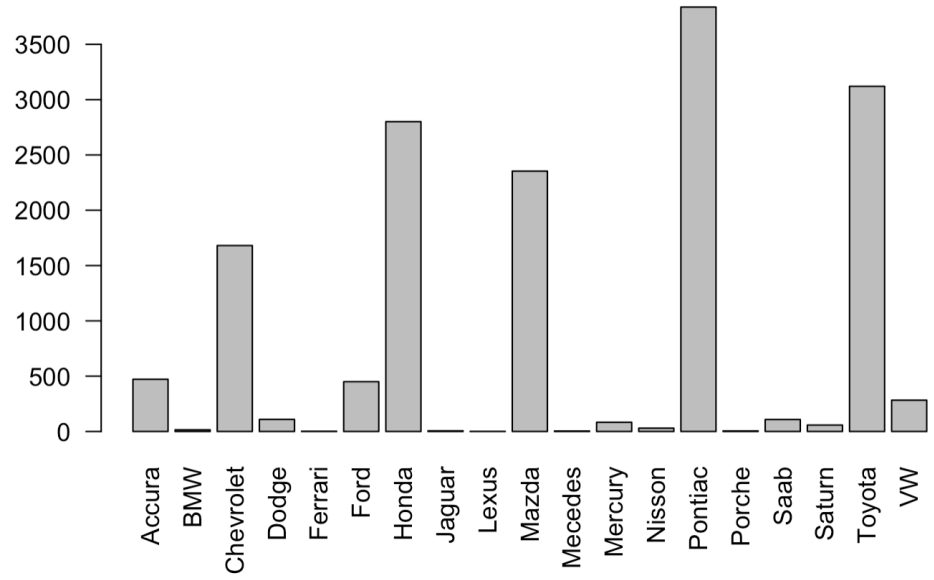
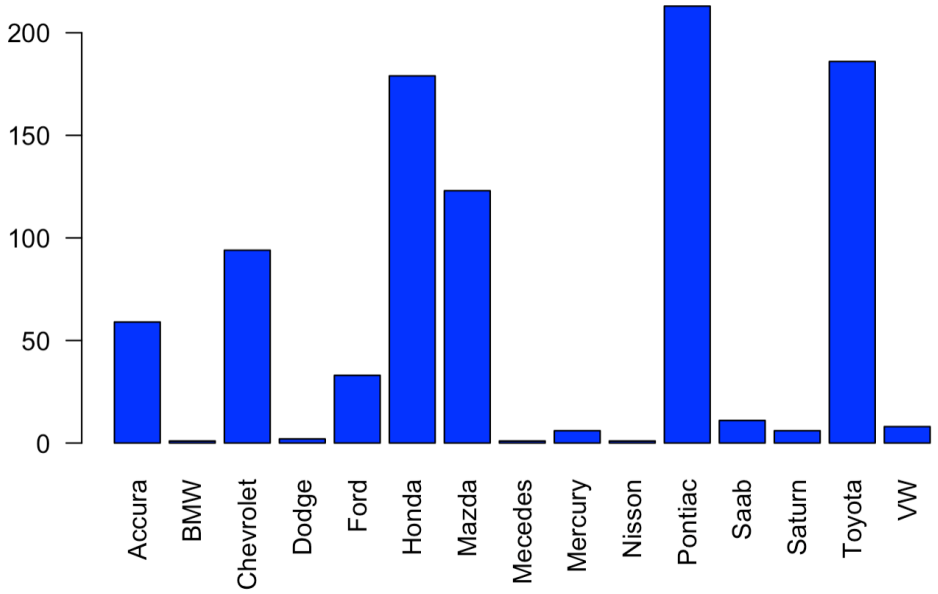
The proportions of various brands in the dataset vary. Below is a plot (created with ggplot in RStudio) that depicts the total number of cars per make – fraudulent and authentic claims including.



The total number of cars varies per brand. The number of claims per brand looks somehow proportional to the number of fraudulent claims:



One would assume that most likely the make of the car has little influence on the outcome - I.e., if a claim is a fraud or not – only by examining the proportions in the plots above. To explore further this assumption, a random forest classifier with data features scoring is created in the next section.

2.2. Random Forest Classifier

To explore the possible relationship between car brand (make) and accident report authenticity, a random forest classification model is applied to the data and the performance of the model is adjusted to its optimal configuration.

a) Initial random forest model results (75% training and 25% testing sets):

```
Call:
 randomForest(formula = FraudFound_P ~ ., data = train, ntree = 300,      importance = TRUE,
proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 300
No. of variables tried at each split: 5

        OOB estimate of  error rate: 5.96%
Confusion matrix:
      0 1  class.error
0 10868 5 0.0004598547
1   684 8 0.9884393064
```

Features importance table (20 top features):

	names <fctr>	values <dbl>
15	Fault	43.0352245
14	BasePolicy	16.0523423
11	AddressChange_Claim	13.4498487
13	PolicyType	12.7464595
9	Deductible	11.5864707
17	Year	10.6404375
6	VehicleCategory	7.7912510
16	VehiclePrice	5.2007369
26	Make	5.1502154
20	NumberOfSuppliments	3.4583539

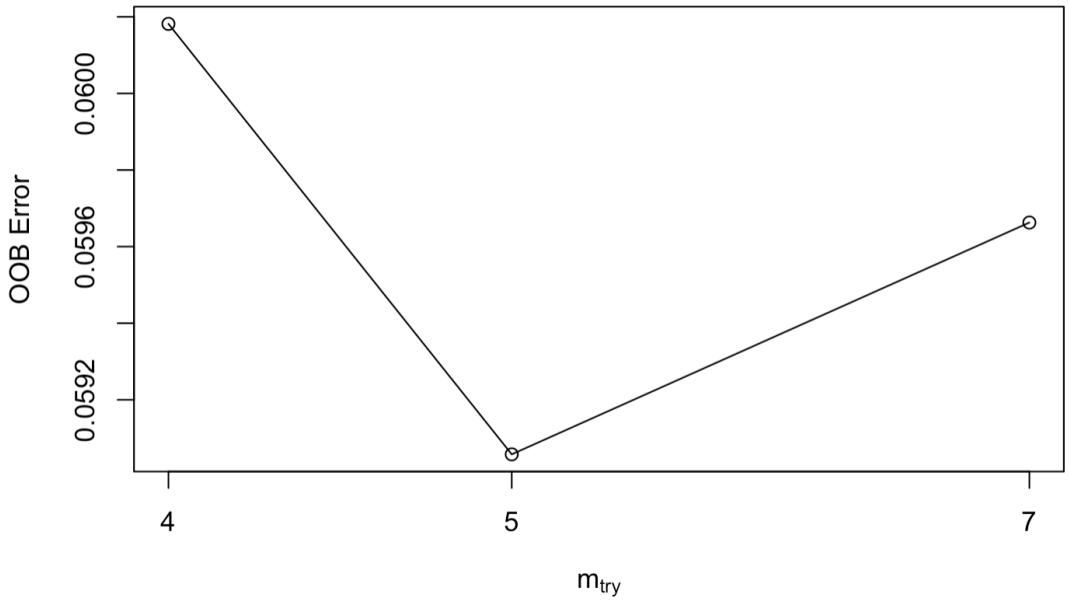
	names <fctr>	values <dbl>
24	WeekOfMonth	2.1352463
19	PastNumberOfClaims	1.9580280
3	Days_Policy_Claim	1.7347071
4	PoliceReportFiled	1.4941234
8	NumberOfCars	1.3205501
25	DayOfWeekClaimed	1.1054128
5	Days_Policy_Accident	0.7749381
10	AccidentArea	-0.3672802
2	AgentType	-0.4409303
27	DayOfWeek	-0.8314687

That the ***make of the cars has impact on the outcome 'claim is fraudulent'***, although less of an impact compared to other top variables.

b) Optimized random forest classification – estimating the best Out of Bag Erros (OOB) (RDocumentation, 2023):

As the optimization relies on random splits, unless one sets the seed in RStudio at a constant number, the final estimated OOB varies between 4 and 7. For the purpose of this report, the final best selected OOB level is at Mtry = 5.

Plot visualizing the Mtry level optimization results:



Random forest tree classifier with Mtry set at 5:

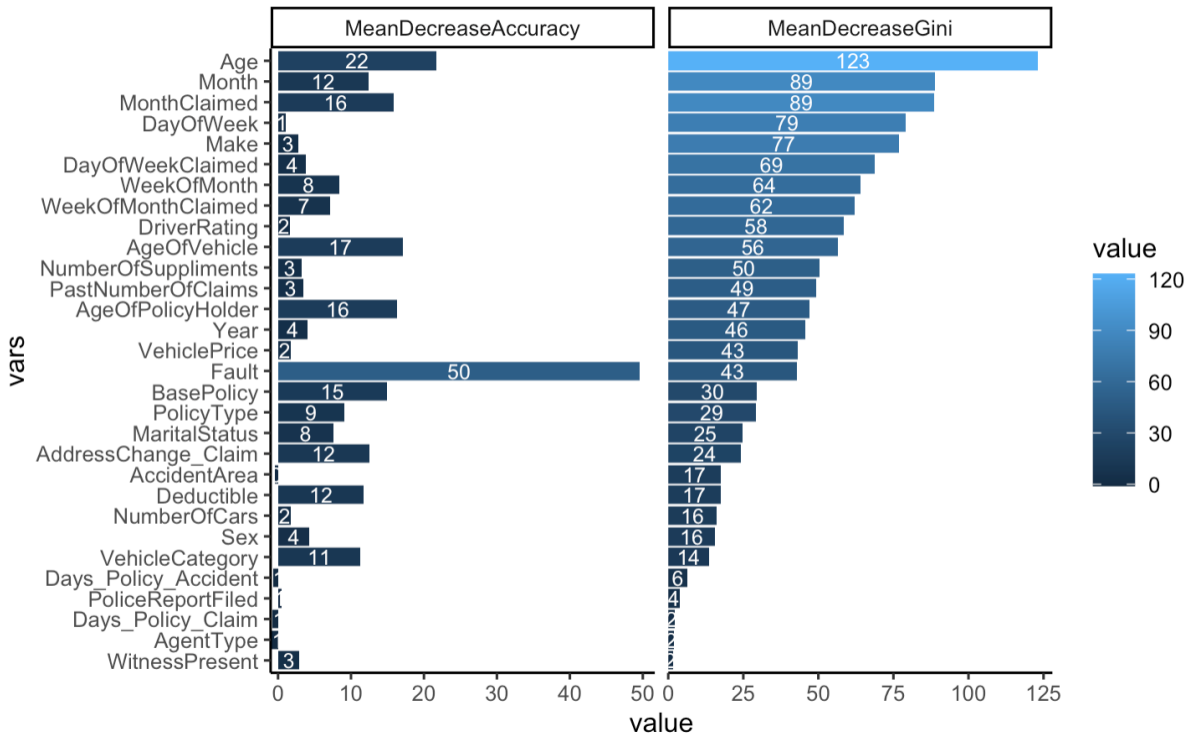
```
 randomForest(formula = FraudFound_P ~ ., data = train, ntree = 300,      mtry = 5,
importance = TRUE, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 300
No. of variables tried at each split: 5

        OOB estimate of  error rate: 5.96%
Confusion matrix:
      0 1  class.error
0 10868 5 0.0004598547
1   684 8 0.9884393064
[1] "RMSE"               "0.0591439688715954"
```

RMSE calculated with the validation set is about 6%. This implies that the model is reliable. The cars' make scores one more time in the top 10 features list:

	names <fctr>	values <dbl>
15	Fault	43.0352245
14	BasePolicy	16.0523423
11	AddressChange_Claim	13.4498487
13	PolicyType	12.7464595
9	Deductible	11.5864707
17	Year	10.6404375
6	VehicleCategory	7.7912510
16	VehiclePrice	5.2007369
26	Make	5.1502154
20	NumberOfSuppliments	3.4583539

One can conclude that the car make is an important factor in the data when it comes to the estimation of the response variable.

Plot of the Gini (purity) index for all variables:



The analysis can be extended by focusing only on mechanical factors, such as the vehicle make, category and price to explore further the relationship between those and the fraudulent claims.

2.3. Logistic Binomial Linear Model

Given that data about the customer behavior may not be available (or trustworthy) a subset of the data that can be easily extracted from the policy, such as vehicle make, model, category, year of production and purchasing cost, can be the most reliable and quick option to make an initial assessment if the accident claim is authentic. Below is an estimation of the strength and direction of influence of those factors derived from a logistic model.

Performance of Step-both vs. Step-Backward:

```{r}
summary(step.both.fit)
summary(step.backward)
```

```
Call:
glm(formula = FraudFound_P ~ VehicleCategory + VehiclePrice +
    AgeOfVehicle + Make, family = binomial(link = "logit"), data = claims_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8855  -0.4075  -0.3459  -0.1676   3.0881
```

The selected variables in the model are significant and the vehicles with the lowest p-values are listed below. However, the coefficients are negative. Car models with the highest chance for authentic claims seem to be:

- Ferrari
- Jaguar
- Lexus
- Porche

Those types of cars are also, on average ,more expensive and there are no fraudulent claims registered in the dataset with the same.

```
MakeChevrolet               -0.71495    0.17974  -3.978 6.96e-05 ***
MakeDodge                   -1.68490    0.73202  -2.302 0.021352 *
MakeFerrari                -11.85252  378.59298  -0.031 0.975025
MakeFord                    -0.54603    0.24376  -2.240 0.025088 *
MakeHonda                   -0.41566    0.17560  -2.367 0.017931 *
MakeJaguar                 -11.15136  211.83100  -0.053 0.958017
MakeLexus                  -12.11988  535.41120  -0.023 0.981940
MakeMazda                   -0.61242    0.17677  -3.465 0.000531 ***
MakeMecedes                  0.32451    1.17823   0.275 0.782992
MakeMercury                 -0.59865    0.45404  -1.318 0.187343
MakeNisson                  -1.45919    1.03120  -1.415 0.157058
MakePontiac                 -0.62842    0.16759  -3.750 0.000177 ***
MakePorche                 -11.33057  233.80298  -0.048 0.961348
MakeSaab                    -0.06460    0.35681  -0.181 0.856326
MakeSaturn                  -0.12044    0.46444  -0.259 0.795389
MakeToyota                  -0.58508    0.17194  -3.403 0.000667 ***
MakeVW                      -1.54118    0.39525  -3.899 9.65e-05 ***
```