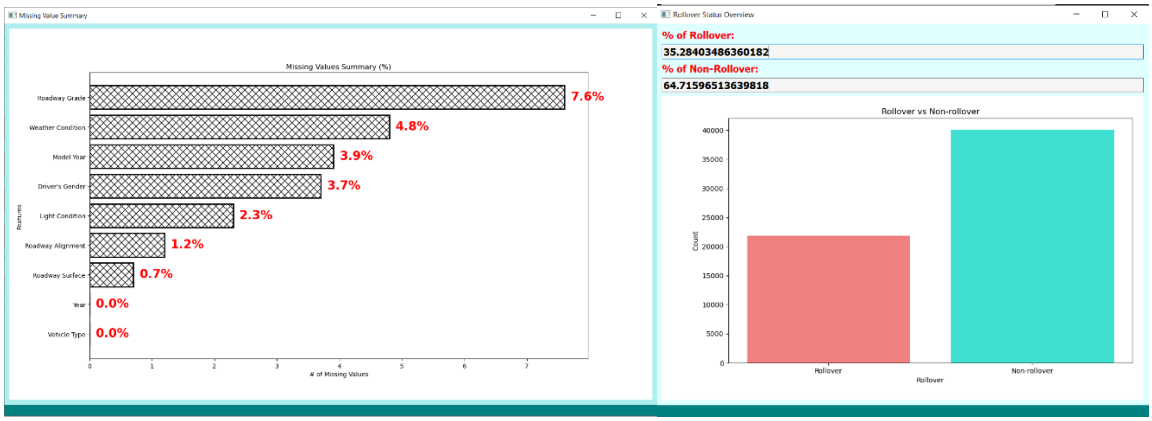
2. **The "Professor Amir Jafari" tab**: applied the QPixmap() and urllib.request.urlopen().read() technique used in the main window photo display. Photo of Professor Jafari credit goes to his GitHub page.



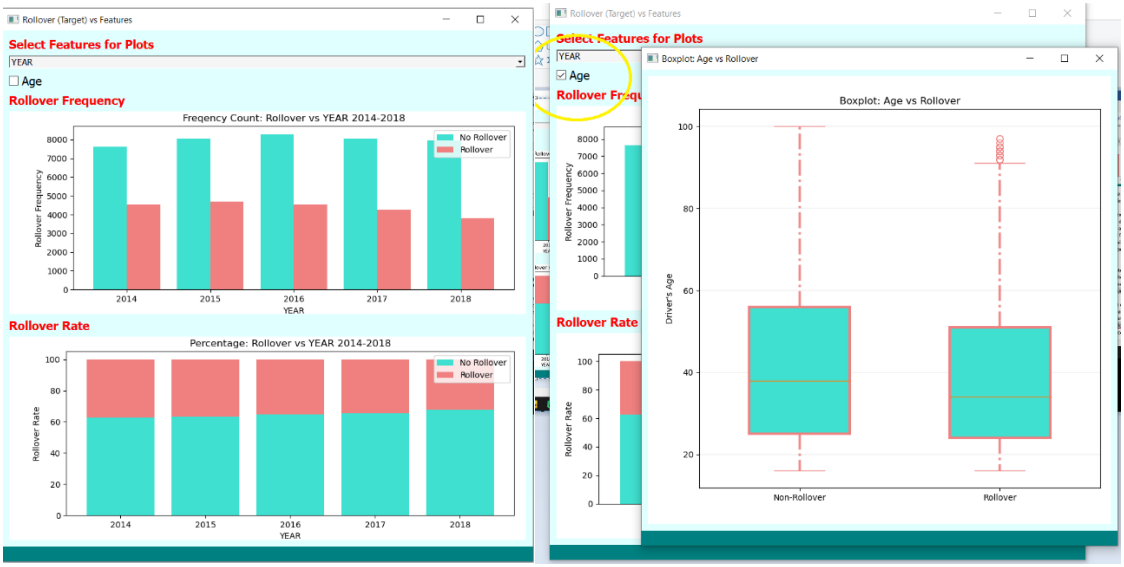
3. **The "Exit" button**: used QAction function for creating the exit button and linked with the triggered.connect() function to execute the closing window action.

6.4 The "EDA" menu

1. **The "Missing Values" tab**: use a horizontal bar plot to display, the number results for each bar are also displayed on top of each bar.

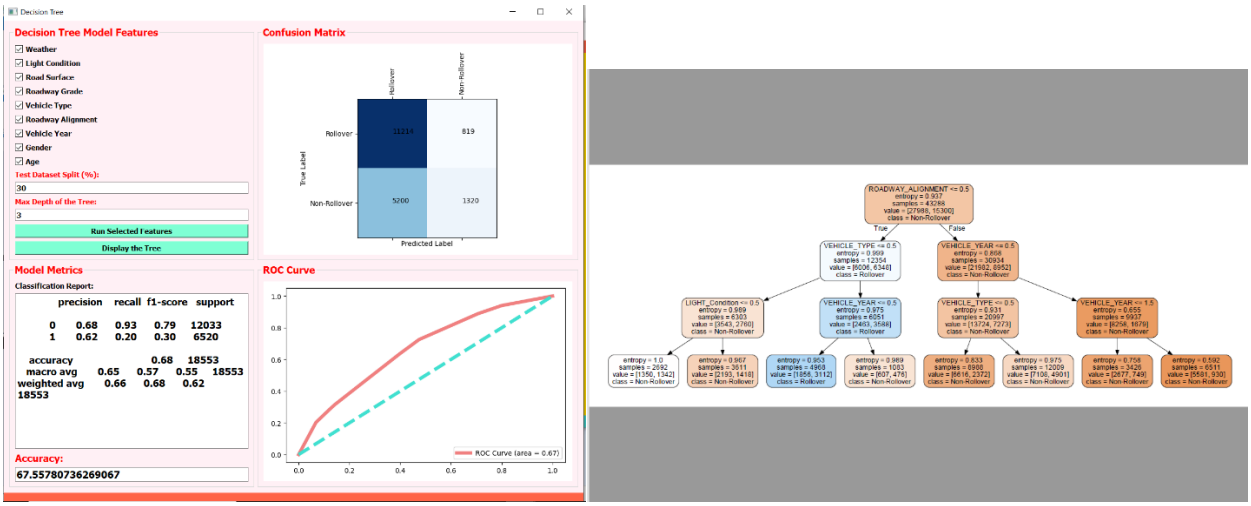


2. **The "Rollover Status Overview" tab**: displays the frequency count of the two Classes (rollover or non-rollover) in the Target variable (Rollover) by a bar chart and displays, QLabel and QLineEdit and setText() function were used to display the results.
3. **The "Rollover (Target) vs Features" tab**: displays the frequency count and percentage comparing against the Target variable. For the Age variable, we showed different techniques to treat Age as both a categorical variable (by grouping age segments) and a numerical variable. For numerical variable, we displayed not only bar plot which is not ideal for continuous data, and also created a stand alone checkbox and link to another (3rd layer) window to display a boxplot for Age vs Y variable, which is more appropriate for numerical variable.

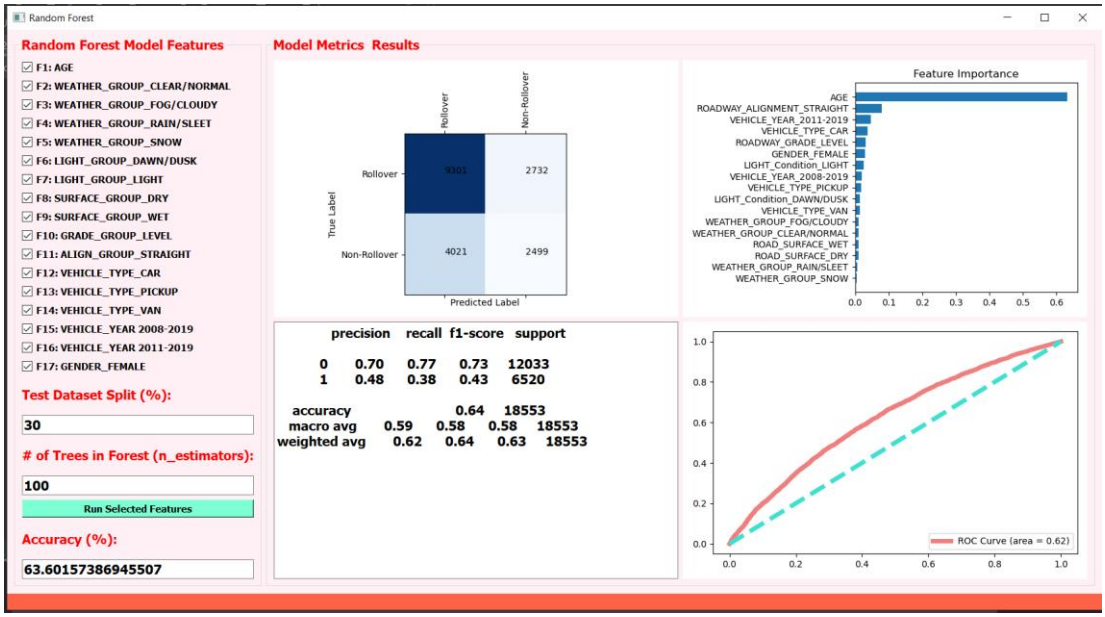


6.5 The "Models" menu

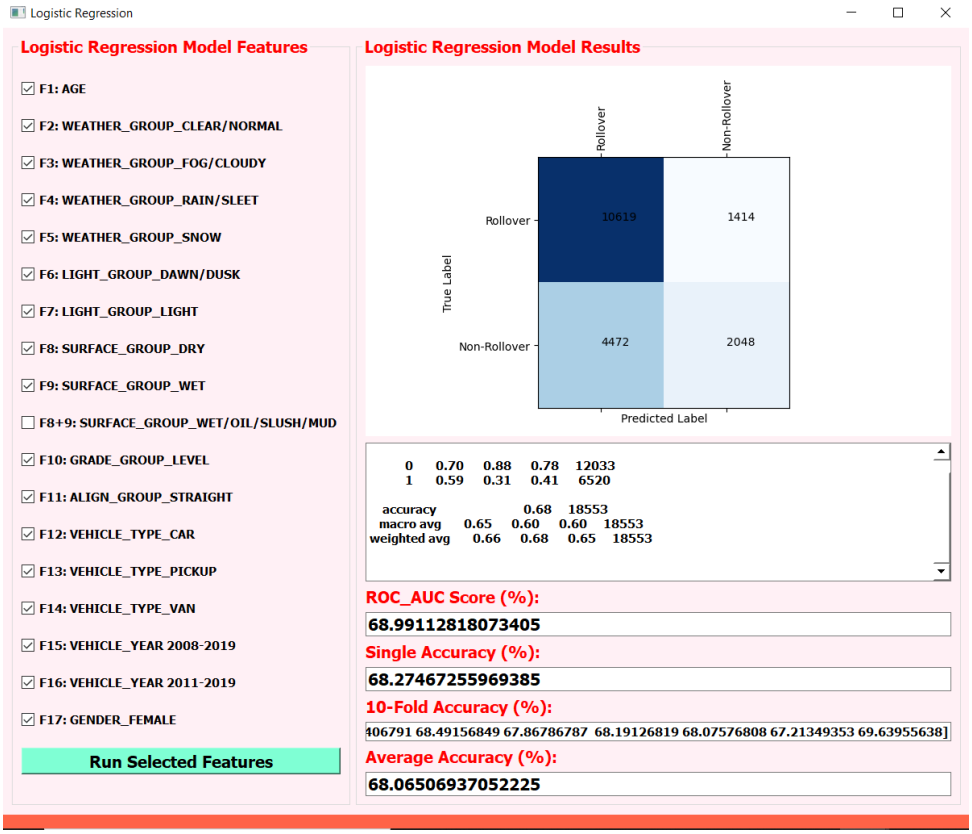
1. **Decision Tree**: layout was split into QGridlayout design which allows features and control buttons and to display on top left corner. Checkbox was used to select the features and it's linked to the display results. QPlainTextEdit() was used to create the classification report text box, and two graphics used for the confusion matrix plot and the ROC Curve. Finally a QPlainTextEdit() was used to display the Accuracy score of the model.



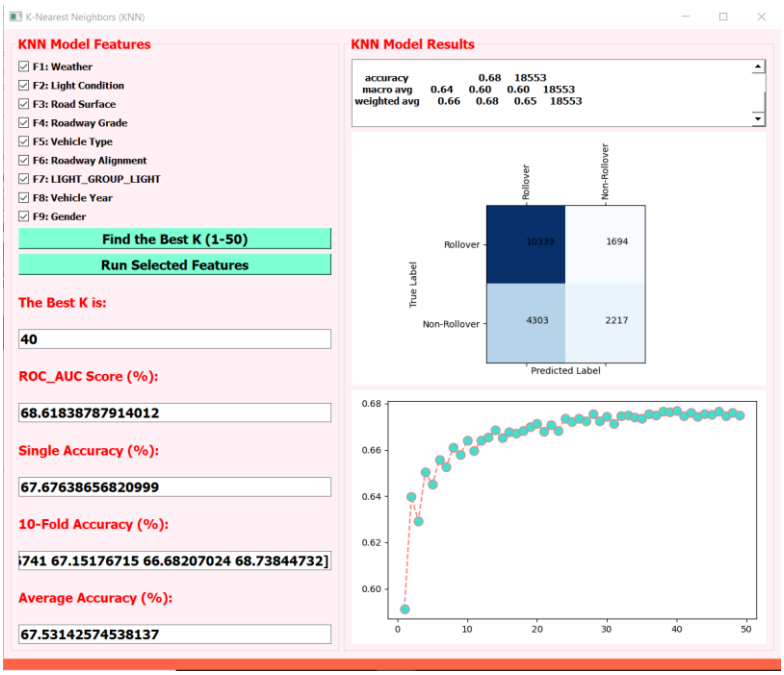
2. **Random Forest**: the layout design was similar to Decision Tree, except it used HBoxlayout and for the right area within the HBox, a QGridlayout design was used to create 4 graphics within the HBoxlayout as a second layer. The other unique thing for the Random Forest model is to take input for the n_estimators of the model. On the features, we used OneHotEncoding to split those features that have multiple labels, this is for the feature importance selection and our model will pick the first 15 important features.



3. **Logistic Regression**: same QHBoxlayout design was used, this time on the right side of the HBox, VBoxlayout was applied to accommodate the ROC_AUC score, the accuracy scores from 10-Fold cross check validation, and the average accuracy score of the 10. The ROC Curve plot was dropped because of the layout design balance, but the ROC_AUC score was included. The other consideration in GUI design for this page was after the P-value analysis on each of the selected 17 features for this model, we decided to drop GENDER_FEMALE and combine SURFACE_DRY and SURFACE_WET given the insignificant influence to the model. GUI design was adapted to make this selection feasible.



4. **KNN**: QHBoxlayout for the first layer, and QVBOXlayout for both sides as the second layer. Something different and unique of this window compared to others is the "Find the best K" button which will trigger a loop from 1-50 and plot a K value graph to show the best K value.



6.6 The "Conclusion" menu: display the ranking of the accuracy scores from 4 models we built, and the accuracy scores were the average from 10-fold validation. While we didn't evaluate our models only based on the accuracy scores, it is showing anyway from GUI's perspective for completion.