



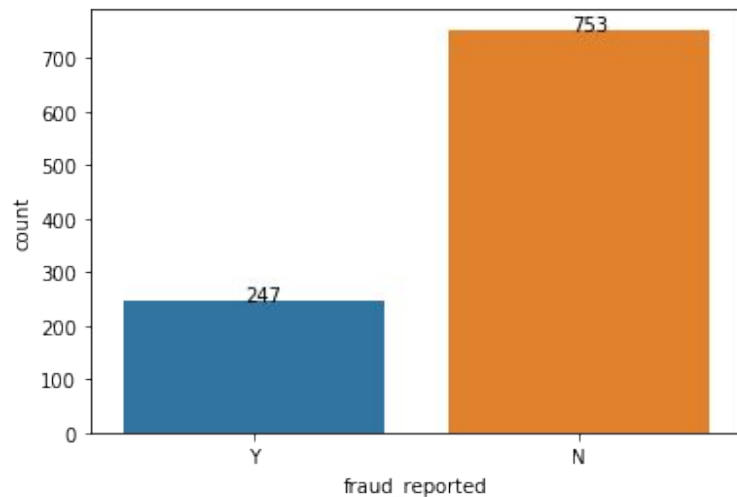
Predictive model for auto claims

Kathy Chang

Target for the task

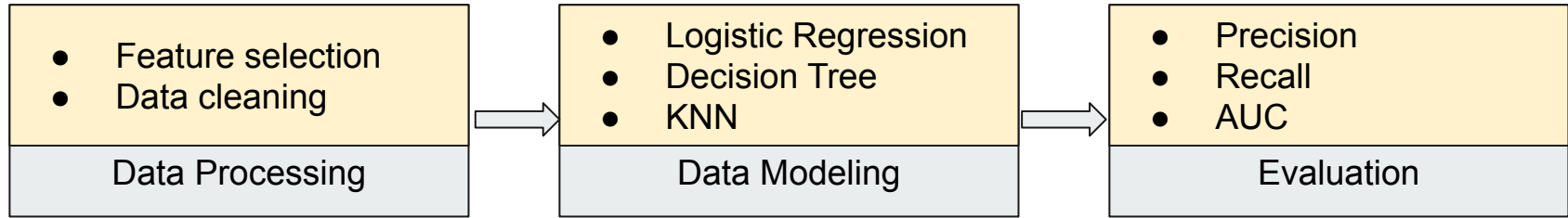
With in the data, there are 1000 cases with 40 features, 247 cases are reported as fraud versus 753 normal cases. We will detect fraud claims using classification methods to get interest insight and answer following questions!

1. Which variables are good predictors of claim amounts?
 - Total claim amount
 - Injury claim amount
 - Property claim amount
 - Vehicle claim amount
2. Which variables are good predictors of fraud as seen in the column called “fraud_reported (Y/N)”?
3. How would you assign a score from 1-100 to each claim as an indicator of the extent to which this claim is fraudulent?

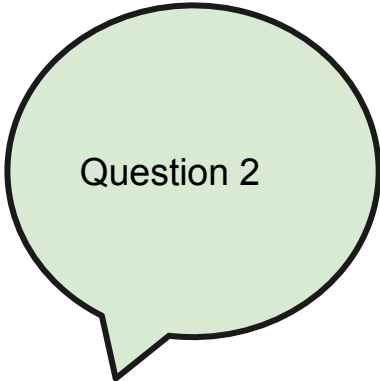




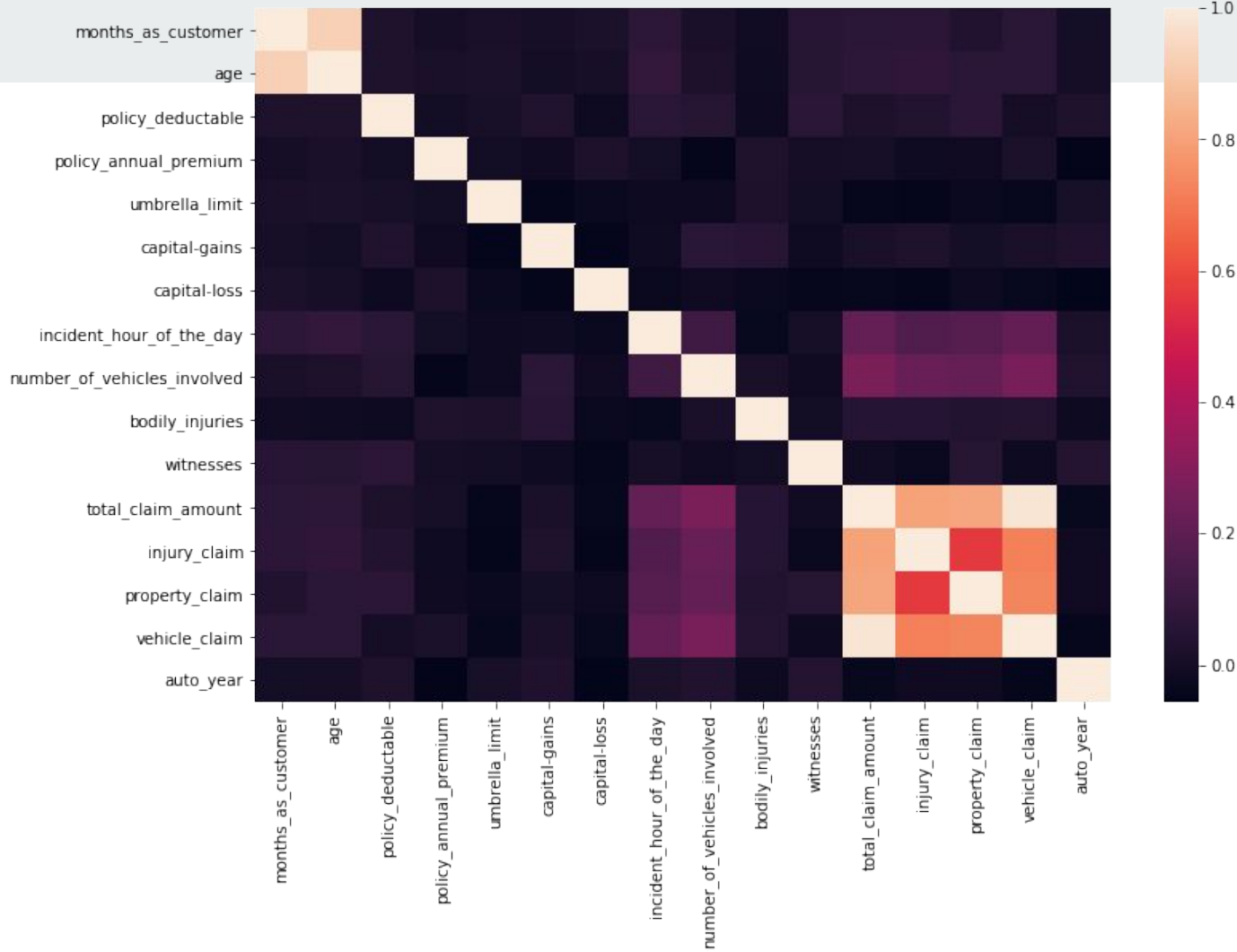
Modeling Pipeline



Question 1



Question 2





Data Processing

Data Cleaning:

- Drop column with all NaN values (_c39) and
- Drop feature with all distinct value in each row (incident location)
- Drop serial number (policy number)



Data Processing - dealing with dtype 'object'

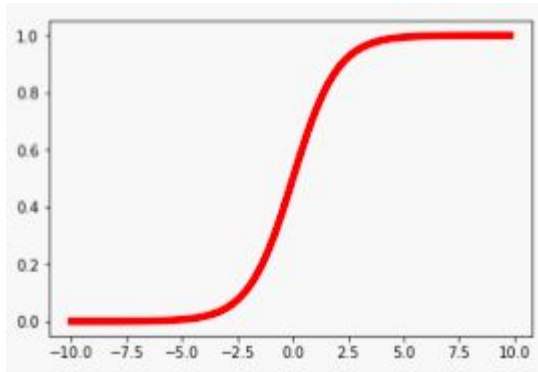
- Ordinal : adjust value according to level (education, incident_severity)
- Non-ordinal : one-hot encoding
- Binary: adjust value of binary feature to 1 vs 0 (sex, fraud_reported)

Add new feature: adjust string data with numerical value(policy_csl to csl_numerator and csl_denominator

After dealing with data types, we get 103 features in total.

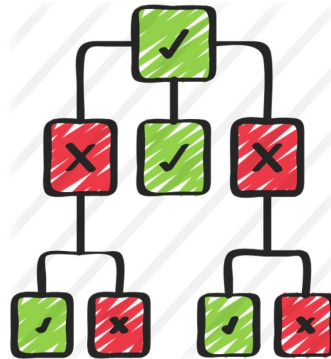
Data Modeling

Split data into training and testing dataset with proportion 75% vs 25%

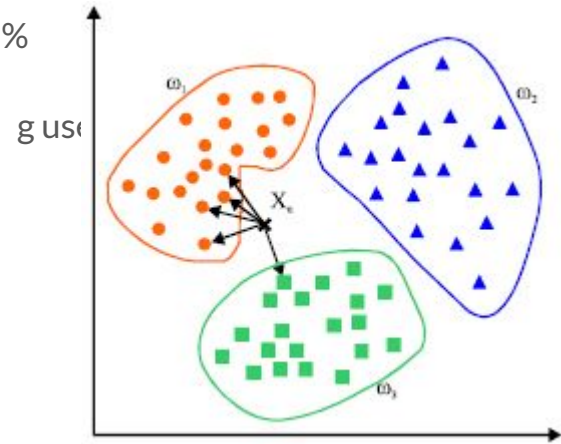


follow

Logistic Regression

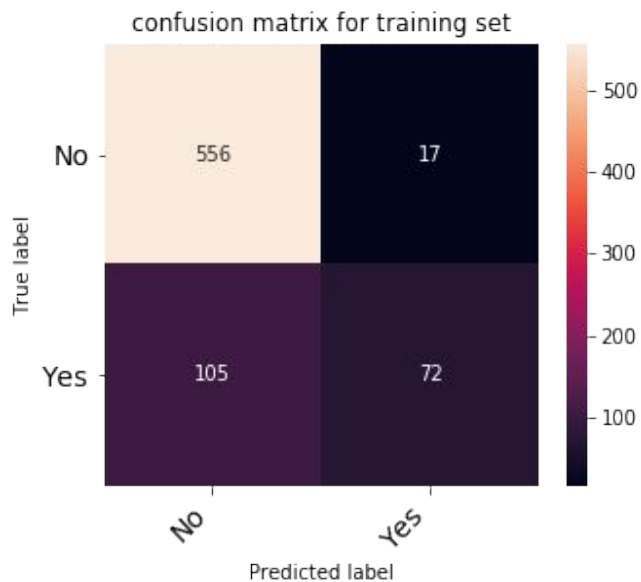


Decision Tree



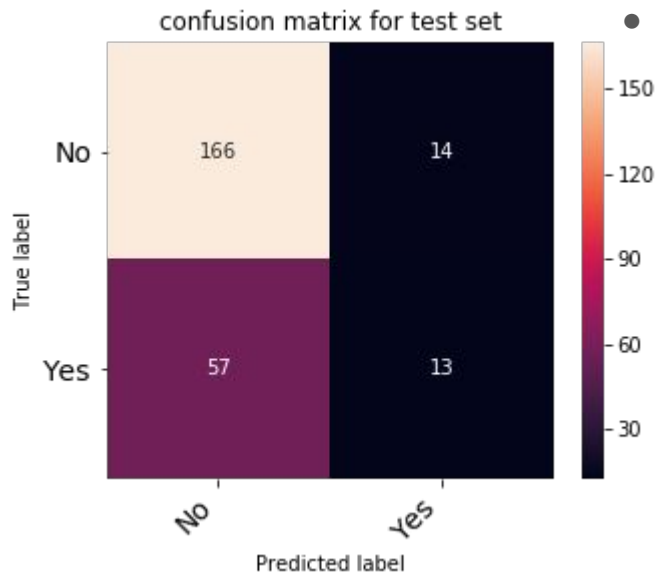
K-Nearest Neighbor

Data Modeling - Logistic Regression



Precision: 0.406780

Recall: 0.808989

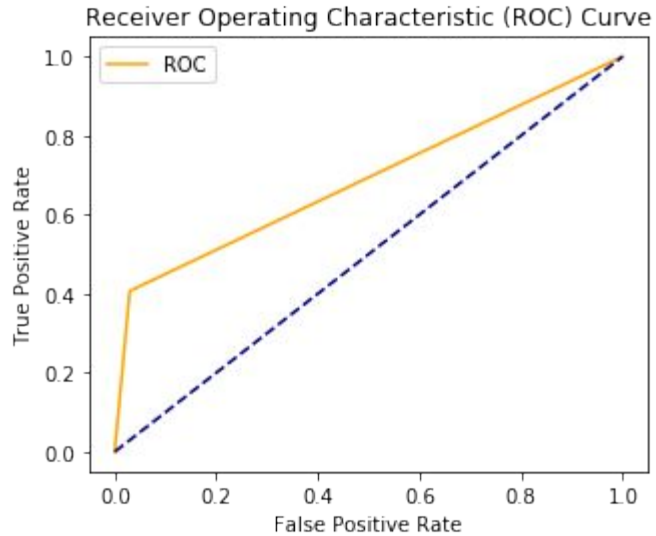


Precision: 0.185714

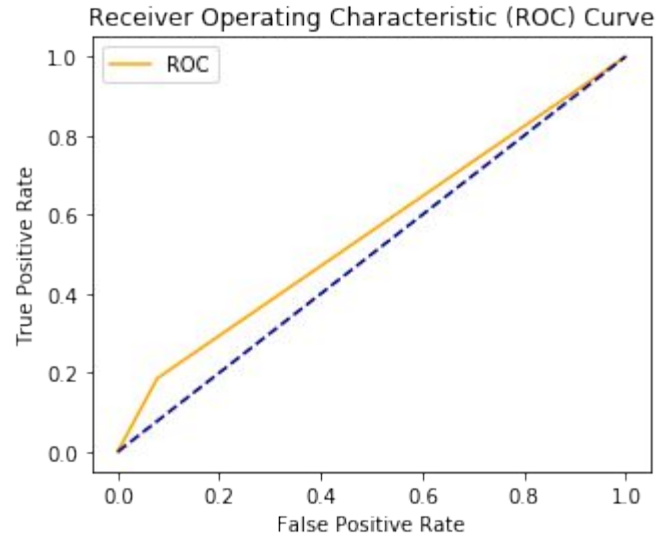
Recall: 0.481481

- Accuracy of Logistic regression classifier on training set: 0.84
- Accuracy of Logistic regression classifier on test set: 0.72

ROC-AUC - Logistic Regression

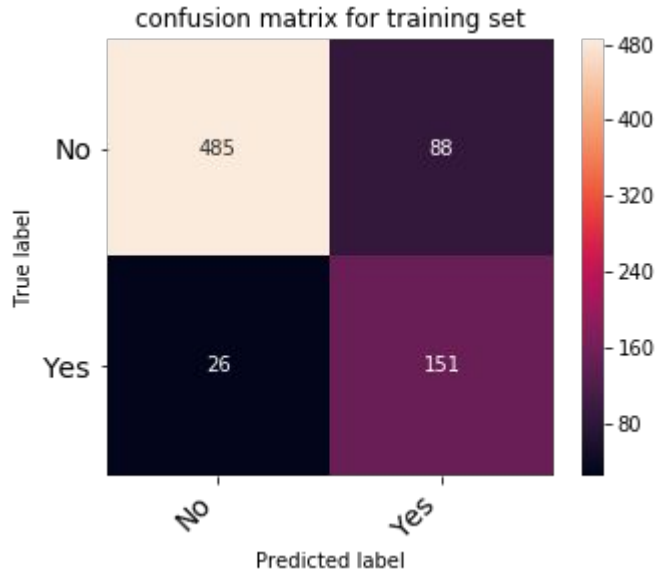


Test AUC: 0.69

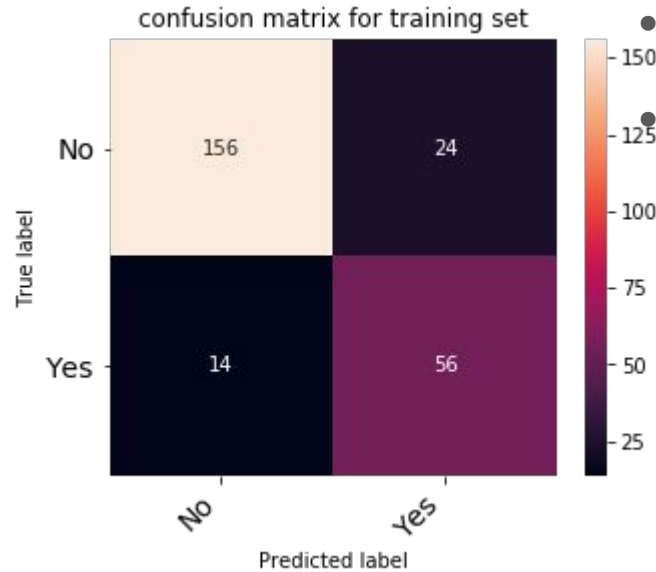


Test AUC: 0.55

Data Modeling - Decision Tree



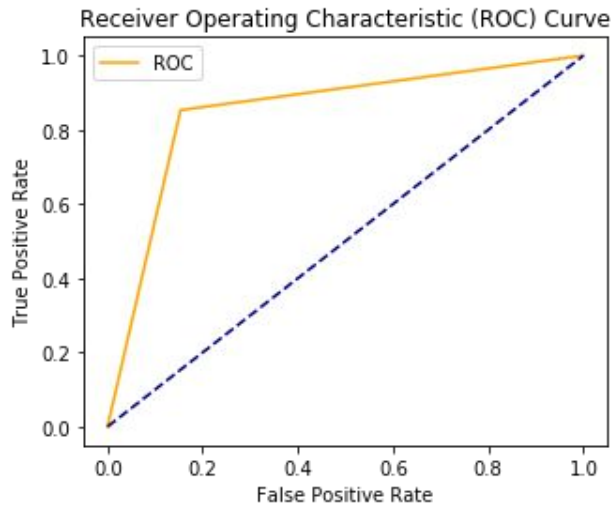
Precision: 0.853107
Recall: 0.631799



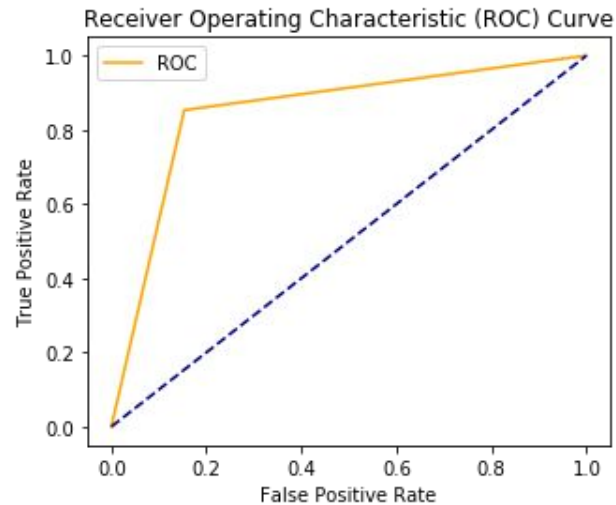
Precision: 0.8
Recall: 0.7

● Accuracy of Decision tree classifier on training set: 0.85
● Accuracy of Decision tree classifier on test set: 0.85

ROC-AUC - Decision Tree



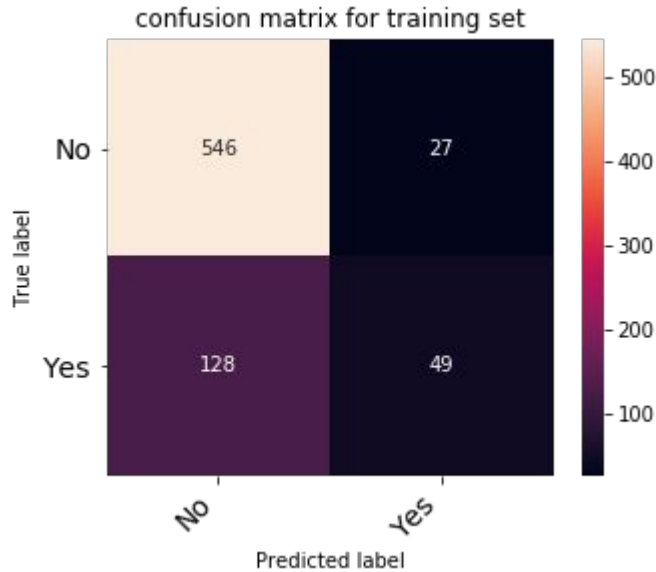
Test AUC: 0.85



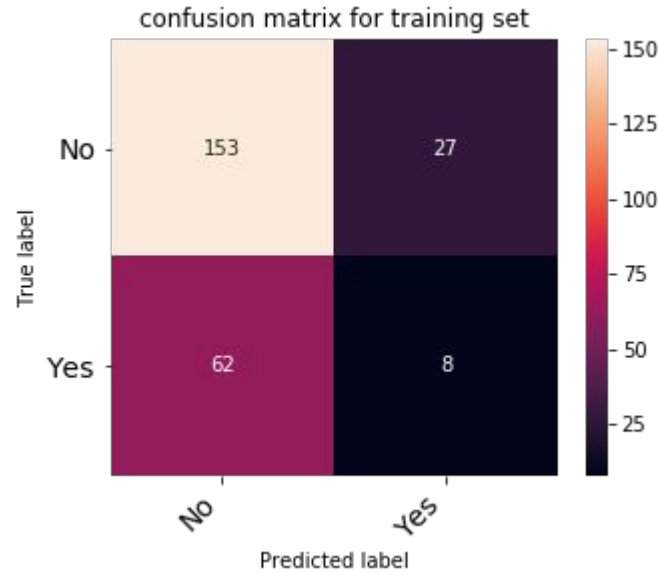
Test AUC: 0.85

Data Modeling - KNN

- Accuracy of KNN classifier on training set: 0.79
- Accuracy of KNN classifier on test set: 0.64

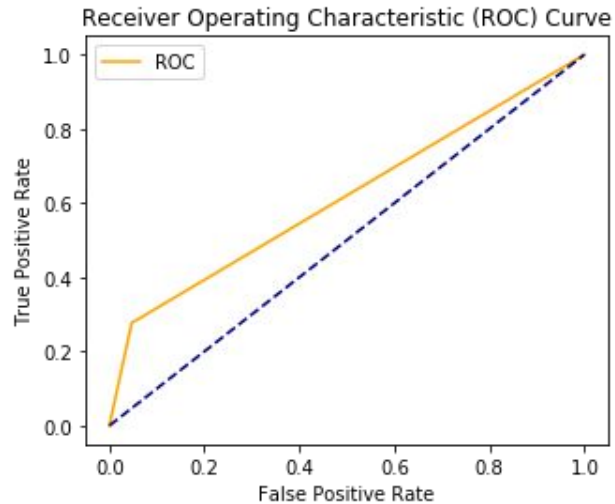


Precision: 0.276836
Recall: 0.644737

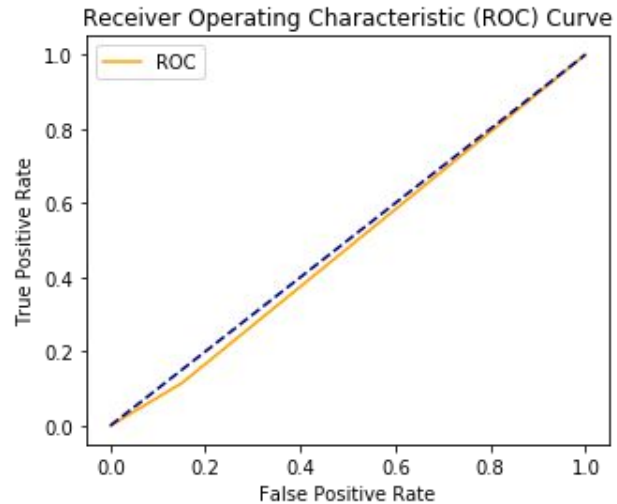


Precision: 0.114286
Recall: 0.228571

ROC-AUC - KNN



Test AUC: 0.61



Test AUC: 0.48



Performance in Training and Testing data

Train/Test	Logistic Regression		Decision Tree		KNN	
Accuracy	84%	72%	85%	85%	79%	64%
Precision	41%	19%	85%	80%	28%	11%
Recall	81%	49%	63%	70%	64%	23%
AUC	69%	55%	85%	85%	61%	48%

We can clearly see the Decision Tree outperforms than the other two models in Accuracy and Precision in both Training and test data. The AUC also explains there is 85% of correctness of Decision Tree classifier.

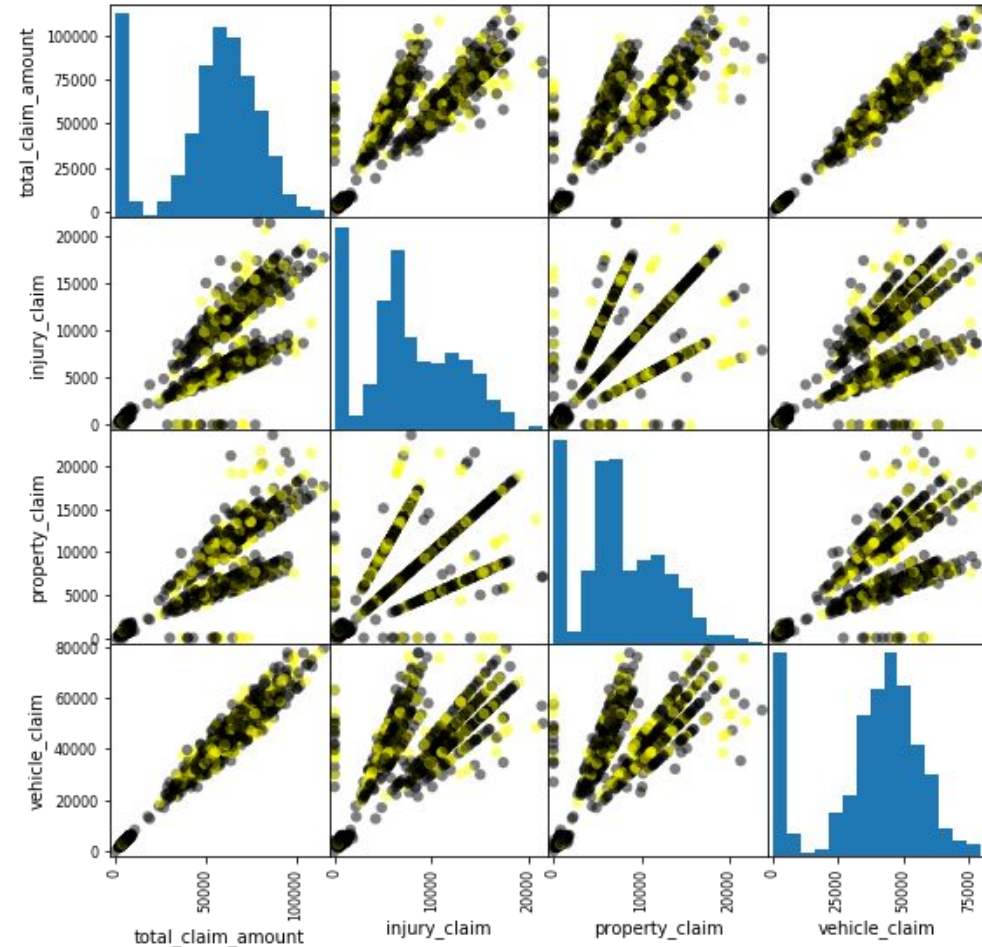
Question 1



Which variables are good predictors of claim amounts?

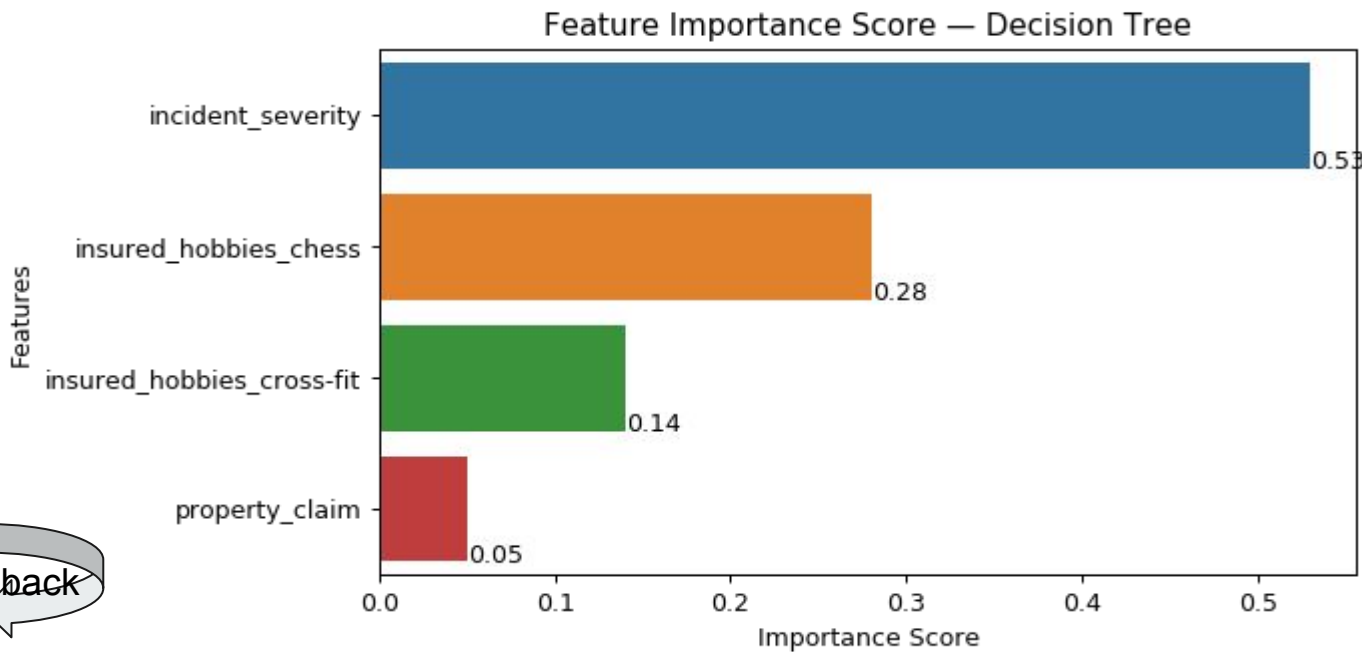
- Total claim amount
- Injury claim amount
- Property claim amount
- Vehicle claim amount

According to the pair plot, the total claim and the vehicle_claim has high positive correlation, and can be a good predictor to total claim.



Question 2

Which variables are good predictors of fraud as seen in the column called “fraud_reported (Y/N)”?



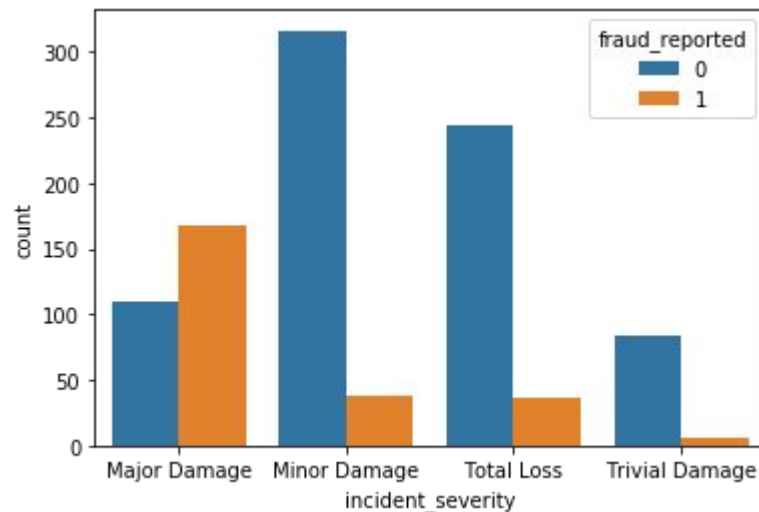
Through the selection in our best model: Decision Tree, when setting the `max_depth = 3`, the top 4 important features are as left:

Question 2

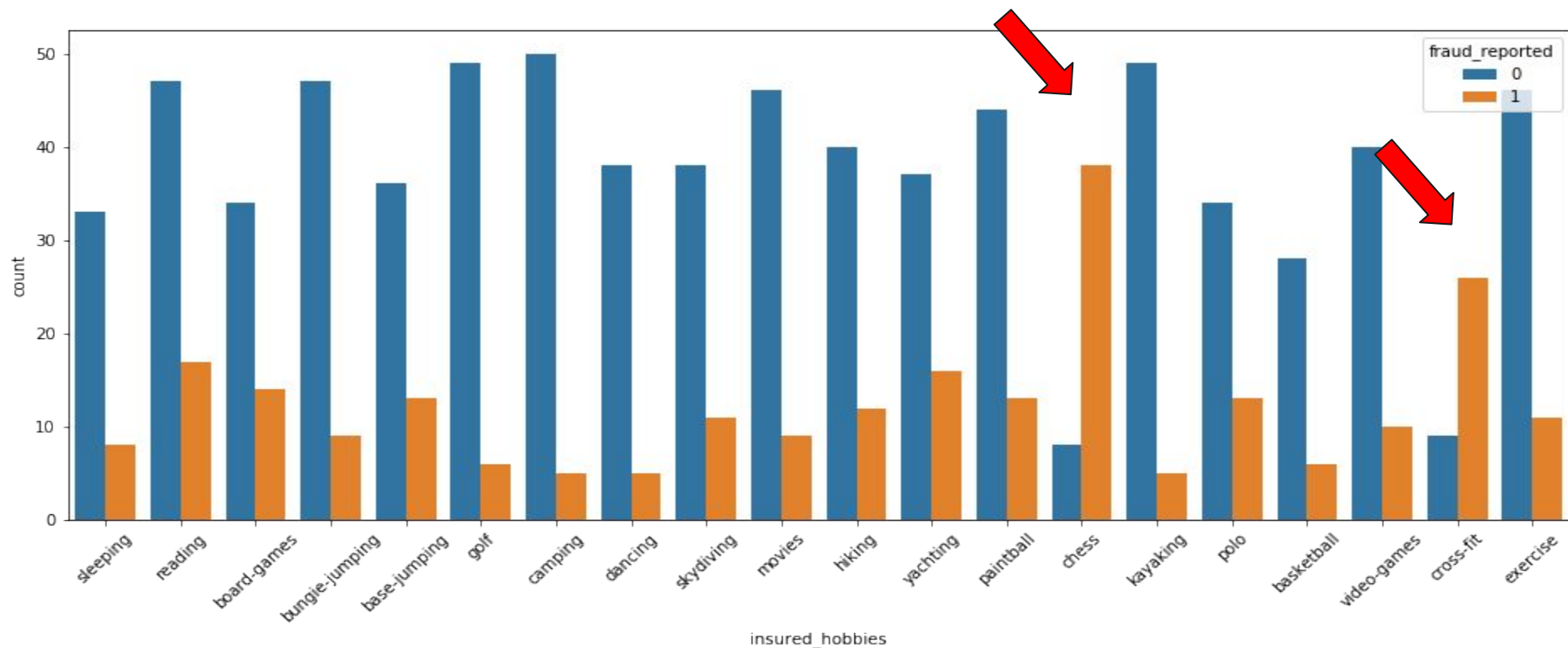
The top 4 influential features are

- Incident severity
- Insured hobby_chess
- Insured hobby_cross_fit
- Property claim

For the incident severity, Major Damage shows particular high cases comparing to the other 3 levels



It is interesting that the fraud claim is correlated to insureds' hobby, and when we look back into the data, these two hobbies shows obvious larger amount than others.





Conclusion

After applying three different model: logistic regression, decision tree, and KNN for distinguishing fraud cases, the decision tree outperforms than the other 2 models.

When applying the feature selection and check what are the important features, 4 features: incident severity, insured hobby_chess, insured hobby_cross_fit, Property claim, are the most important features and count for 100% of importance

If we are going to build fraud detector application, it could be a feasible way to weigh on these four feature. For numerical data, we can set up a threshold, if the cases get higher number than the threshold, then give higher weight. With categorical data, if the case falls in these features, add weight to the case.