CAR INSURANCE

"To File, or Not to File..."

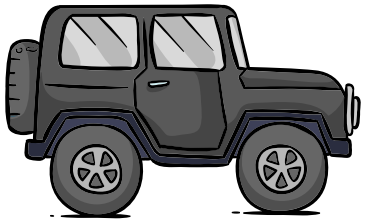
"This is a Question!"

Brenna Yin, Ting Ye, Yijun Yang

CONTENT

Data Visualisation

General Understanding of the data

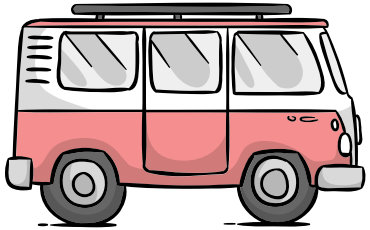


Logistic Regression

Model training and Performance

K-Nearest Neighbor with PCA

PCA analysis; Knn after PCA

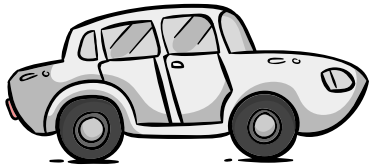


Naive Bayes (Multinomial)

Concept; Model training and Performance

Random Forest

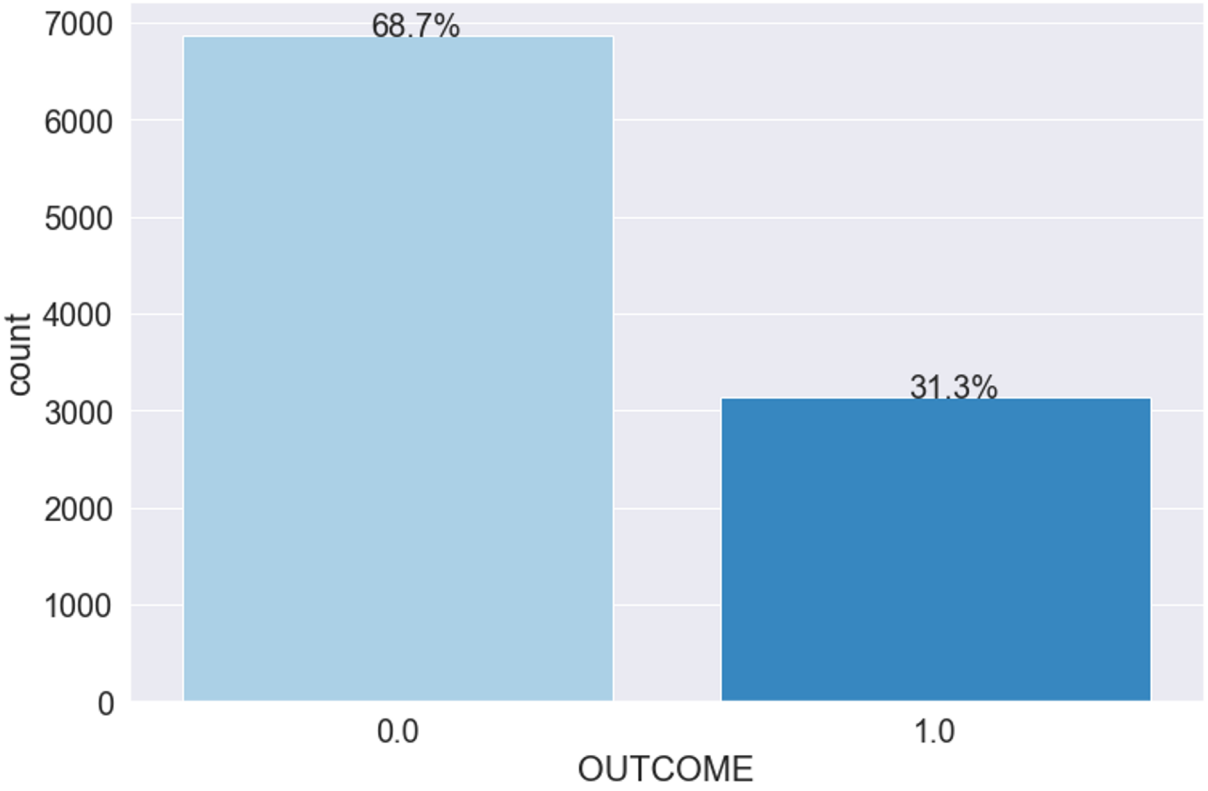
Model training and Performance



Discussion

Comparison among models; Limitations

EDA: Claim or not



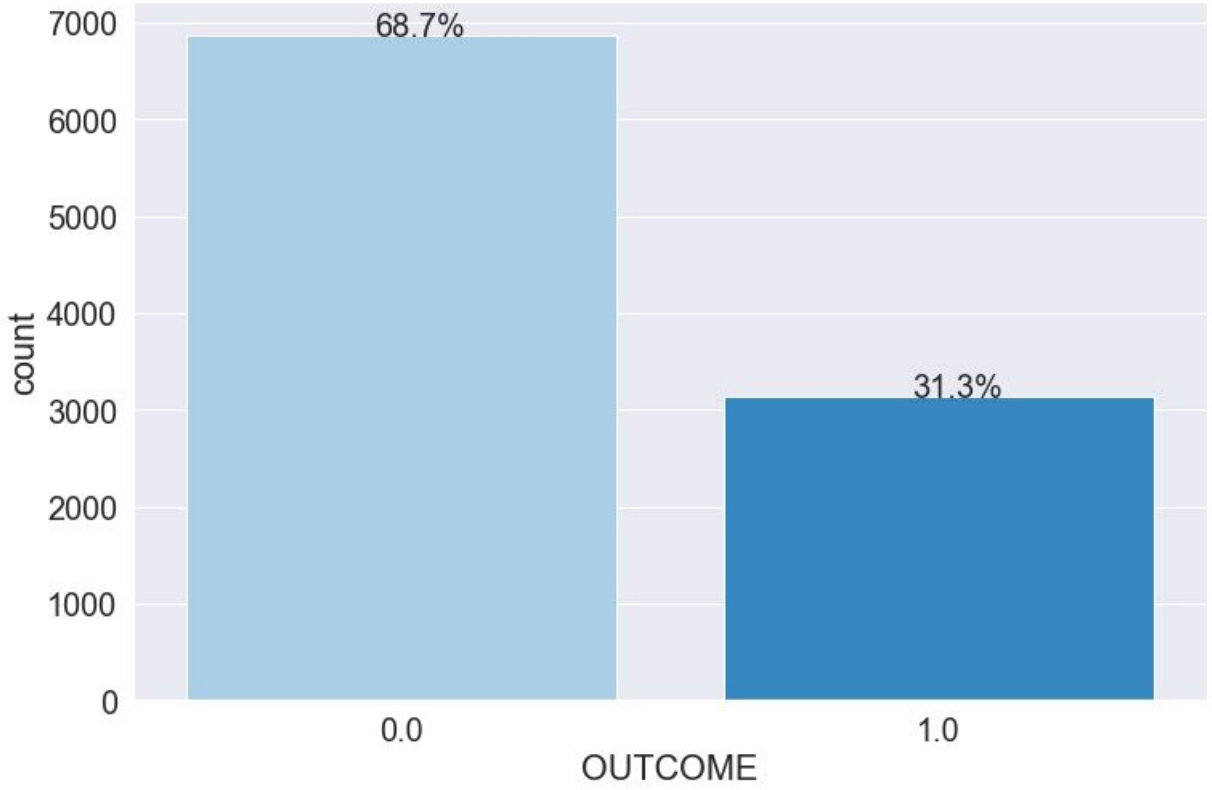
Outcome 1

Does file a claim

Outcome 0

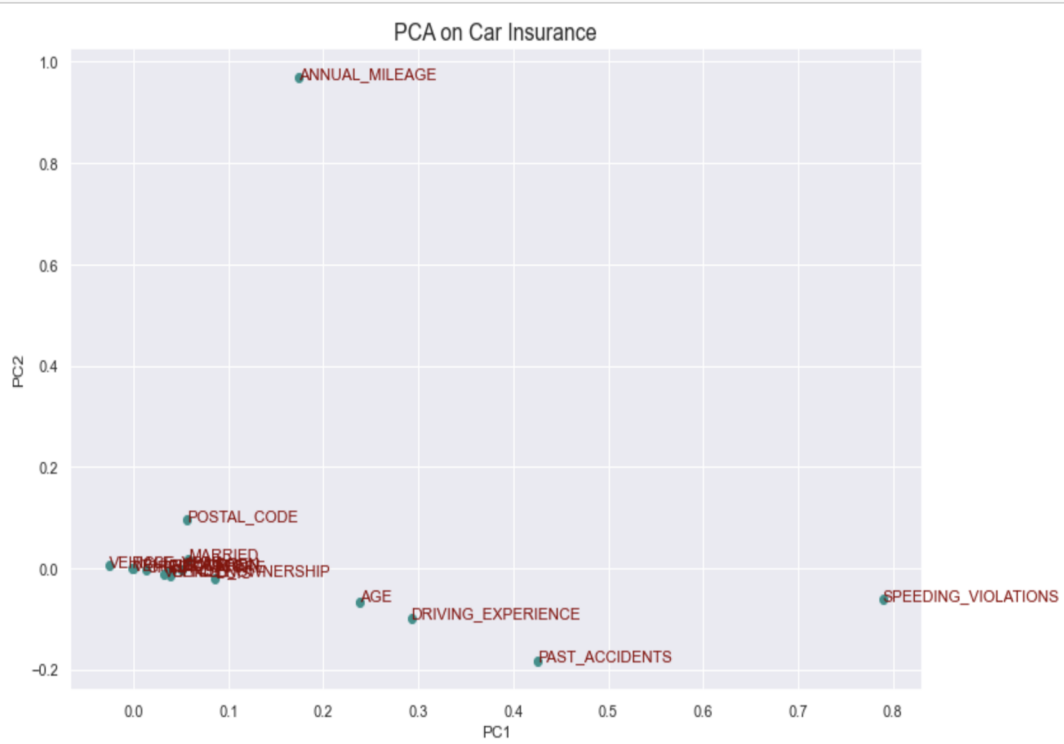
Does not file a claim

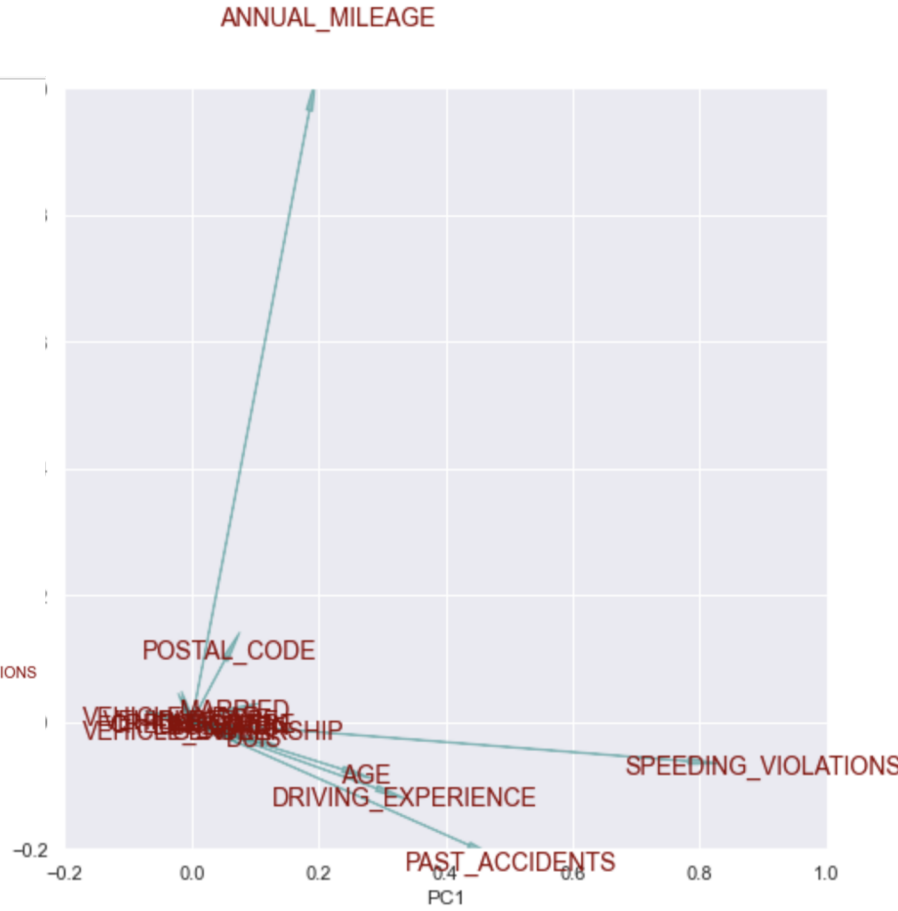
Claim or Not?



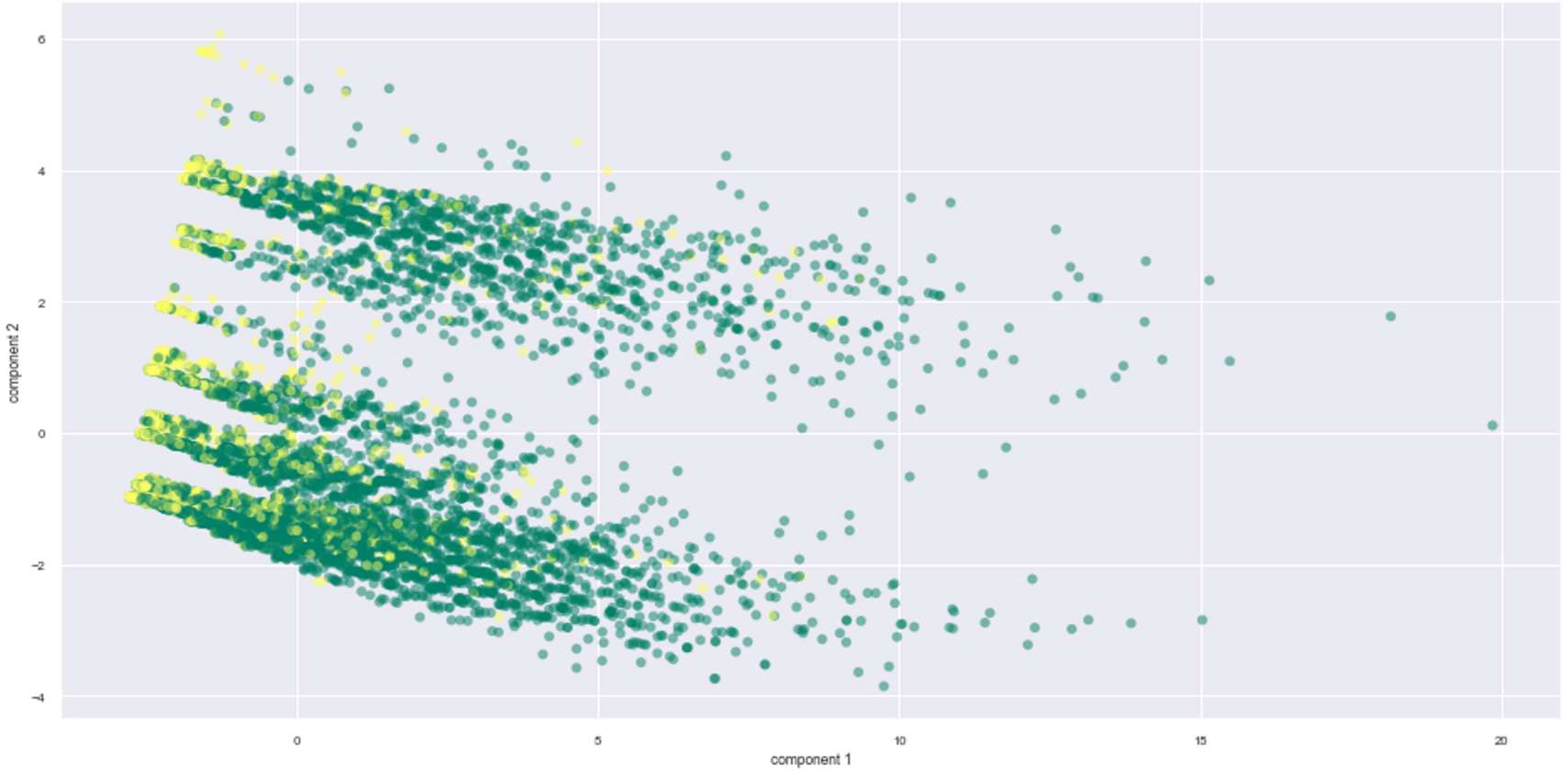
Age
Gender
Race
Education
Income
Credit Score
Vehicle Ownership
Vehicle Year
Married
Children
Postal Code
Annual Mileage
Vehicle Type
Speeding Violations
DUIS
Past Accidents

PCA loading plot 2 components

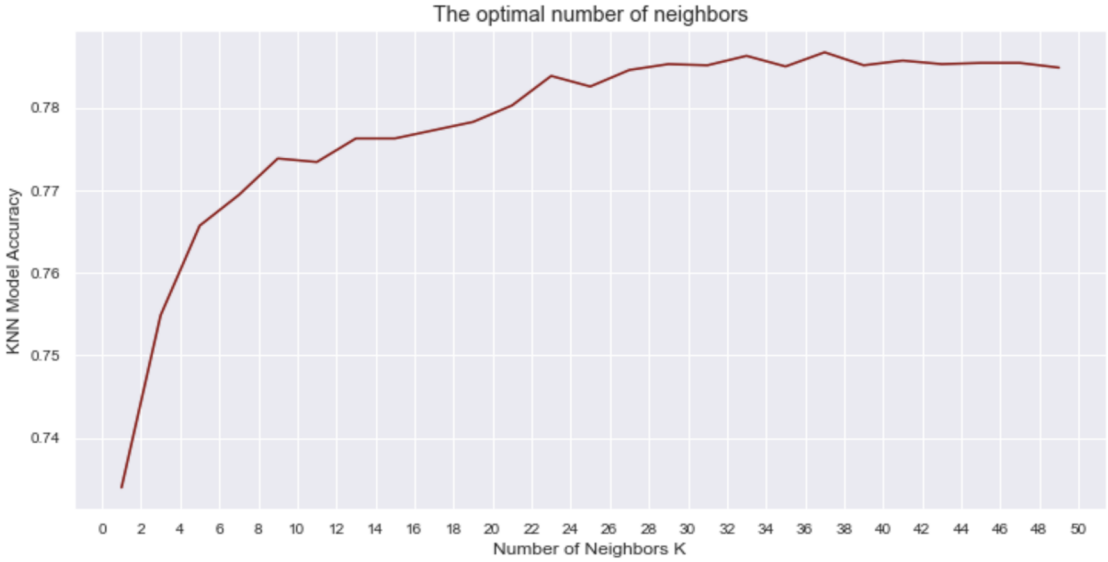




PCA: Data on 2 principle components



KNN Classification with PCA(components=2)

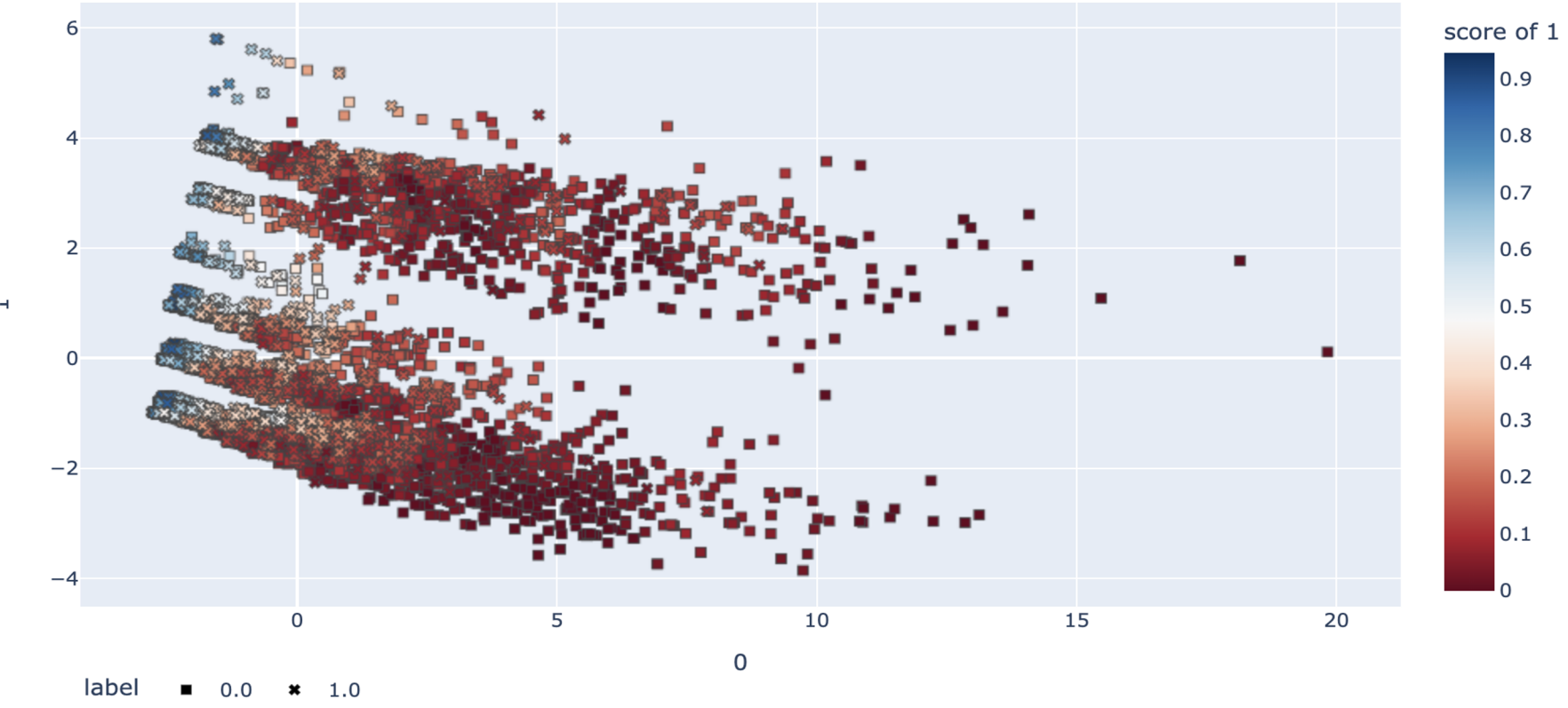


10 fold cross validation

The optimal number of neighbors is **K=37**

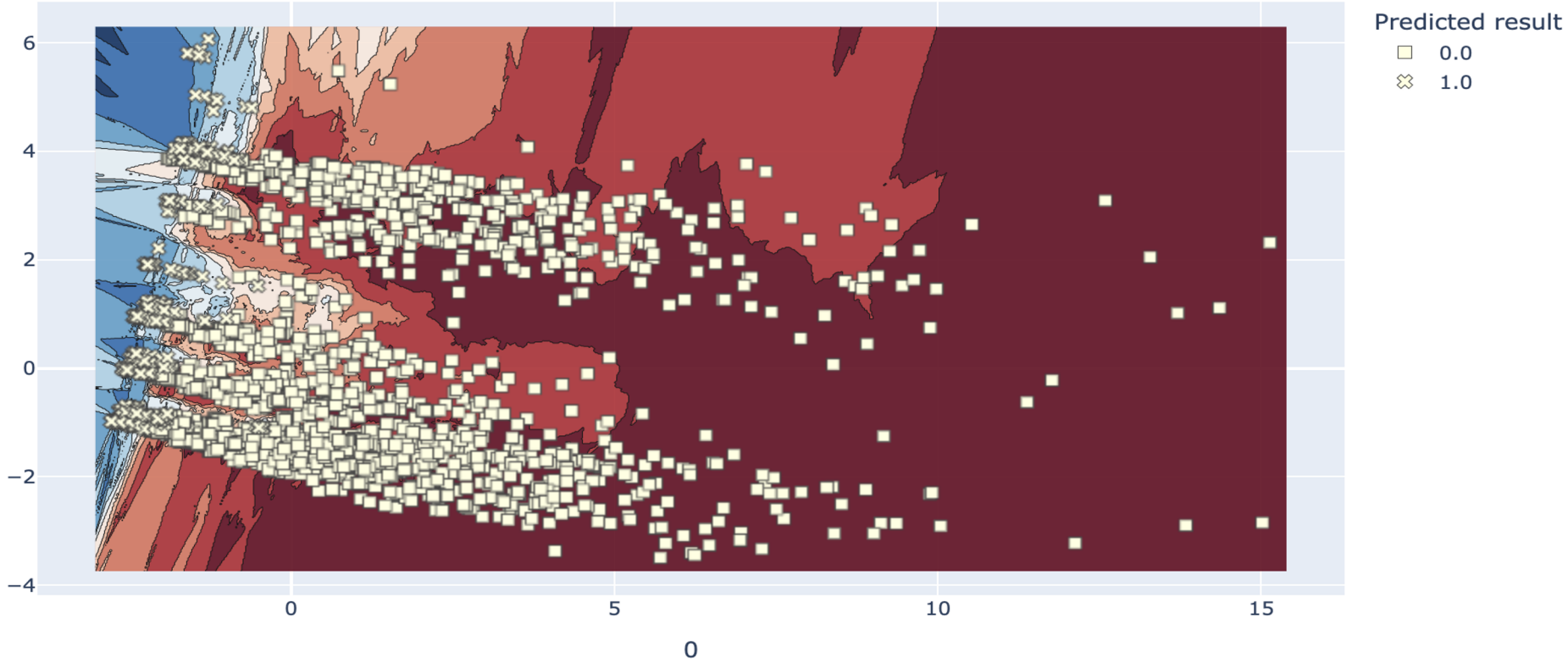
KNN Classification with PCA(components=2)

PCA transformed train data

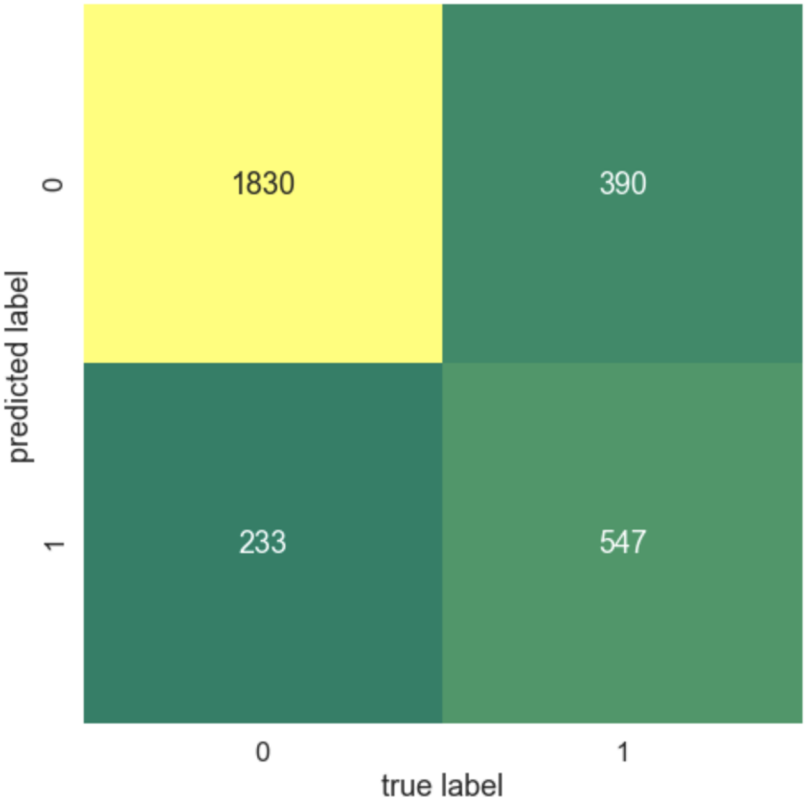


KNN Classification with PCA(components=2)

KNN predicted result on PCA transformed test data



KNN Classification with PCA(components=2)



Accuracy — 79.23%

Best K — 37

KNN Model Optimization

When Component is 2, Cumulative explained variance is 0.7

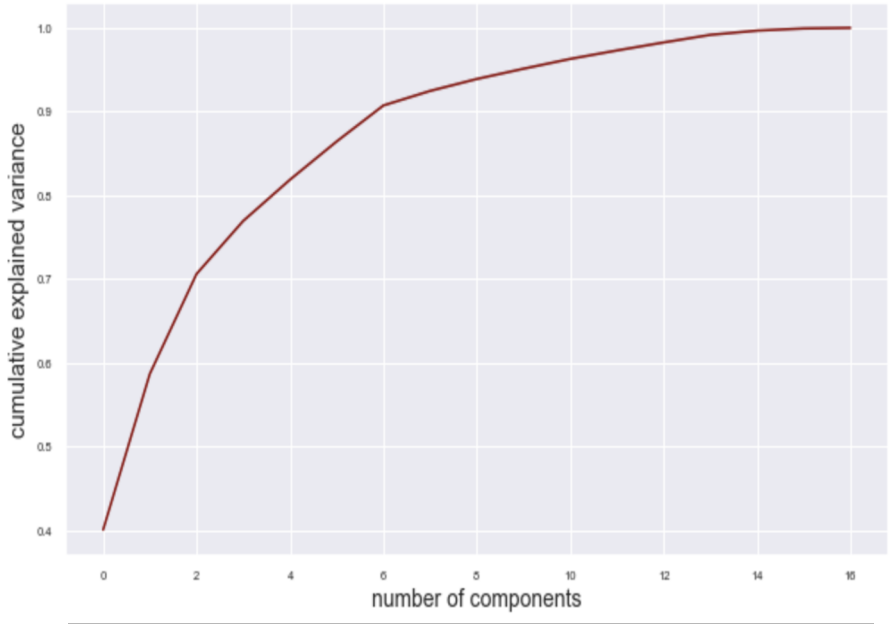
01

We need better components number

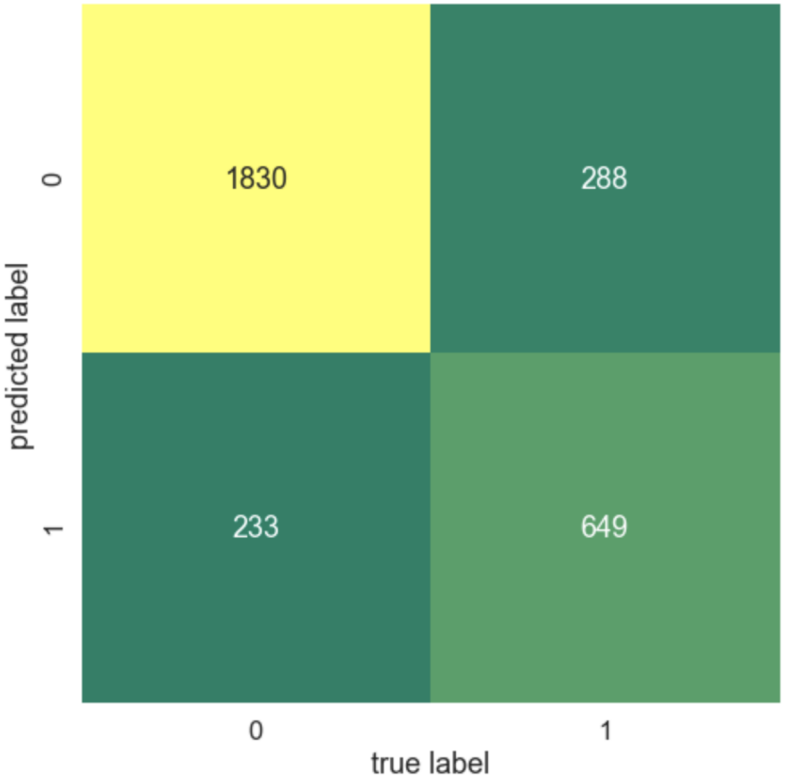
02

We need a optimal combination of Components number and K value

03



KNN Optimization with best (components,K) combination



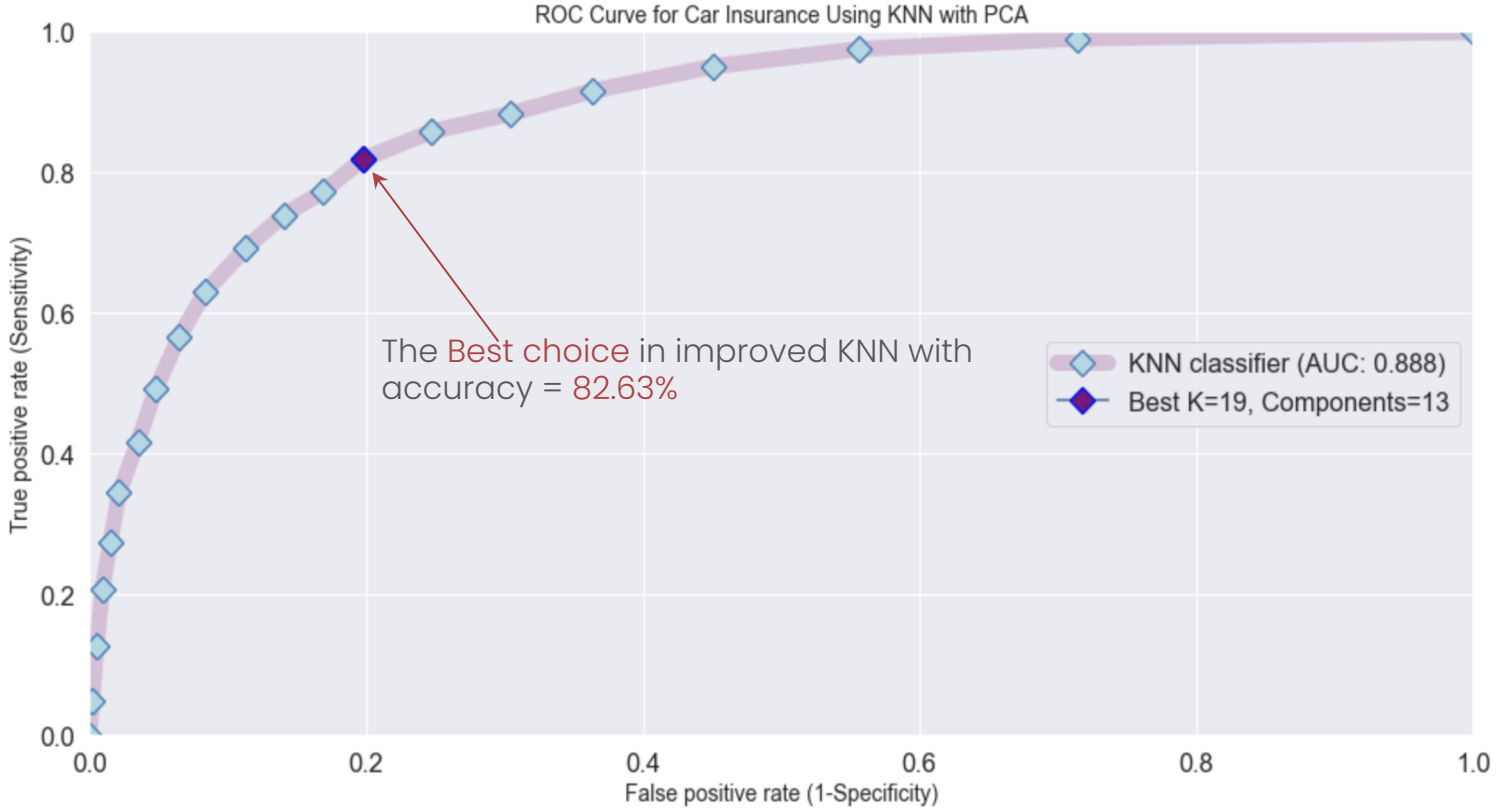
Improved from 79.23% (components=2, K=37)

Accuracy — **82.63%**

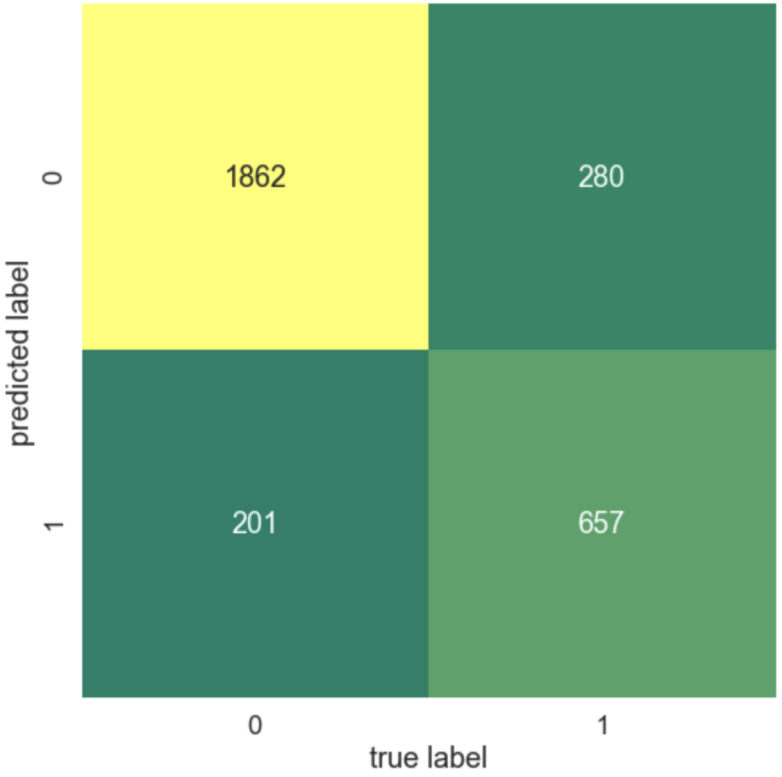
Best Combination — **Components=13 K=19**

Calculated the best combination by circular iteration

Improved KNN ROC & AUC Evaluation



Random Forest (Default parameters)

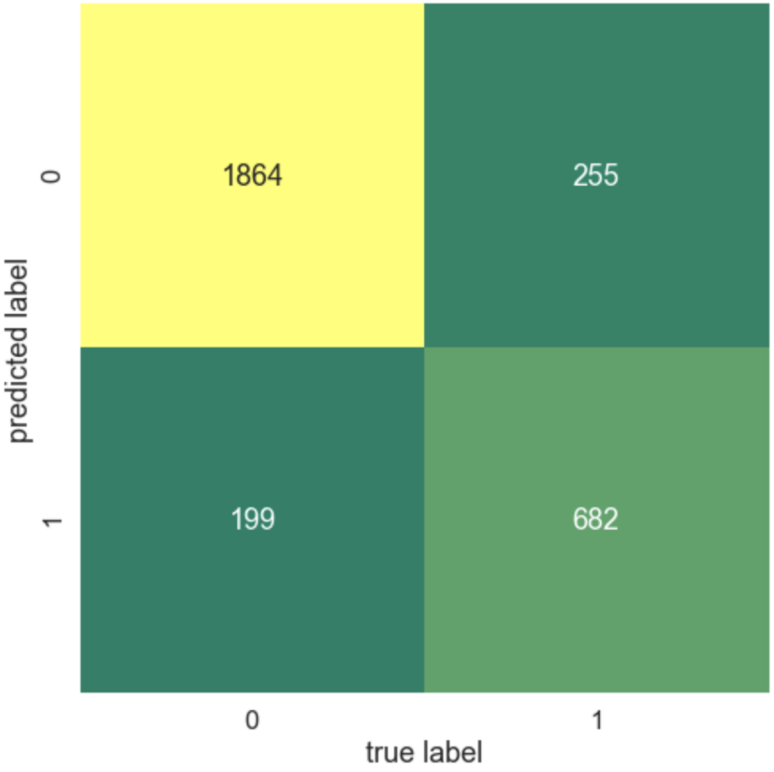


Accuracy — **84%**

Default parameters

max_depth=None
n_estimators=10
min_sample_leaf=1
min_sample_split=2

Random Forest Hyperparameter Tuning



Improved from 84% (Default Parameters)

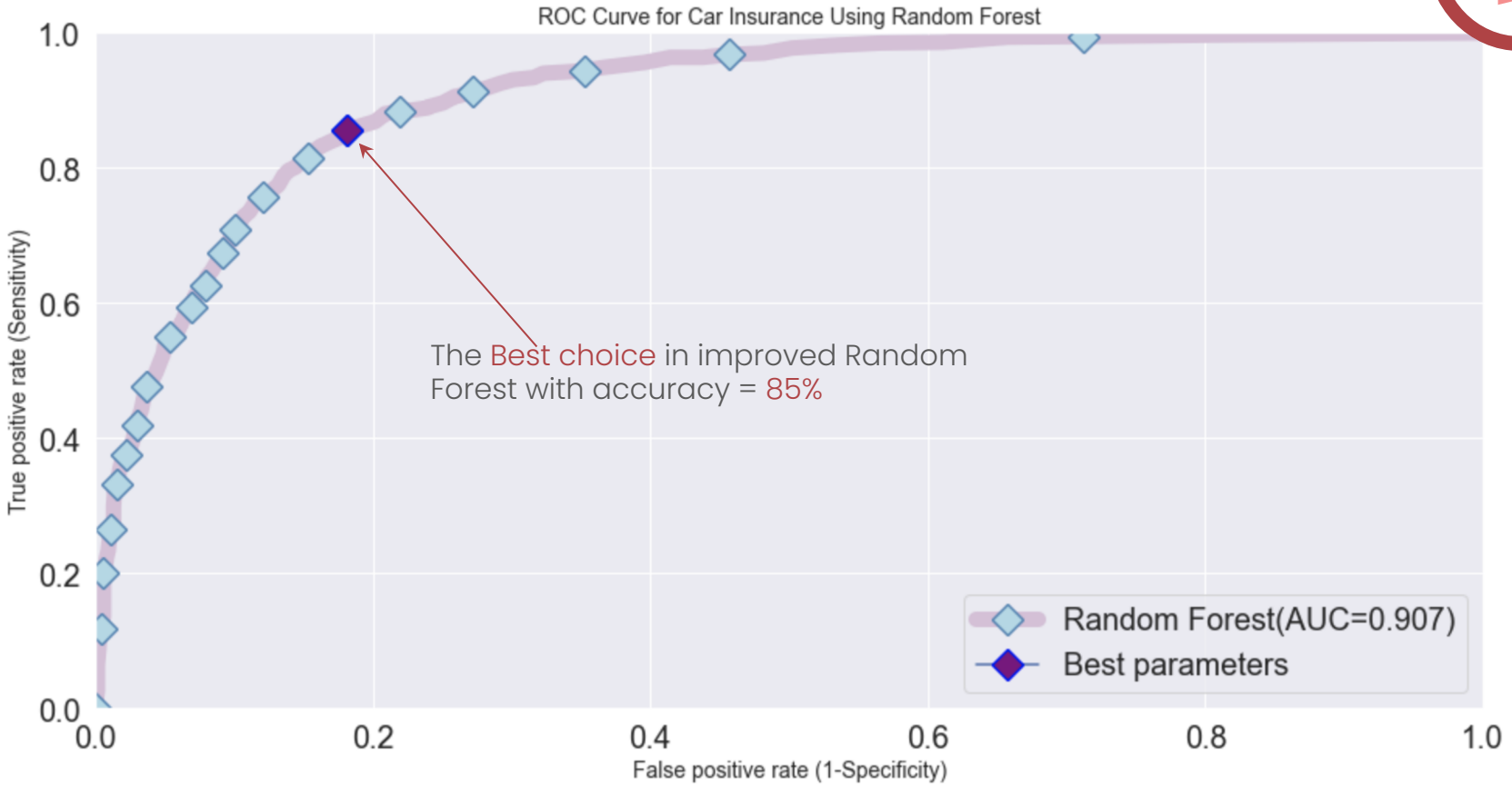
Accuracy — **85%**

Parameters

max_depth=10
n_estimators=4
min_sample_leaf=2
min_sample_split=50

Calculated by GridSearch (3 Fold cross validation)

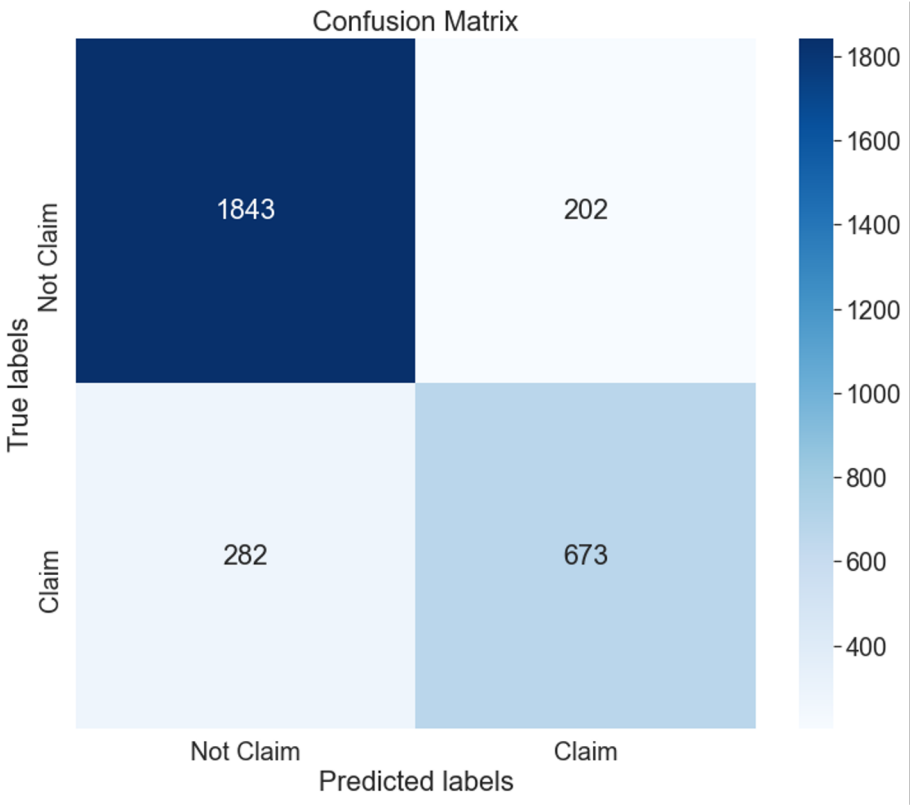
Improved Random Forest ROC & AUC Evaluation



Logistic Regression

Gender	Male group has **2.619** times the odds of female group of filing a claim.
Driving Experience	With **10 years** of extra in driving experience, probability of filing a claim **drops by 84%**.
Vehicle Ownership	Self-owned cars have **83% less** likely than not self-owned cars of filing a claim.
Vehicle Year	Old cars have **6.0 times** the odds of new cars of filling a claim.
Area	San Diego has **1.66 times** the odds of Oviedo of filing a claim, than Baltimore, than New York.
Speeding Violation	With **1 case** of extra in speeding violation, probability of filing a claim **drops by 10%**.

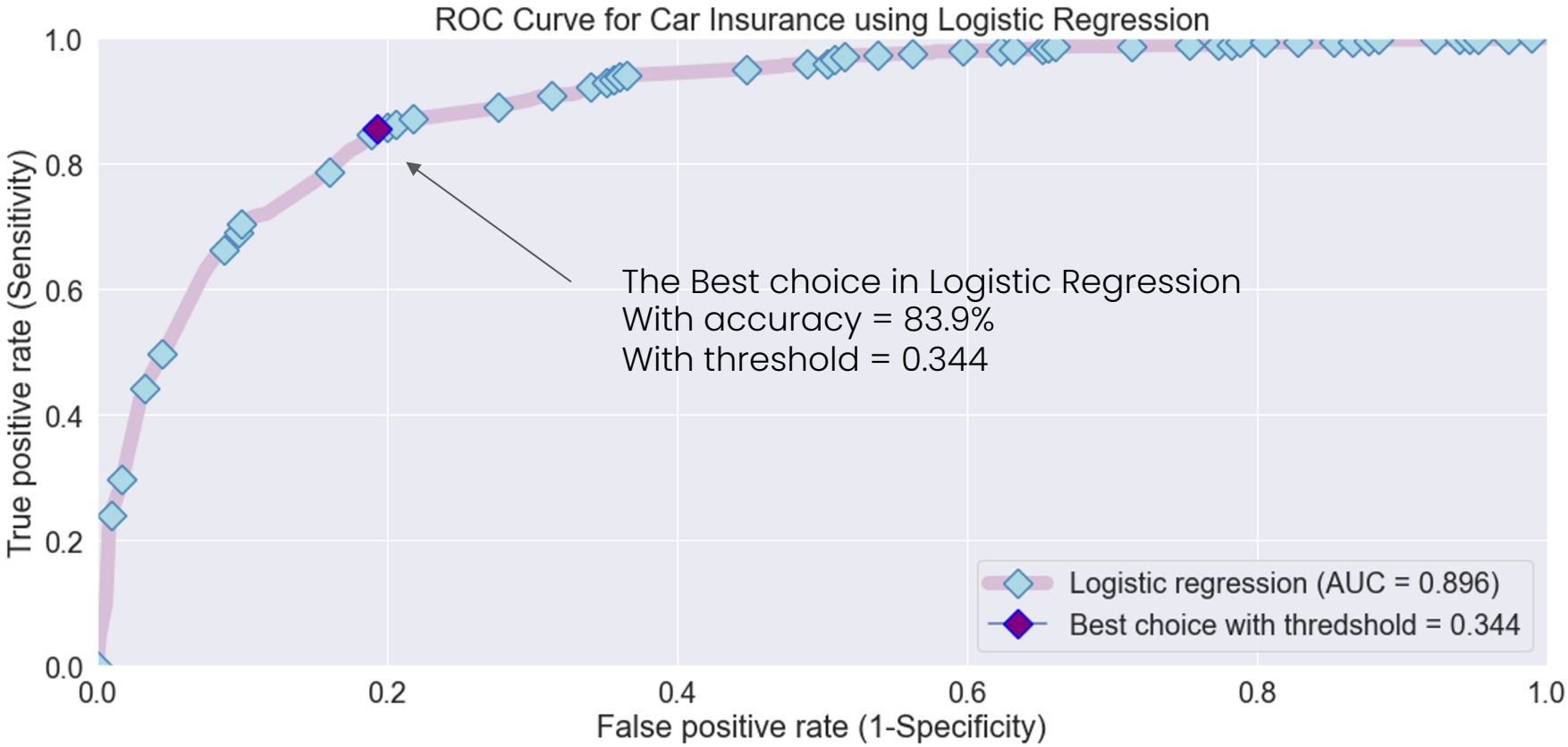
Confusion Matrix



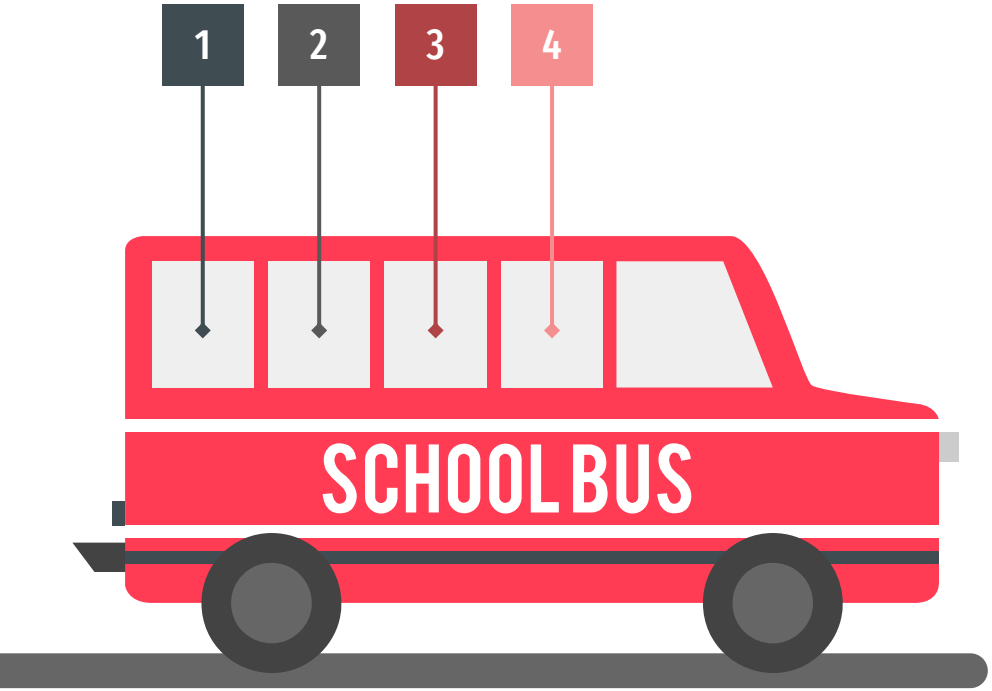
accuracy — 83.9%

threshold — 0.344

ROC & AUC Evaluation of Logistic Regression



Naive Bayes – A Classifier Based on Bayes' Theorem



1

Prior

What is the "prior" assumption of events' probability distribution?

2

Likelihood

What is the probability that the sample would occur if class being defined?

3

Evidence

What is the probability that such sample would occur generally?

4

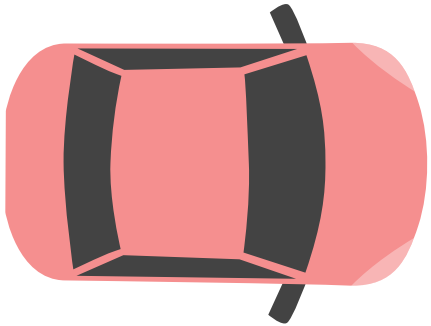
Posterior

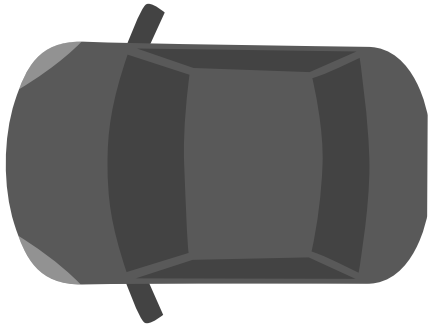
What is the probability that given this sample, that the sample is in a certain class?

Important Preconditions

Independence

Naive Bayes Classifier requires predictors are independent

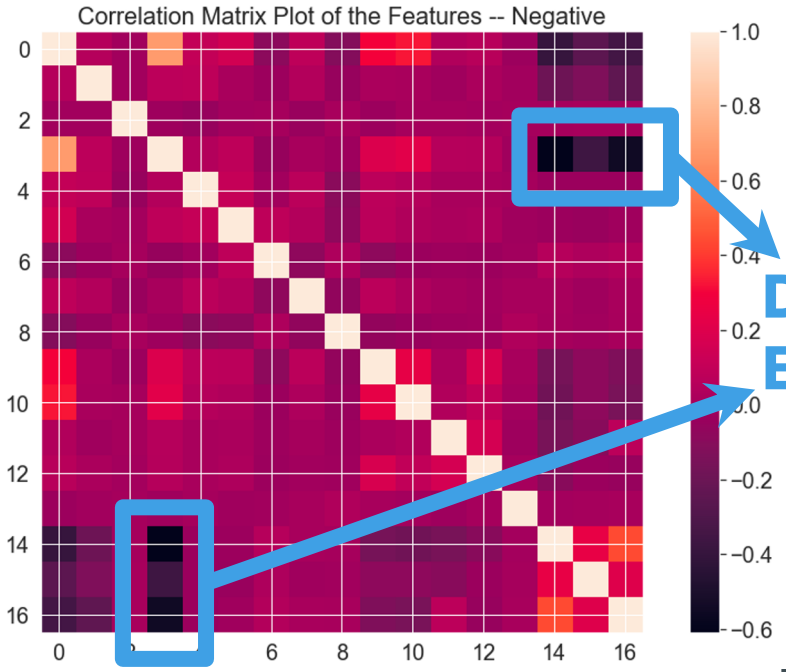




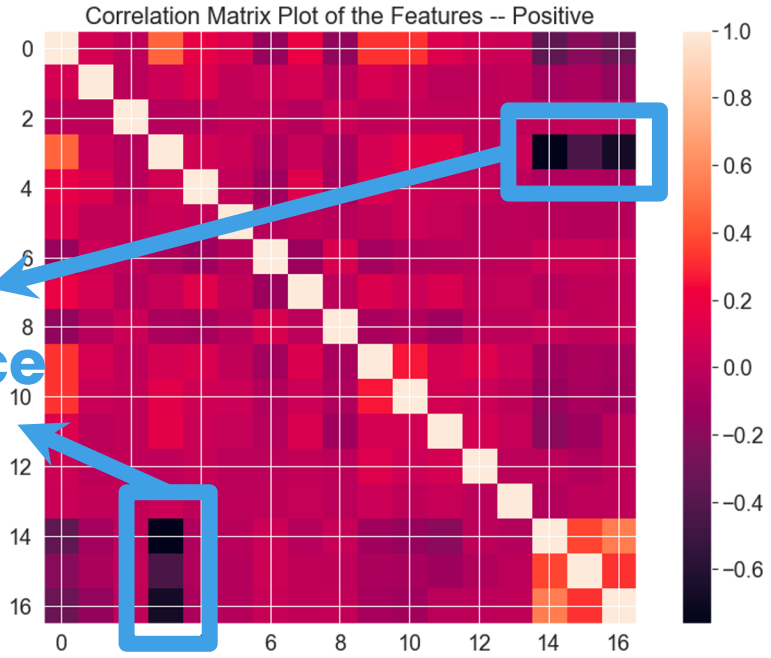
Prior Distribution

$p(C_k)$ needs to be defined in advance.

Independence



Driving Experience



Prior Distribution – Multinomial NB

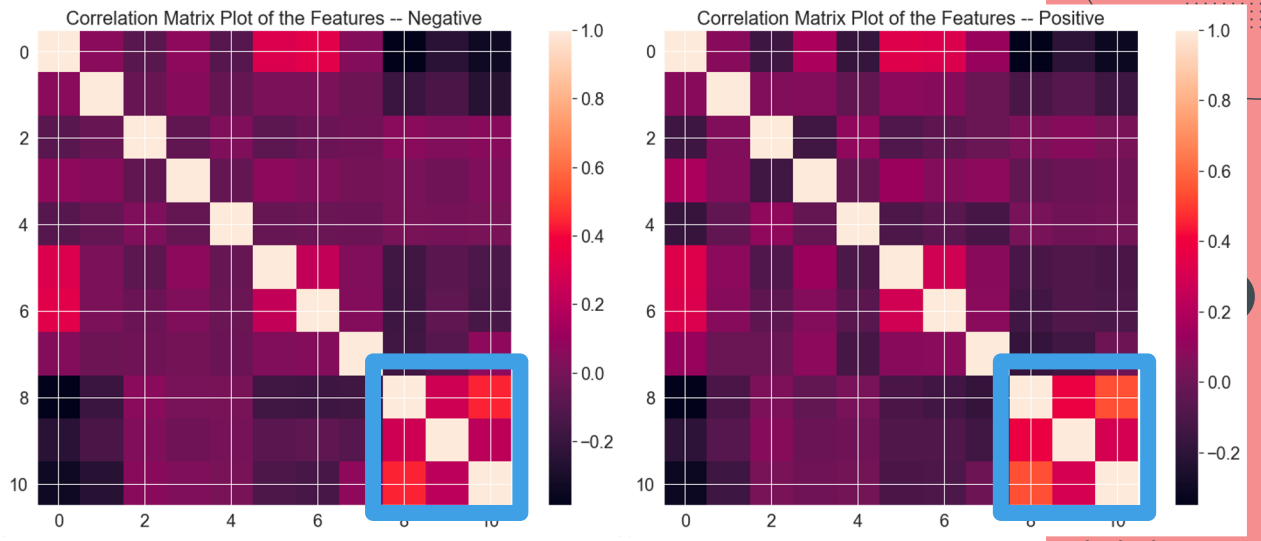
	Min	25%	50%	75%	Max
Credit Score	0.053358	0.417191	0.525033	0.618312	0.960819
Speeding Violations	0	0	0	2	22
DUIS	0	0	0	0	6
Past Accidents	0	0	0	2	15

→ Percentile

→ "Yes/No"

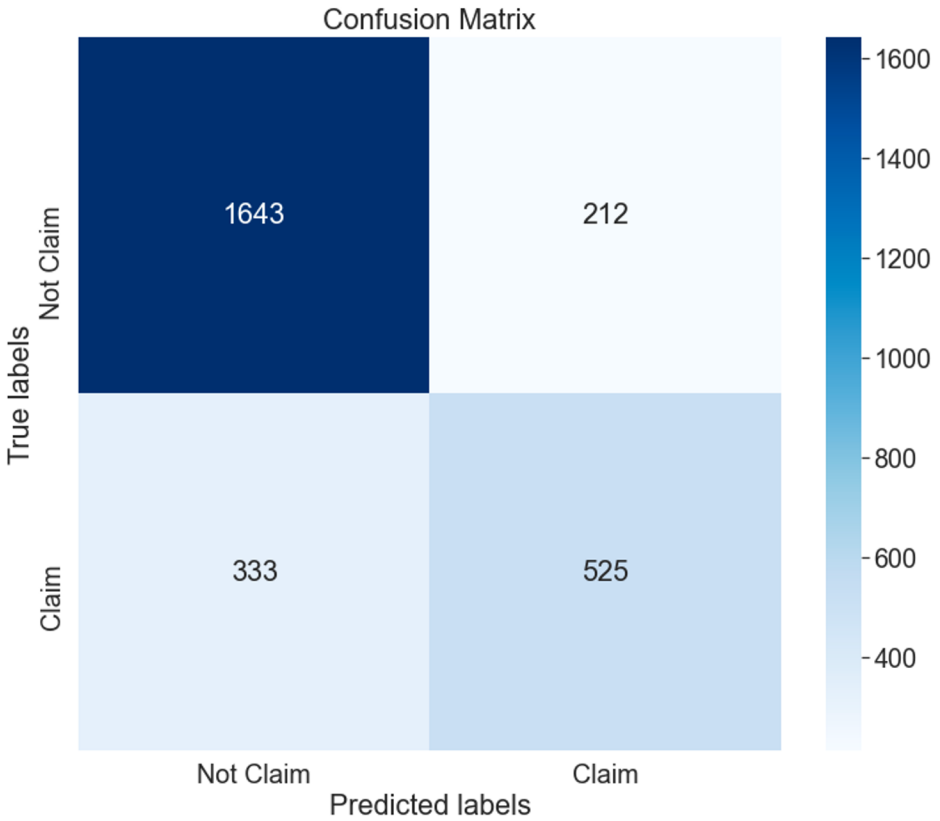
After Backward Selection

- Age
- Gender
- Credit Score
- Vehicle Ownership
- Vehicle Year
- Married
- Children
- Postal Code
- Speeding Violations
- DUIS
- Past Accidents



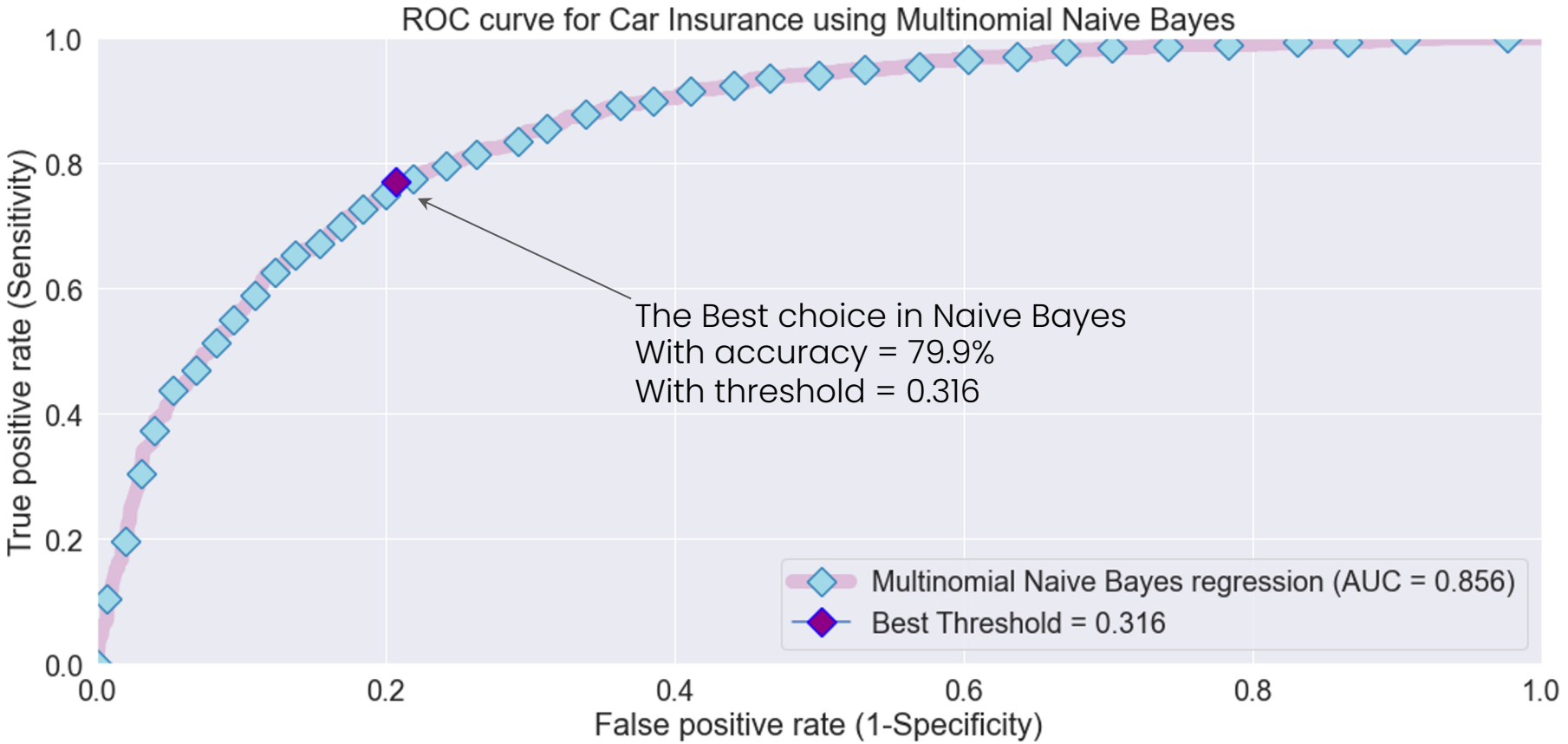
Speeding Violations/DUIS/Past Accidents

Confusion Matrix



accuracy	79.9%
threshold	0.316

ROC & AUC Evaluation of Naive Bayes



Comparison

Model	Accuracy	AUC
Knn with PCA	82.6%	88.8%
Random Forest	85.0%	0.907
Logistic Regression	83.9%	0.896
Naive Bayes (Multinomial)	79.9%	0.856

Random Forest Performs the Best!

Limitations

1

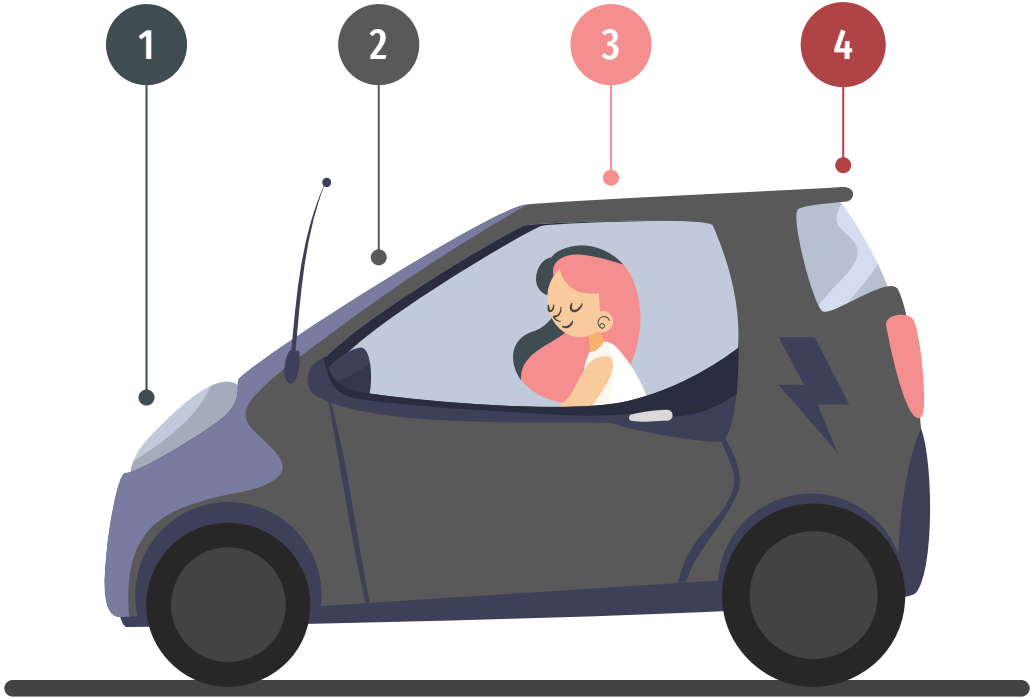
Knn with PCA

Mercury is the closest planet to the Sun and the smallest

2

Random Forest

Despite being red, Mars is a cold place full of iron oxide dust



3

Logistic Regression

Good for binary outcomes. Weak in complex models with much variables.

4

Naive Bayes

Prior Distribution; Independence is not guaranteed.

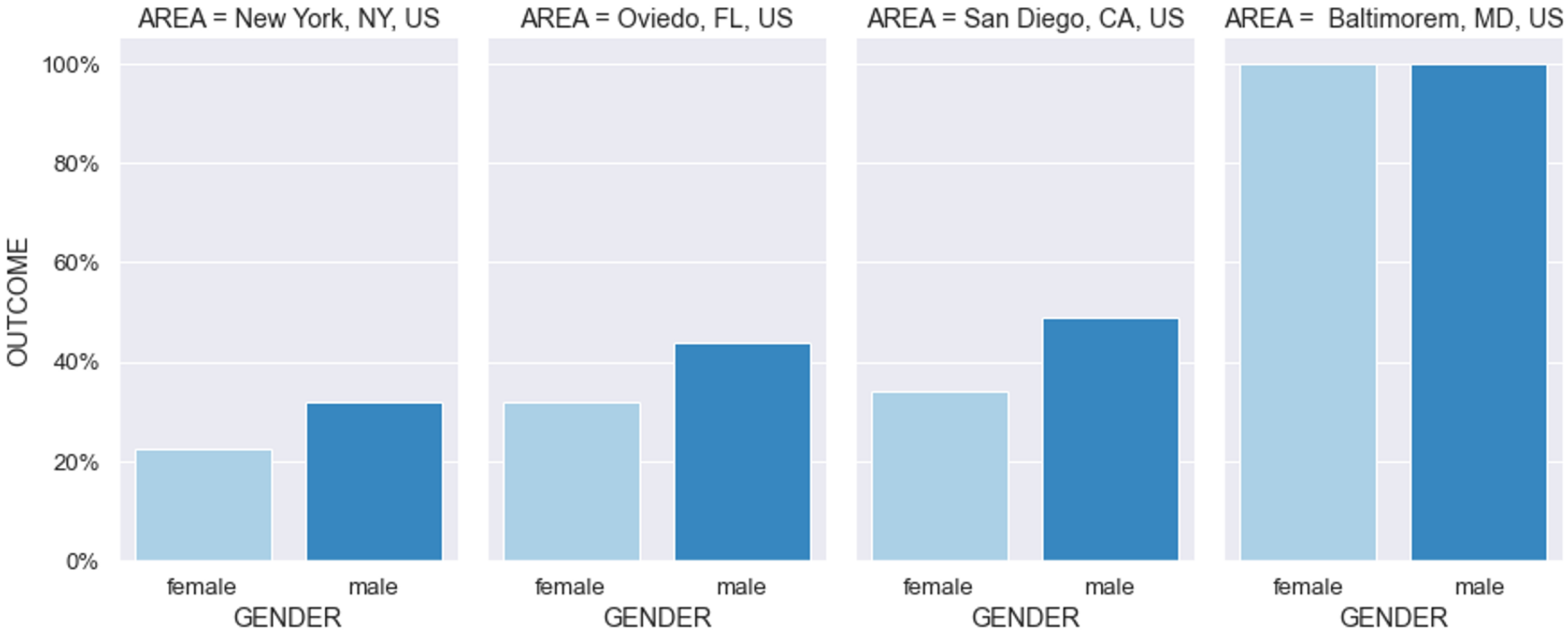
Data Cited from: *https://www.kaggle.com/sagnik1511/car-insurance-data*

Thank You!

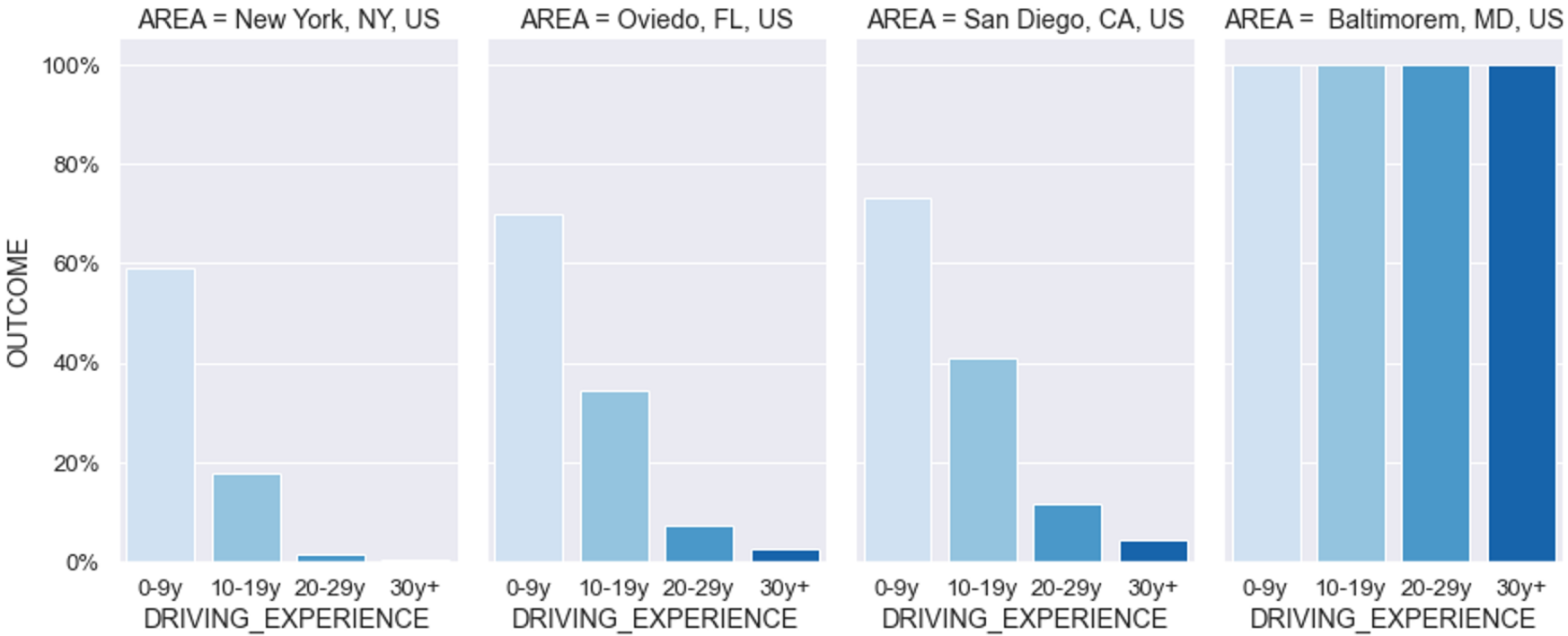
Any Questions?

Appendix

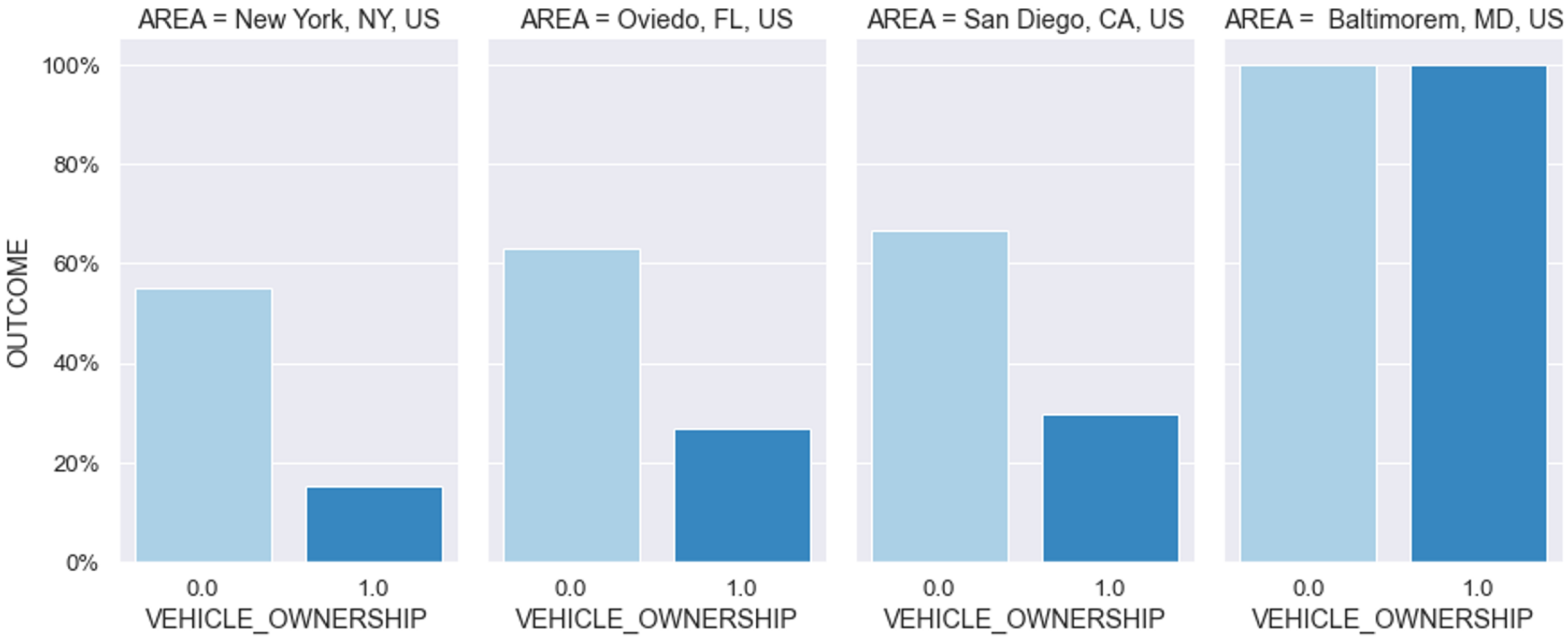
EDA: Gender


AREA = New York, NY, US
AREA = Oviedo, FL, US
AREA = San Diego, CA, US
AREA = Baltimorem, MD, US
OUTCOME
100%
80%
60%
40%
20%
0%
female
male
GENDER

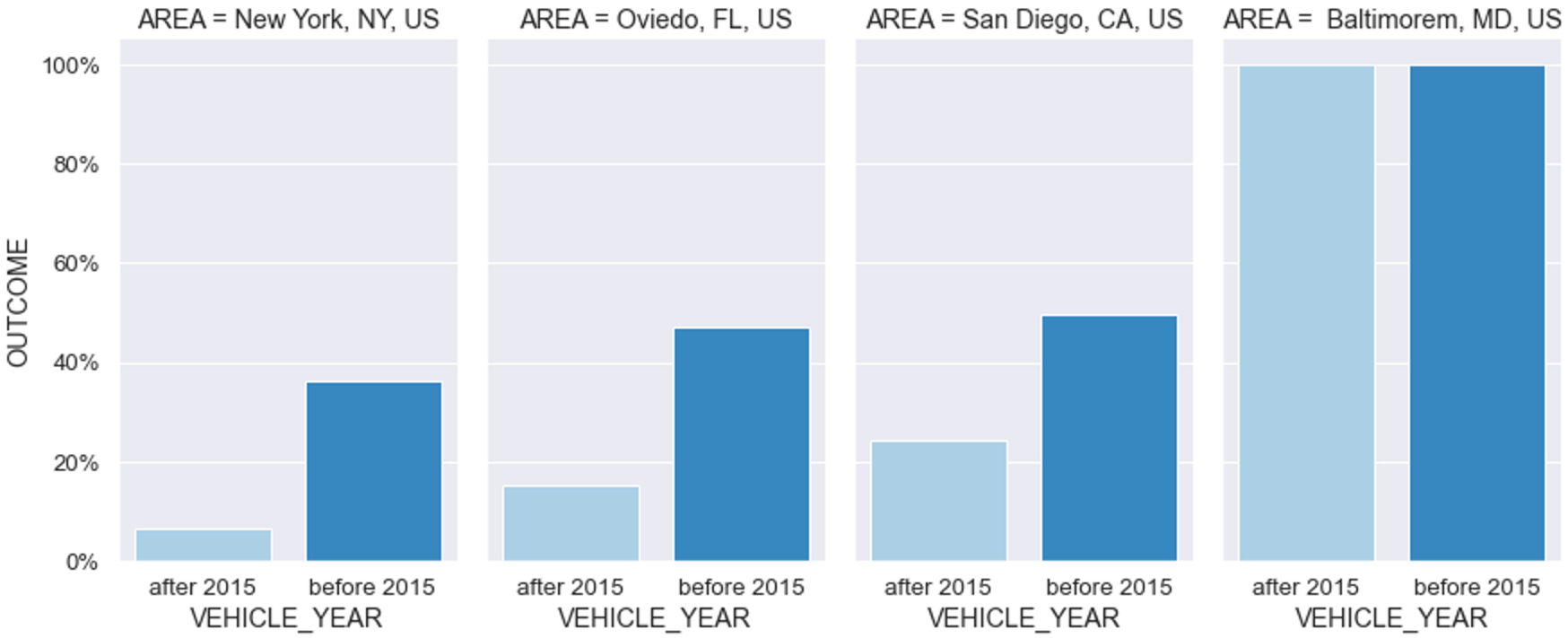
EDA: Driving experience



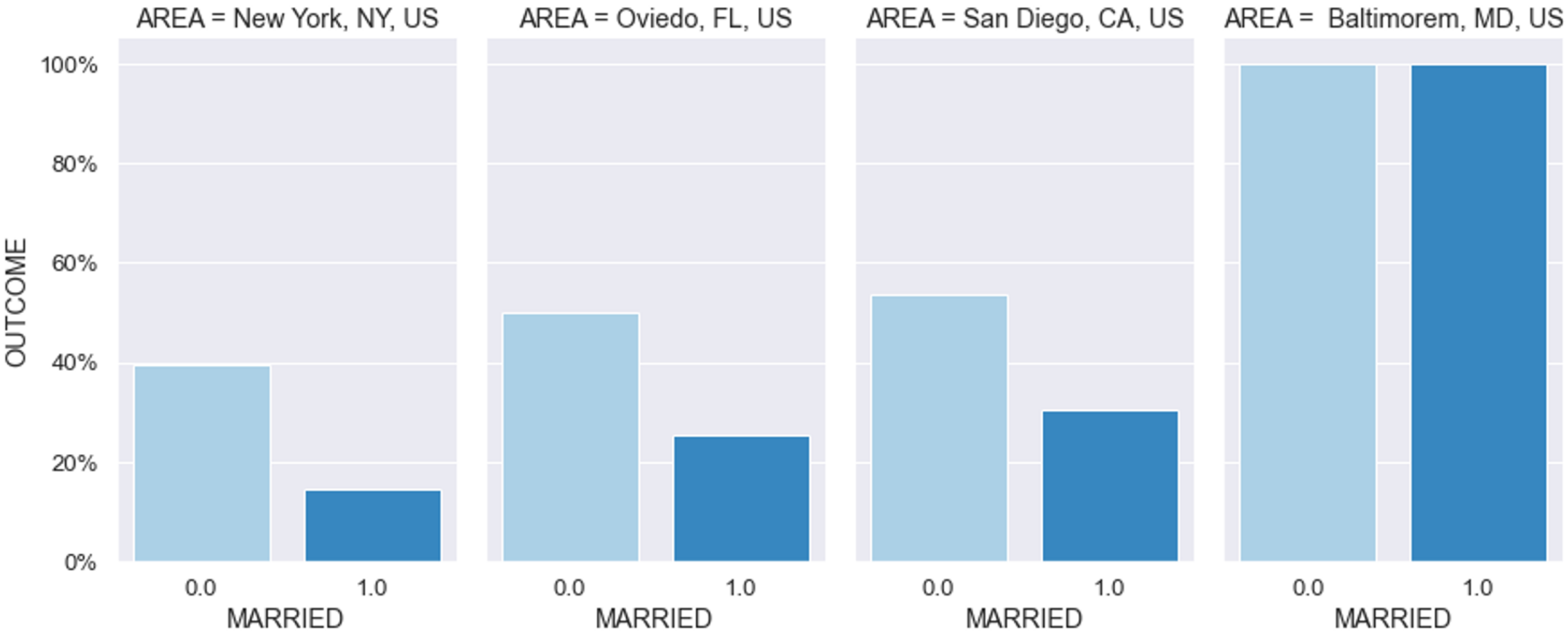
EDA: Vehicle ownership



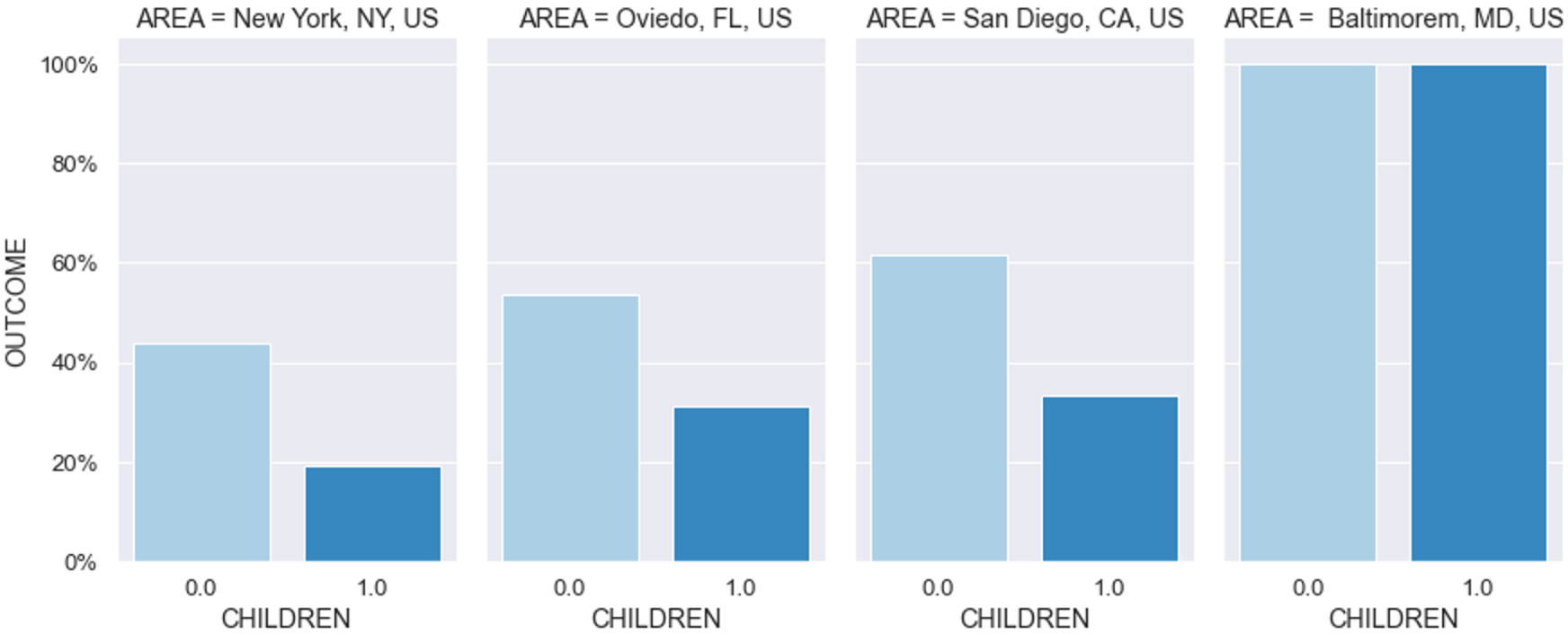
EDA: Vehicle year



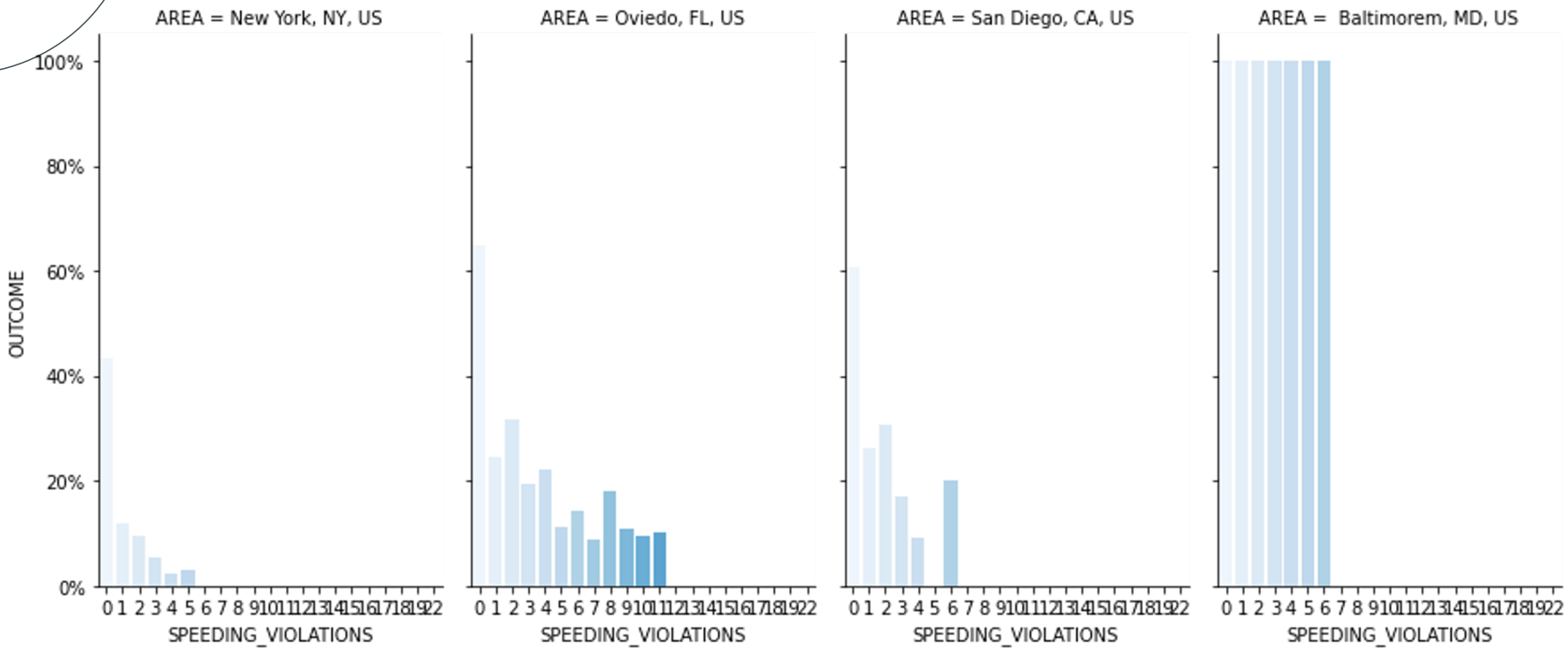
EDA: Marital status

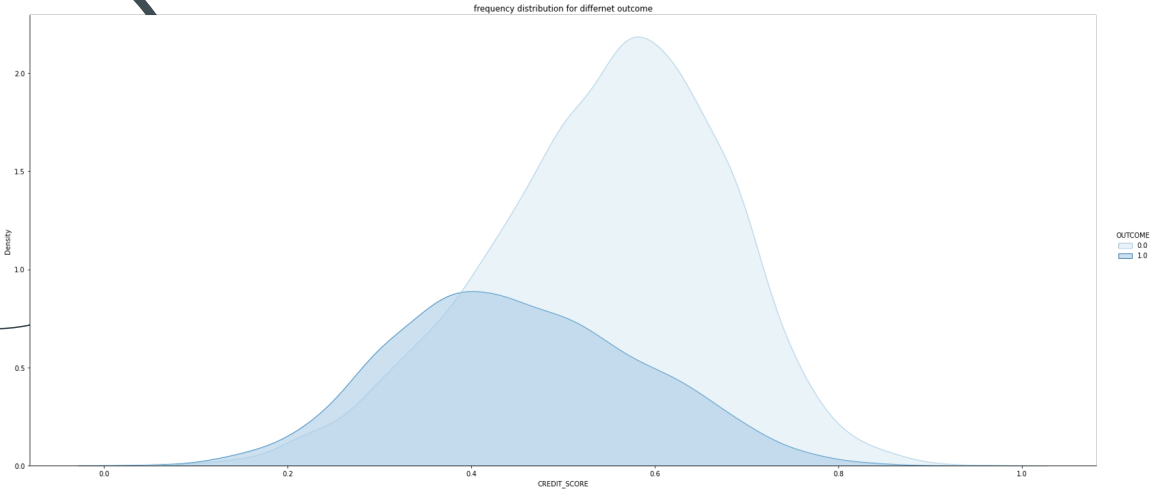


EDA: Children

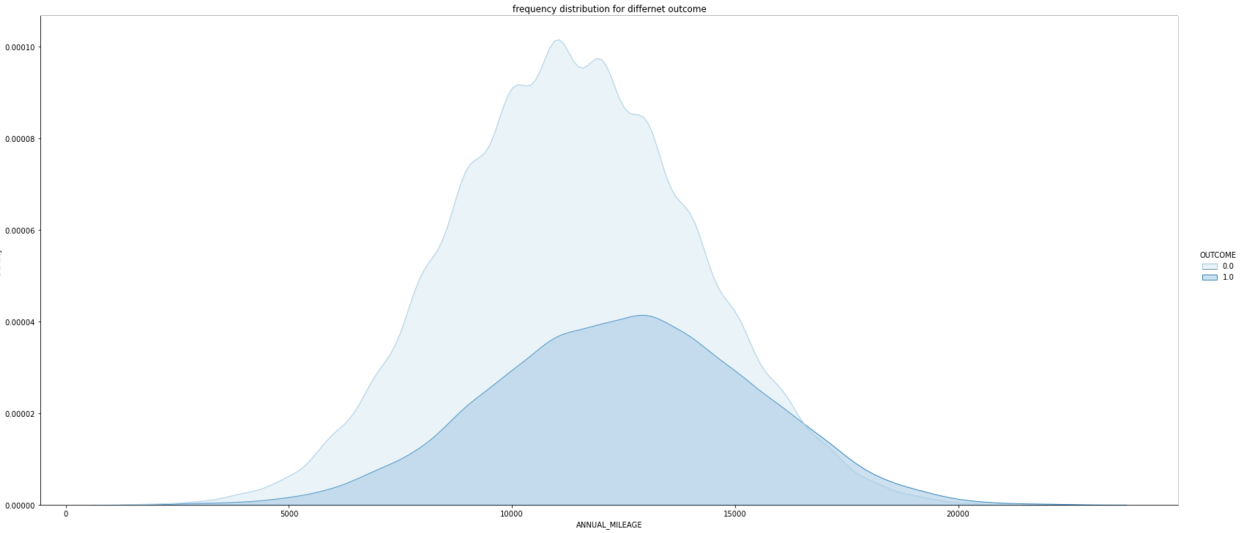


EDA: Speeding Violation

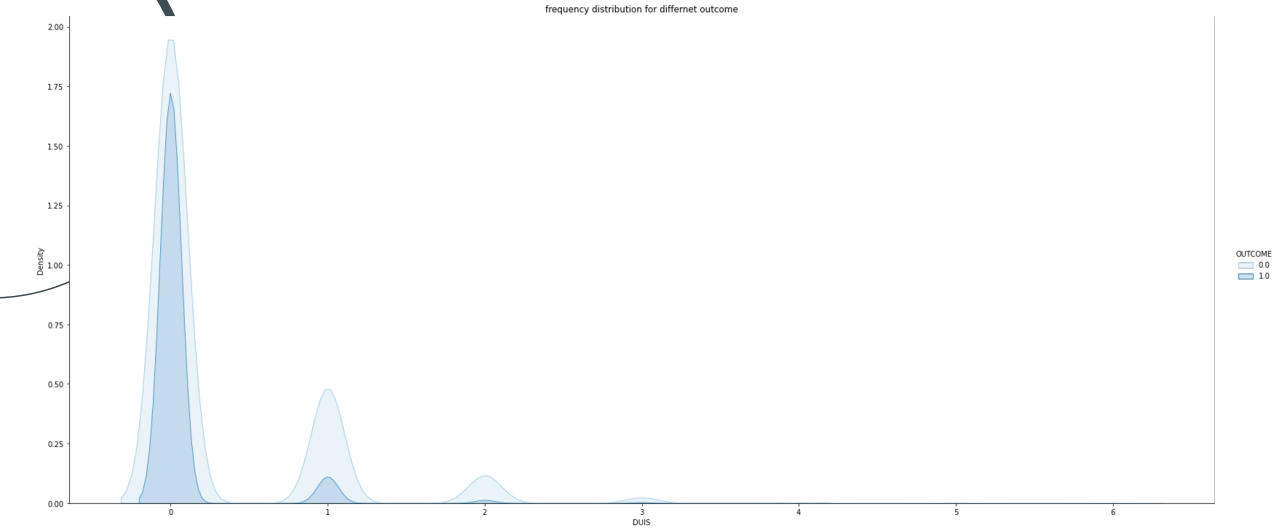




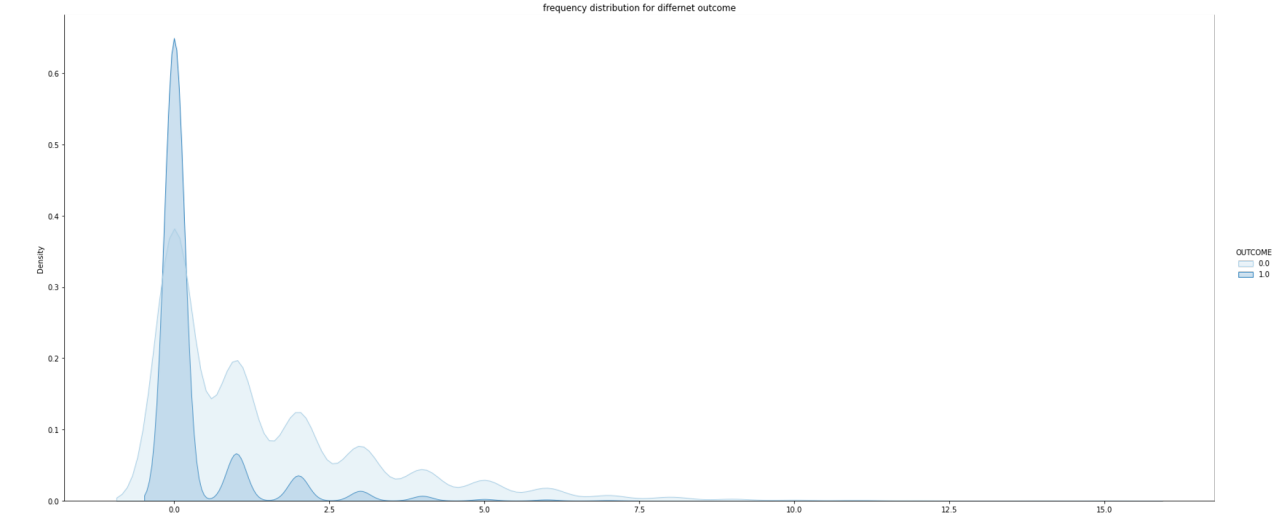
EDA: Credit Score



EDA: Annual Mileage

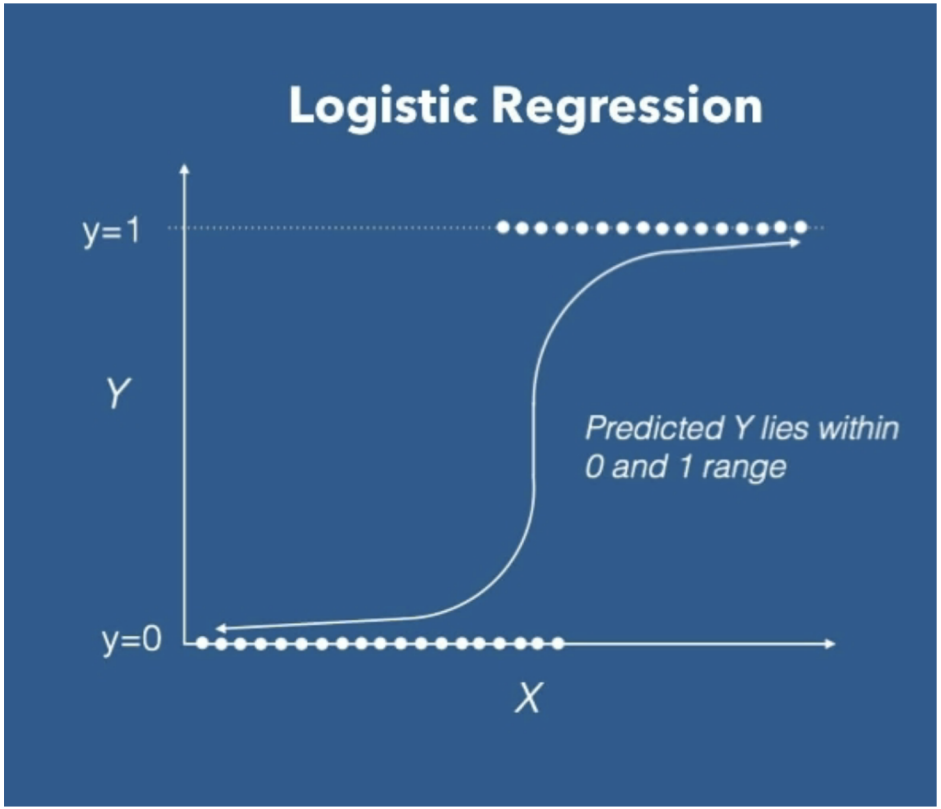


EDA: DUIS



EDA: Past Accident

Logistic Regression



Outcome 1

Does file a claim

Outcome 0

Does not file a claim

Select Features for Logistic Regression

With backward selection and 10-folds cv

Gender

Female or male

Driving Experience

Number of years in driving vehicle

Ownership

Owned the vehicle or rent the vehicle

Vehicle Year

When was the car made? Is it a new car or an old car?

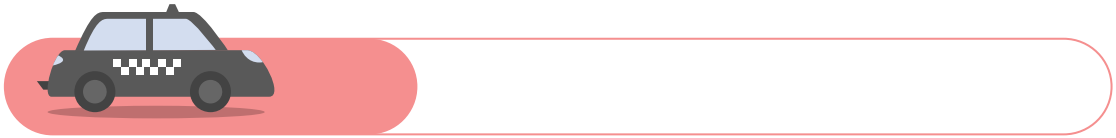
Area

Where does the driver live. East coast? West coast? Middle area?

Speeding Violation

Number of speeding violation that

Understand the Coefficient



Female

Male

Gender

Male group has 2.619 times the odds of female group of filing a claim.

Understand the Coefficient

X years of driving experience

(X+10) years of driving experience

Driving Experience

With 10 years of extra in driving experience, probability of filing a claim drop by 84%.

Understand the Coefficient

Self-owned cars

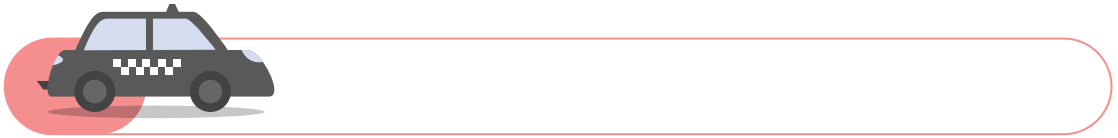
Not self-owned cars

Vehicle Year

Self-owned cars has **83% less** likely than not self-owned cars of filing a claim.

Understand the Coefficient

Old car
made before 2015

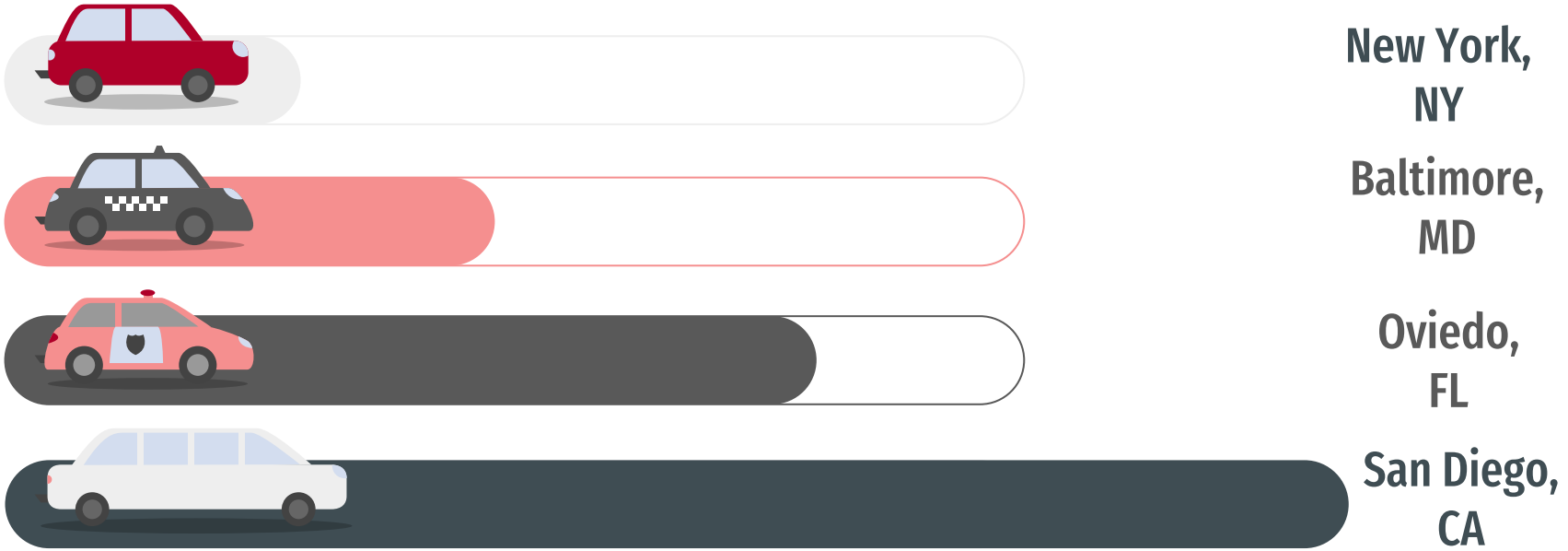


New car
made after 2015

Vehicle Ownership

Old cars have **6.0 times** the odds of new cars of filling a claim.

Understand the Coefficient



Area

San Diego has 1.66 times the odds of Oviedo area of filing a claim, than Baltimore, than New York.

Understand the Coefficient

X cases of speeding violation

(X-1) cases of speeding violation

Speeding Violation

With 1 case of extra in speeding violation, probability of filing a claim drop by 10%.

Bayes Theorem

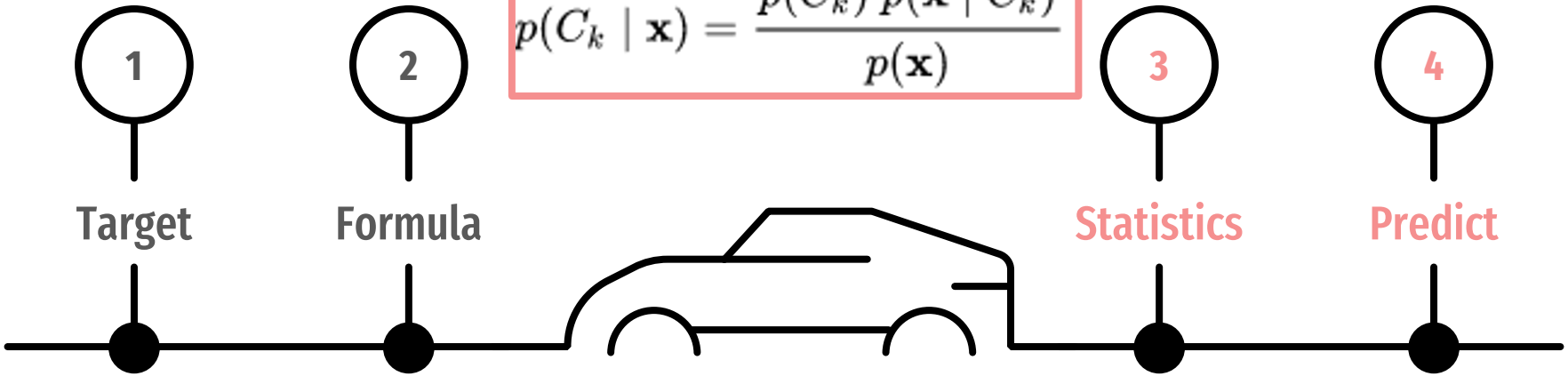
Conditional Probability

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

- C_k – The classification
- x – Sample Appearance

How it works?

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$



Prediction: $p(C_k| x)$

What we need to make prediction?

Probability is calculated based on statistic pattern

Applying the Bayes' theorem.