Project 5:

Health Insurance Cross Sell Prediction

Group 9: Xiran Lin

Overview

- Project Introduction
- Data Description
- EDA and Visualization
- Model Selection
- Unfair Data
- Model Testing
- Conclusion
- Q&A

Project Introduction:

Purpose: An Insurance company that has provided Health Insurance to its customers now they need your help in building a model to predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company.

Given Information: To predict, whether the customer would be interested in Vehicle insurance, you have information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc.

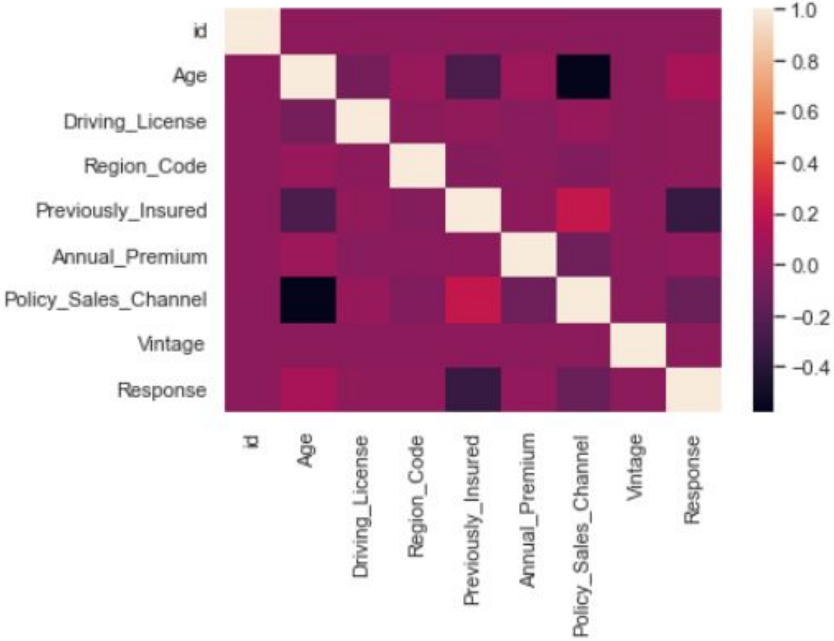
Data Description:
Train Set vs Test Set

Variable	Definition
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Driving_License	0 : Customer does not have DL, 1 : Customer already has DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
Vehicle_Age	Age of the Vehicle
Vehicle_Damage	1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
Annual_Premium	The amount customer needs to pay as premium in the year
Policy*Sales*Channel	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
Vintage	Number of Days, Customer has been associated with the company
Response	1 : Customer is interested, 0 : Customer is not interested

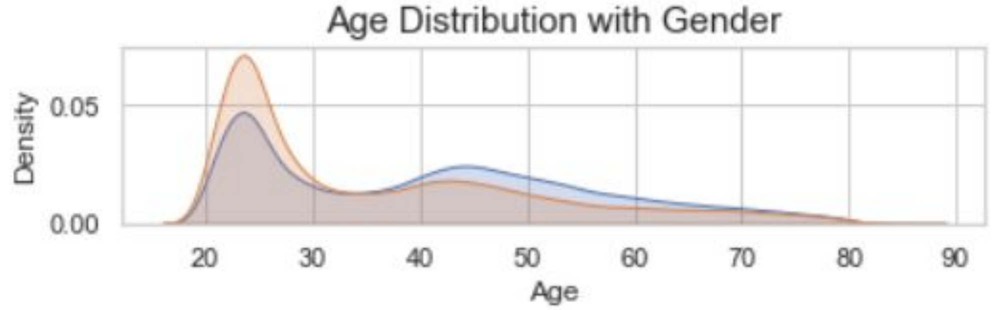
Exploratory Data Analysis (EDA) & Visualization

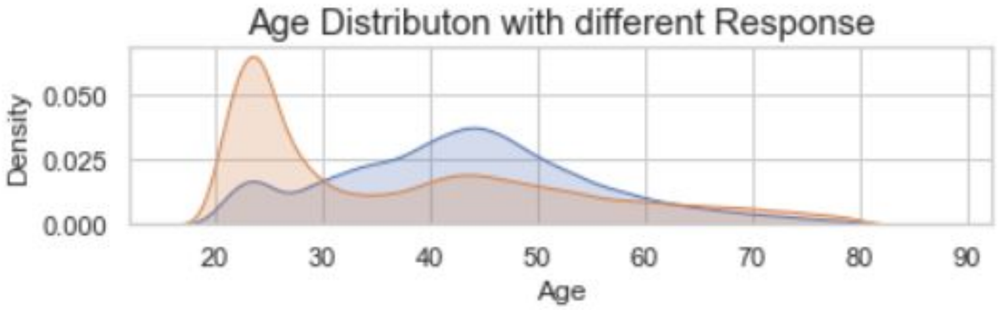
Checking for Missing Values

```
id                      0
Gender                  0
Age                     0
Driving_License         0
Region_Code             0
Previously_Insured      0
Vehicle_Age             0
Vehicle_Damage          0
Annual_Premium          0
Policy_Sales_Channel    0
Vintage                 0
Response                0
dtype: int64
```



EDA and Visualization

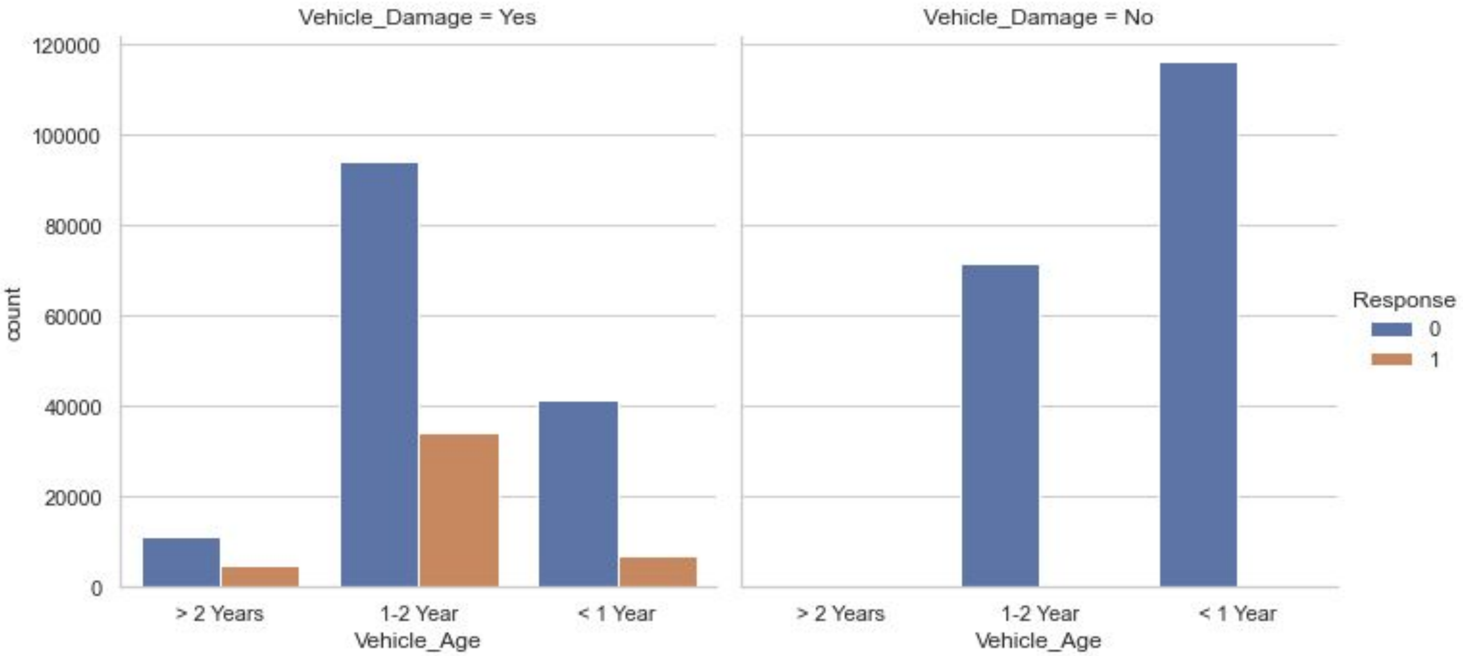




Red: Male
Blue: Female

Red: interest
Blue: no interest

EDA and Visualization



Model Selection

According to the previous analysis, this problem can be identified as Binary Classification, that is whether customers will be interested in Vehicle Insurance.

And we have a dataset with 300,000+ record, which means we are unlikely to choose SVM Classifier to train due to too much time.

So we will most likely make a selection from following models:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Gradient Boost

Unfair data

Negative response data accounts for 87.84%

(1 - 46710/334399) = 87.84%

Prepare balanced data

```
df_response0 = preprocessed_data[preprocessed_data['Response']==0]
df_response1 = preprocessed_data[preprocessed_data['Response']==1]

print(f'Number of Response 0: {len(df_response0)}')
print(f'Number of Response 1: {len(df_response1)}')
```

```
Number of Response 0: 334399
Number of Response 1: 46710
```

Models' Accuracy

Confusion Matrix

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

Accuracy:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy of Logistic Regression: 0.7740847784200385

Accuracy of Decision Tree: 0.7045600513808606

Accuracy of Random Forest: 0.76043673731535

Accuracy of Gradient Boosting: 0.792763862128024

Models' Recall

Confusion Matrix

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

Recall: $$R = \frac{TP}{TP + FN}$$

Recall of Logistic Regression: 0.90656

Recall of Decision Tree: 0.6936533333333333

Recall of Random Forest: 0.8293333333333334

Recall of Gradient Boosting: 0.9293866666666667

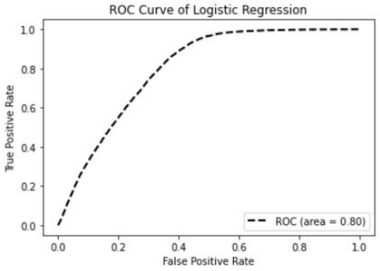
ROC & AUC of Models

Confusion Matrix

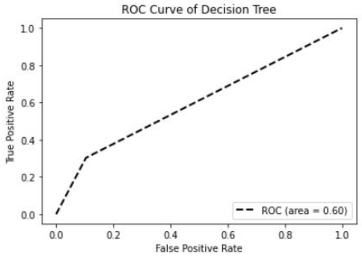
		Predicted class		
		− or Null	+ or Non-null	Total
True	− or Null	True Neg. (TN)	False Pos. (FP)	N
class	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

Name	Definition
False Pos. rate	FP/N
True Pos. rate	TP/P
Pos. Pred. value	TP/P*
Neg. Pred. value	TN/N*

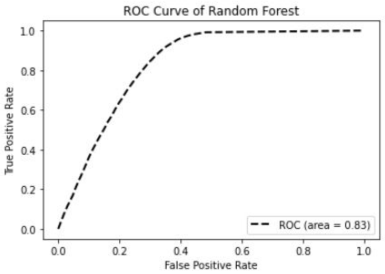
AUC of Logistic Regression: 0.7979603626176262



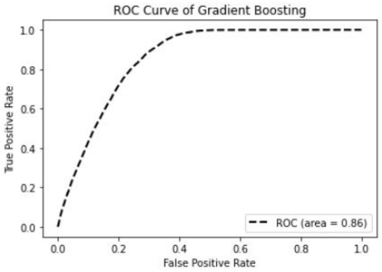
AUC of Decision Tree: 0.5992201328588924



AUC of Random Forest: 0.8307982668315834



AUC of Gradient Boosting: 0.855610387216959



Conclusion

1. Our goal: Response prediction
2. EDA: Decide models
3. Data Cleaning
4. Model selection
5. Model testing

Thank you for your attention !