Health Insurance Cross Sell Prediction

Team 03
(Bruce) Chang-Hung Hou | Chenli Qiu | Lequn Yu | Qiqi Tang | Yuchen Feng | Shamika Kalwe

Problem Statement

To build a model to predict whether the existing health insurance customers will also be interested in Vehicle Insurance provided by the same company.

Cross Sell

Data Source

link

The kaggle link mentions **Analytics Vidhya** as its source for this dataset and problem. It also mentions relevant license for public sharing.

Structure of the Dataset (1/2)

Rows	**381,109**	**Columns**	**12**

No.	Variable	Definition
1	id	Unique ID for the customer
2	Gender	Gender of the customer
3	Age	Age of the customer
4	Driving_License	0 : No, 1 : Yes
5	Region_Code	Unique code for the region of the customer
6	Previously_Insured	0 : No, 1 : Yes
7	Vehicle_Age	Age of the Vehicle
8	Vehicle_Damage	0 : No, 1 : Yes (damaged in the past)
9	Annual_Premium	Health Insurance Premium per year
10	Policy*Sales*Channel	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
11	Vintage	Number of Days, Customer has been associated with the company
12	**Response**	**0 : Not Interested, 1 : Interested**

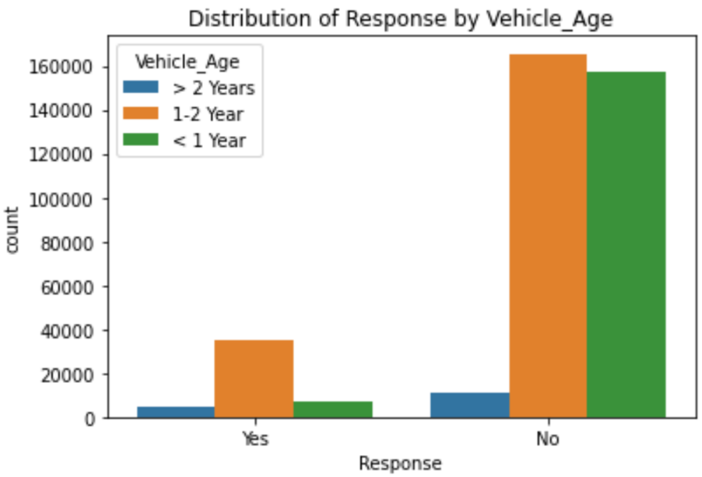
Target Variable

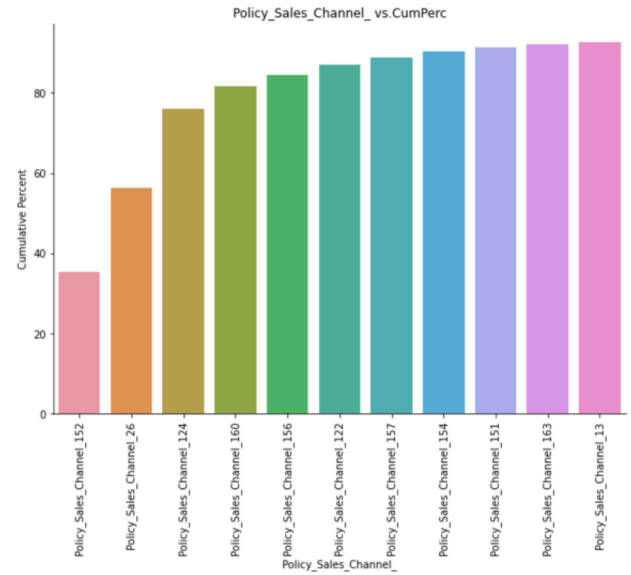
Stakeholders in the problem

- Company
  - Targeted marketing
  - Increase ticket size per customer
  - Save marketing cost
  - Know and understand their customers better
- Customer
  - Having multiple services from one company reduces hassel

Exploratory Analysis

Distribution of Response by Vehicle Age



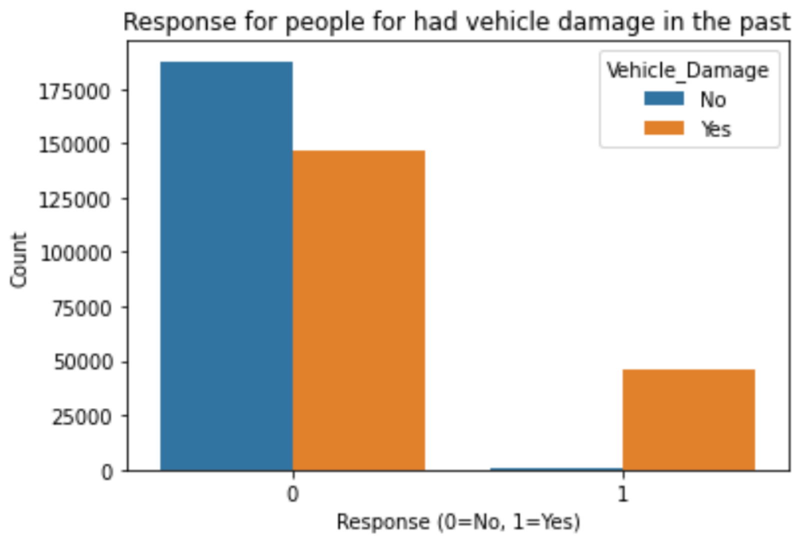


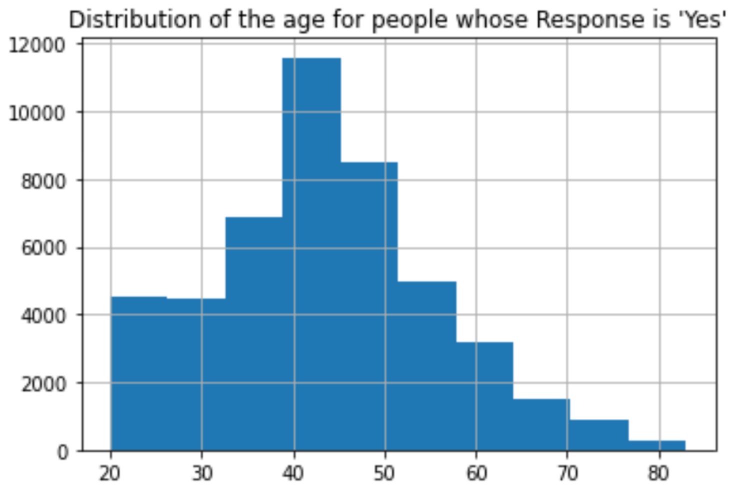
- New car owners have less preference to by an insurance

- 1-2 Year vehicle owners are more willing to buy car insurances

- Top 9 sales channel which customers prefer to buy insurances

Different Responses of People whether they have accidents

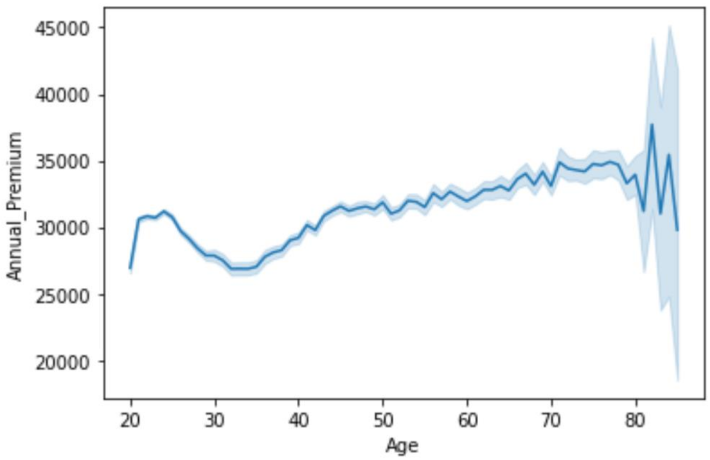


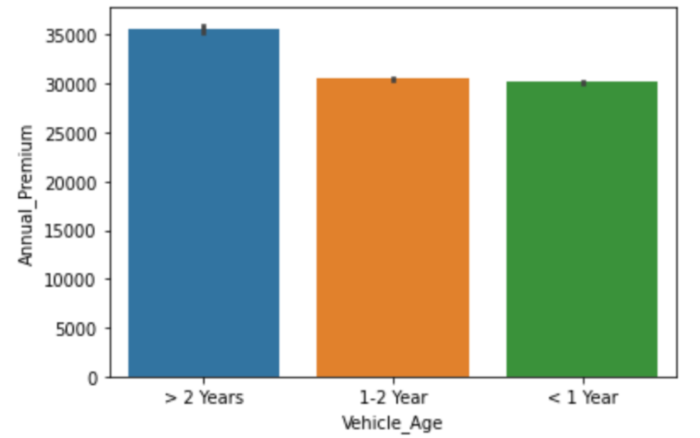


- People who do not have accidents in the past are unwilling to buy insurances
- People who have accidents in the past are willing to buy insurances

- Young people want to buy insurances may because of few driving experiences
- 40-50 years old people’s responses to purchase insurances are the strongest

The variances of annual premium





- Elderly people prefer to spend money on their insurances

- Vehicles which are used more than 2 years have the highest annual payments among other vehicles

Customer Profiling

Feature	**Target Variable = 0** ***Not Interested*** *in Vehicle Insurance (popular/mean value)*	**Target Variable = 1** ***Interested*** *in Vehicle Insurance (popular/mean value)*
Gender	Male	Male
Age	38 years	43 years
Driving_License	Present	Present
Region_Code	28	28
Previously_Insured	Yes	No
Vehicle_Age	1-2 Year	1-2 Year
Vehicle_Damage	No	Yes
Annual_Premium	INR 30,419	INR 31,604
Policy_Sales_Channel	152	26
Vintage	154	154

Highlighted are the differences

Modeling

- No sampling
- Down sampling
- SMOTE sampling

Lasso with vs. without SMOTE

Without SMOTE:

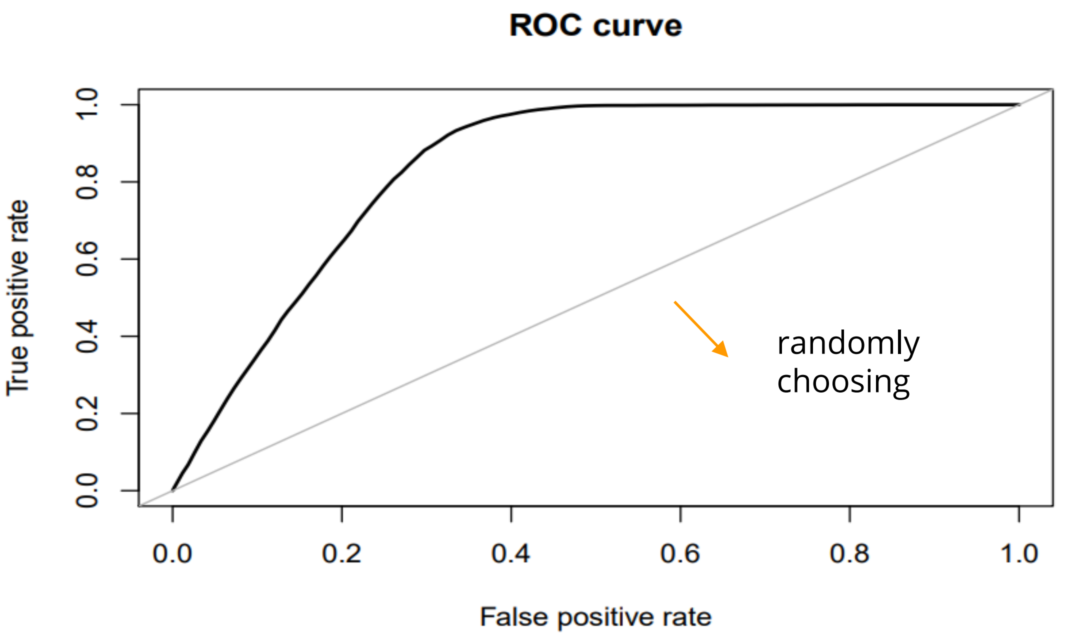
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 66959  9262
         1     1     0

               Accuracy : 0.8785
                 95% CI : (0.8761, 0.8808)
    No Information Rate : 0.8785
    P-Value [Acc > NIR] : 0.5072

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.000e+00
            Specificity : 1.000e+00
         Pos Pred Value : 0.000e+00
         Neg Pred Value : 8.785e-01
             Prevalence : 1.215e-01
         Detection Rate : 0.000e+00
   Detection Prevalence : 1.312e-05
      Balanced Accuracy : 5.000e-01

       'Positive' Class : 1
```

With SMOTE:

```
          Reference
Prediction     0     1
         0 46212   965
         1 20691  8390

               Accuracy : 0.716
                 95% CI : (0.7128, 0.7192)
    No Information Rate : 0.8773
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3081

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8968
            Specificity : 0.6907
         Pos Pred Value : 0.2885
         Neg Pred Value : 0.9795
             Prevalence : 0.1227
         Detection Rate : 0.1100
   Detection Prevalence : 0.3814
      Balanced Accuracy : 0.7938

       'Positive' Class : 1
```

Modeling: Lasso Regression

ROC curve :



Area under the curve(AUC) : 0.838

Ridge with vs. without SMOTE

Without SMOTE:

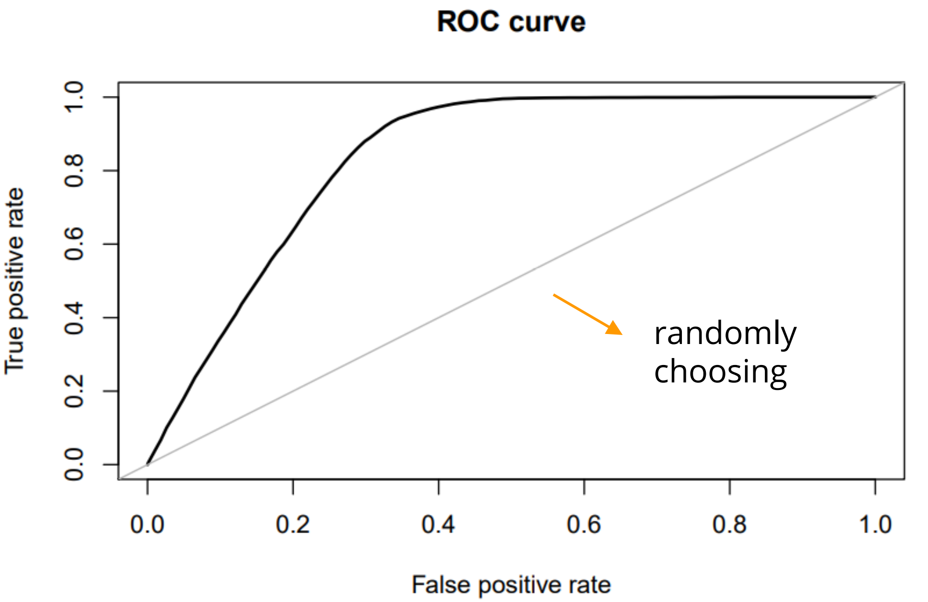
```
          Reference
Prediction     0     1
         0 66960  9262
         1     0     0

               Accuracy : 0.8785
                 95% CI : (0.8761, 0.8808)
    No Information Rate : 0.8785
    P-Value [Acc > NIR] : 0.5028

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.8785
             Prevalence : 0.1215
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 1
```

With SMOTE:

```
          Reference
Prediction     0     1
         0 46525  1050
         1 20378  8305

               Accuracy : 0.719

                 95% CI : (0.7158, 0.7222)
    No Information Rate : 0.8773
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3088

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8878
            Specificity : 0.6954
         Pos Pred Value : 0.2895
         Neg Pred Value : 0.9779
             Prevalence : 0.1227
         Detection Rate : 0.1089
   Detection Prevalence : 0.3761
      Balanced Accuracy : 0.7916

       'Positive' Class : 1
```

Modeling: Ridge Regression

ROC curve:



Area under the curve(AUC) : 0.837

Comparison: Lasso vs. Ridge

	Lasso	Ridge
Sensitivity	89.68%	88.78%
Area under the curve(AUC)	83.8%	83.7%

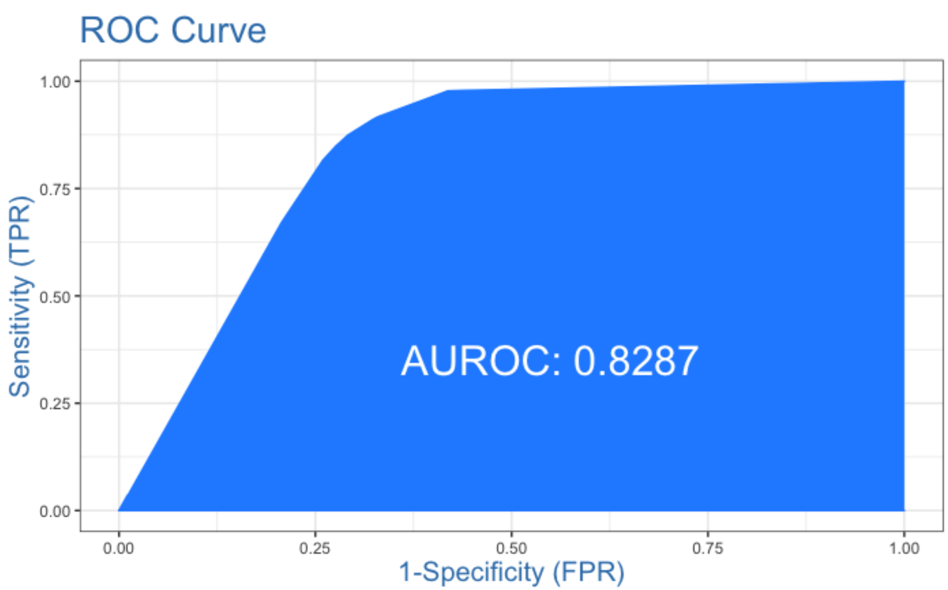
Modeling: Decision Tree (rpart)

Before balancing the data

```
Confusion matrix:
                Actual

Predicted    0       1
         0 67001   9221
         1     0      0
```

After balancing the data

```
Confusion matrix:
                Actual
Predicted       0       1
          0   48335   1412
          1   18439   7932
Accuracy Overall (%): 73.92075
Sensitivity (1) (%):  84.8887
Specificity (0) (%):  72.38596
```



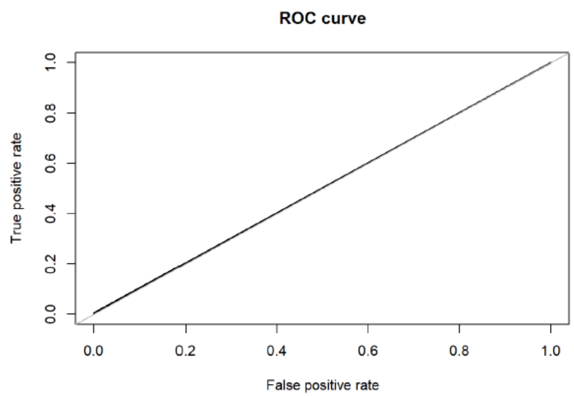
Feature importance:

```
Vehicle_Damage
Previously_Insured
Age
Policy_Sales_Channel
Vehicle_Age
Region_Code
Annual_Premium
Driving_License
Gender
```

Modeling: Random Forest

No Sampling

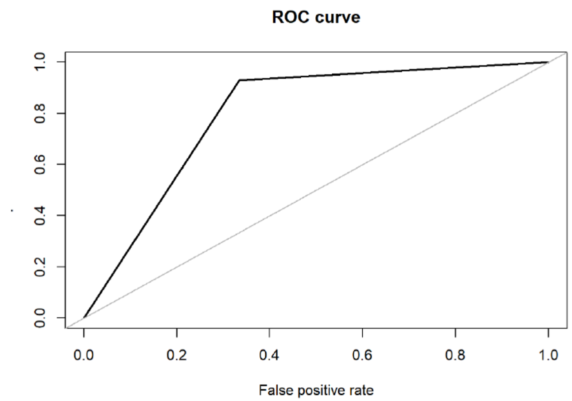
Accuracy: 87.78%
Specificity 0: 99.93%
Sensitivity 1: 0.42%
AUC: 0.502




```
Confusion matrix:
       0    1   class.error
0 234022 175 0.0007472342
1  32434 145 0.9955492802
```

Down Sampling

Accuracy: 79.84%
Specificity 0: 66.6%
Sensitivity 1: 93.08%
AUC: 0.798

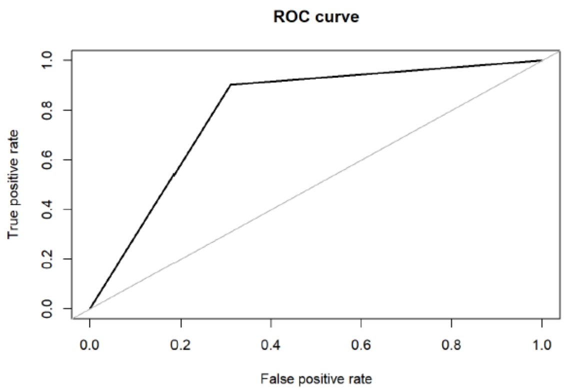



```
Confusion matrix:
      0     1 class.error
0 21686 10893  0.33435649
1  2256 30323  0.06924706
```

SMOTE

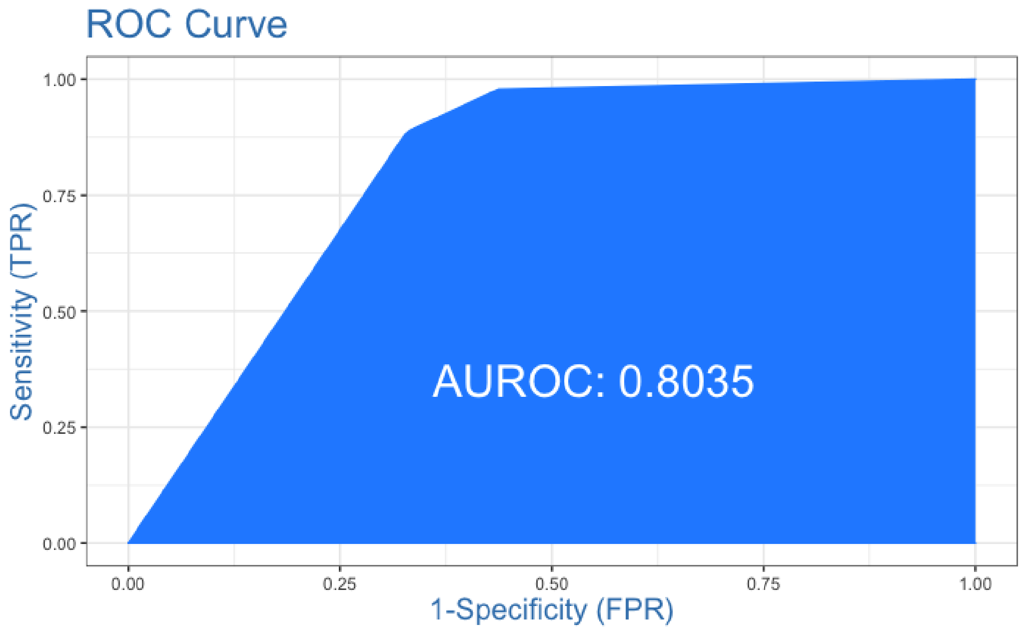
BEST Model of ALL

Accuracy: 82.26%
Specificity 0: 70.24%
Sensitivity 1: 94.29%
AUC: 0.797




```
Confusion matrix:
      0     1 class.error
0 45764 19394  0.29764572
1  3721 61437  0.05710734
```

Modeling: Boosting with SMOTE



Confusion matrix:

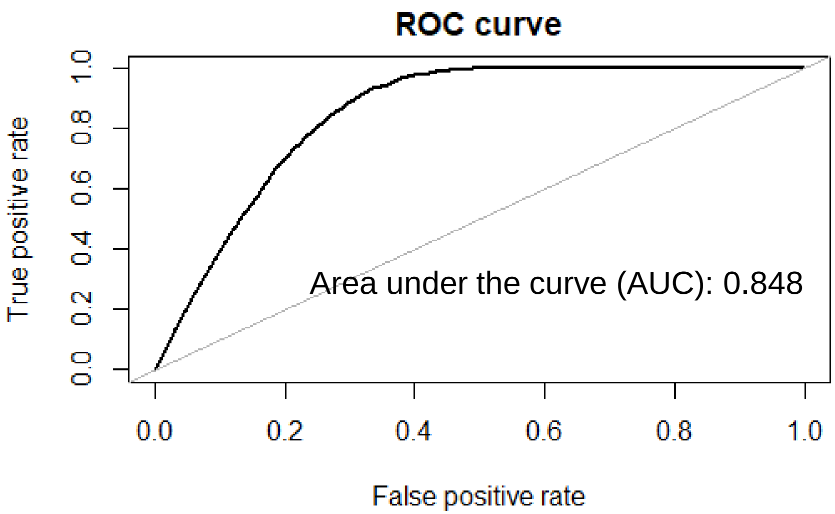
Predicted		Actual 0	Actual 1
	0	44961	1114
	1	21813	8230

Accuracy Overall (%)
69.87966

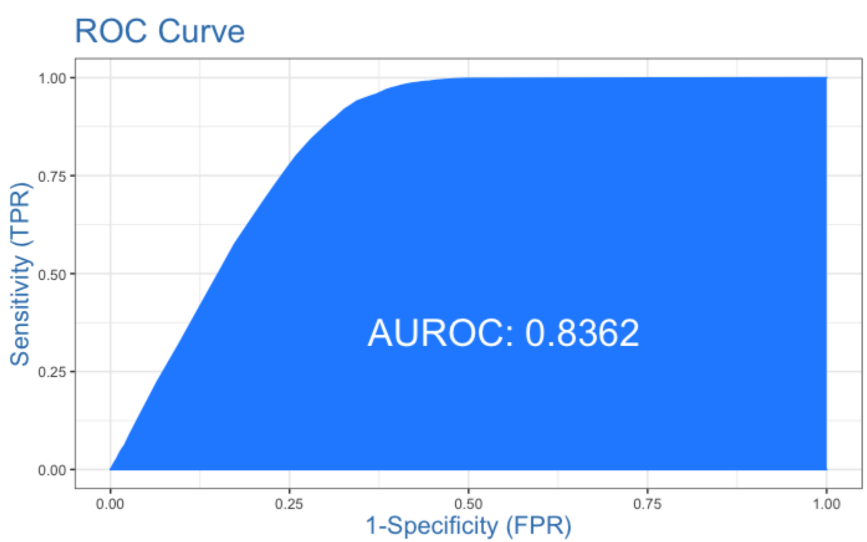
Sensitivity (1) (%)
88.07791

Specificity (0) (%)
67.33309

Logistic without & with SMOTE




```
Confusion matrix:
              Actual
Predicted      0      1
        0  66929   9256
        1     31      6
Accuracy Overall (%): 87.82
Sensitivity (1) (%): 00.064781
Specificity (0) (%): 99.953704
```




```
Confusion matrix:
              Actual
Predicted      0      1
        0  49473   1870
        1  17291   7474
Accuracy Overall (%): 74.82393
Sensitivity (1) (%): 79.98716
Specificity (0) (%): 74.10131
```

Challenges and Possible Solutions

Area	Challenges	Solutions
Dataset	The size of the train dataset was too high, thus modeling was time consuming	Using methods like SMOTE we were able to reduce the size and remove the imbalance
	Dataset was highly imbalanced (88%-12%)	
	Policy_Sales_Channel and Region_Code were categorical variables with many categories (150+, 50+ respectively) which resulted in many columns after one-hot encoding. This resulted in poor model performance	We converted the categorical variables into 'factor' class and then fed the data to the models
Modeling	Coding in R, as we had little experience with it	
	Difficulty in choosing models as Y was categorical	We decided to use: DT, RF, Boosting, Linear, lasso, ridge, Log R

Conclusions

- From EDA we observe some key differences between the two profiles of customers
- Imbalance can disturb modeling a lot. So depending on how important is to correctly classify 1 or 0, the imbalance must be treated.
- Removing imbalance in the data and using **Random Forest** model gives us the most decent model, with Sensitivity of 95% (overall accuracy: 82%, Specificity: 70%).
- Top 5 features are:
  - Vehicle_Damage
  - Previously_Insured
  - Age
  - Annual_Premium
  - Policy_Sales_Channel

Thank you!

Any Questions?