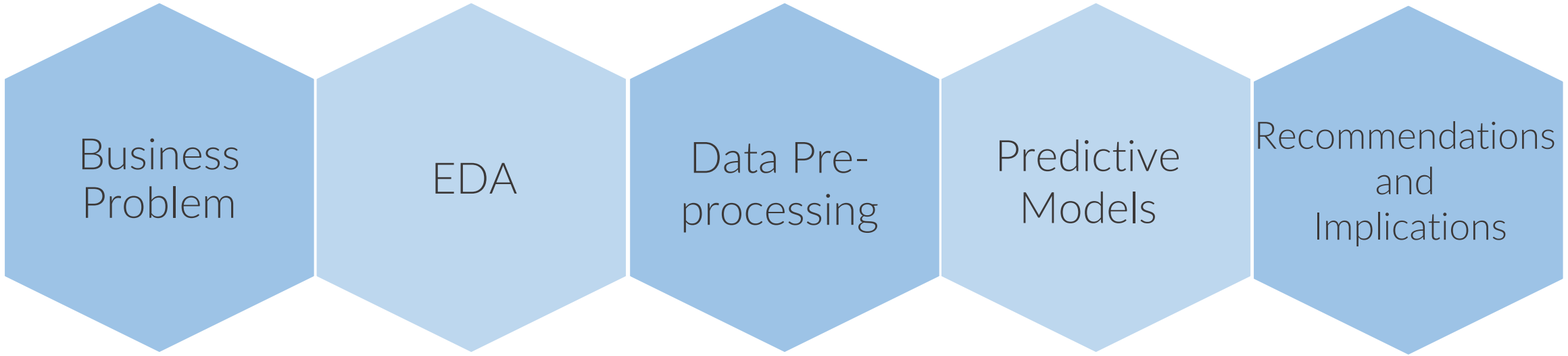
Health and Car Insurance Cross Sell prediction 📈

Contents


Business Problem
EDA
Data Pre-processing
Predictive Models
Recommendations and Implications

Motivations

From companies' side:

Help the targeted insurance company to better develop strategies for:

- Marketing communications
- Resource allocation
- Insurance pricing
- Customer's loyalty

Other insurance companies could also utilize the analytical models and results.

From customers' side:

- Customers interested in vehicle insurance can receive well-tailored information or potentially better deals.
- Customers falling outside the targeted group will receive general marketing emails or calls concerning vehicle insurance from the firm.

Business Problem

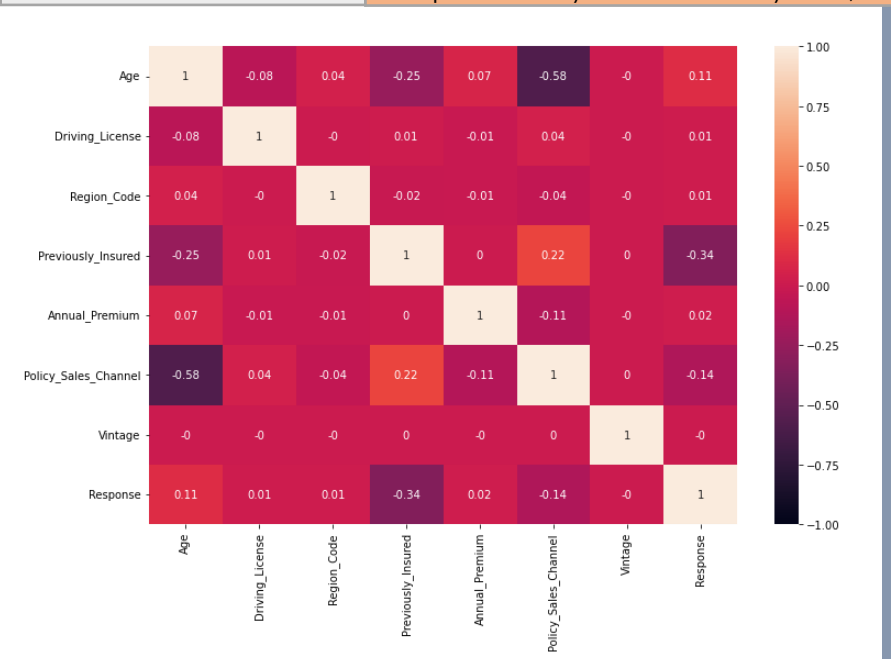
Whether a policyholder of the company's health insurance would be interested in buying vehicle insurance

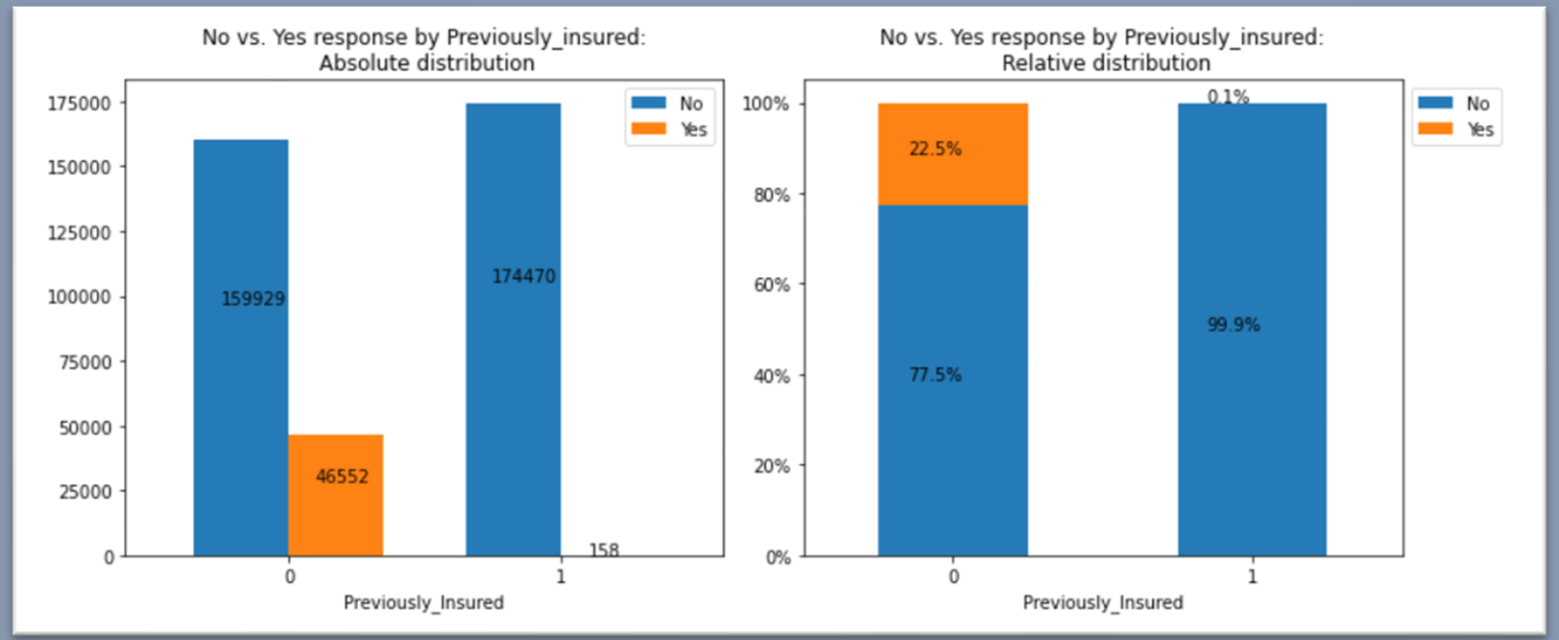
Possible Hypotheses before Exploring the data:

1. Customers who already have their cars insured by other insurance companies in the past, might be interested in updating/ changing vehicle insurance to this company (variable: Previously_Insured)
2. How much money the customer is now paying for insurance might affect their decision (variable: Annual_Premium)
3. Customers who have old vehicles might be interested (variable: vehicle age)
4. Customers who owns driving licence could be interested (variable: Driving_License)
5. Customers who have their vehicle damaged in the past might be interested (variable: vehicle damaged)

Data Description

#	Column	Non-Null Count (381109)	Dtype	
0	id	non-null	int64	Unique ID for the customer
1	Gender	non-null	object	Gender of the customer
2	Age	non-null	int64	Age of the customer
3	Driving_License	non-null	int64	0 : Customer doesn't have DL 1 : Customer already has DL
4	Region_Code	non-null	float64	Unique code for the region of the customer
5	Previously_Insured (*other company)	non-null	int64	1 : Customer already has Vehicle Insurance 0 : Customer doesn't have Vehicle Insurance
6	Vehicle_Age	non-null	object	Age of the Vehicle
7	Vehicle_Damage	non-null	object	1 : Customer got vehicle damaged in the past. 0 : Customer didn't get vehicle damaged in the past
8	Annual_Premium	non-null	float64	The amount customer needs to pay as premium annually
9	Policy_Sales_Channel	non-null	float64	Anonymized Code for the channel of outreaching to the customer
10	Vintage	non-null	int64	Number of Days, Customer has been associated with the company
10	Response	non-null	int64	1 : Customer is interested 0 : Customer is not interested





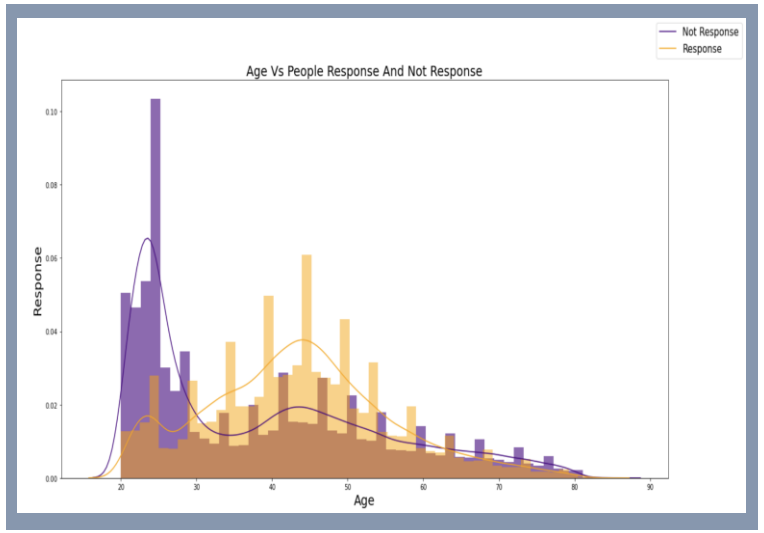
Observing the above heatmap, “Previously_insured” is one of the most important features, since it has the highest negative correlation with “response”.

Heat Map

Exploratory Data Analysis (EDA)

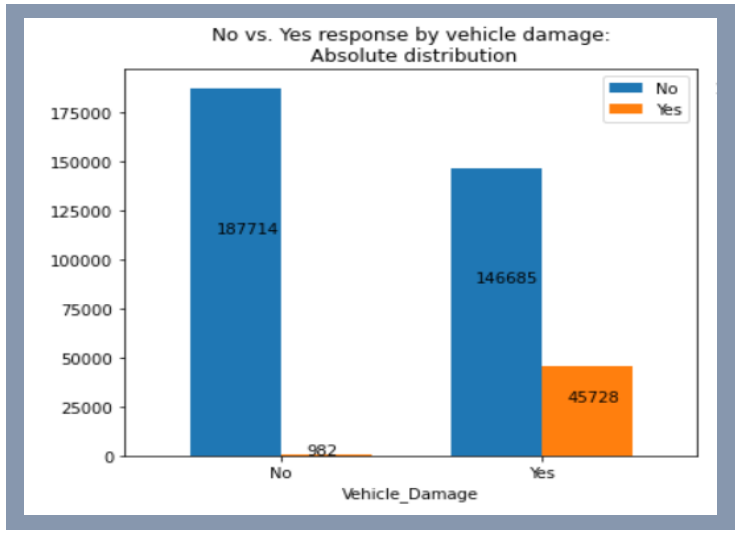
Customers who were not previously hold the car insurance issued by other company are more interested to our car insurance.

Bar Chart



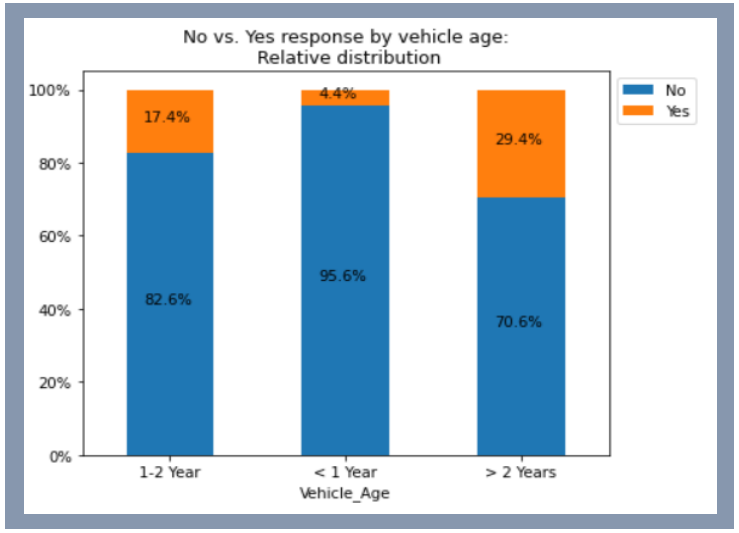
Response v.s. Age

Customers whose ages are between 30~60 are more likely to response. Customer whose ages are between 20~30 are less likely to response.



Response v.s. Vehicle Damage

Customers who experienced vehicle damage before are more willing to response.

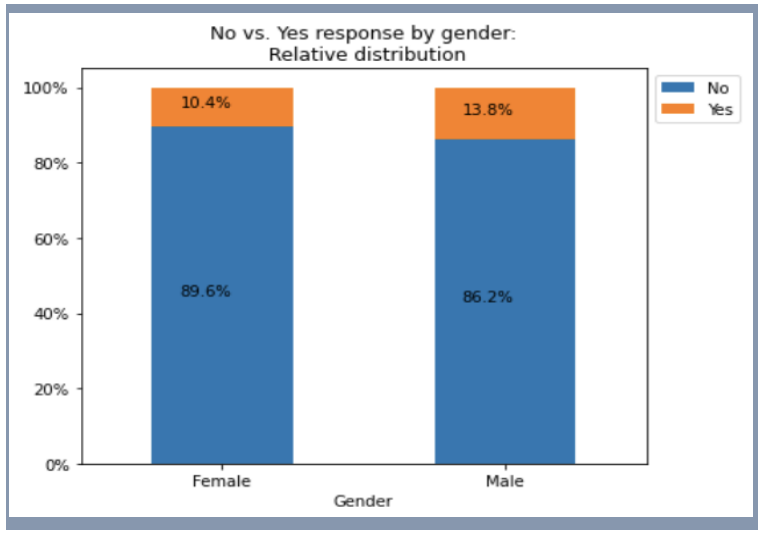


Response v.s. Vehicle Age

The older a vehicle is, the more likely its owner is to response.

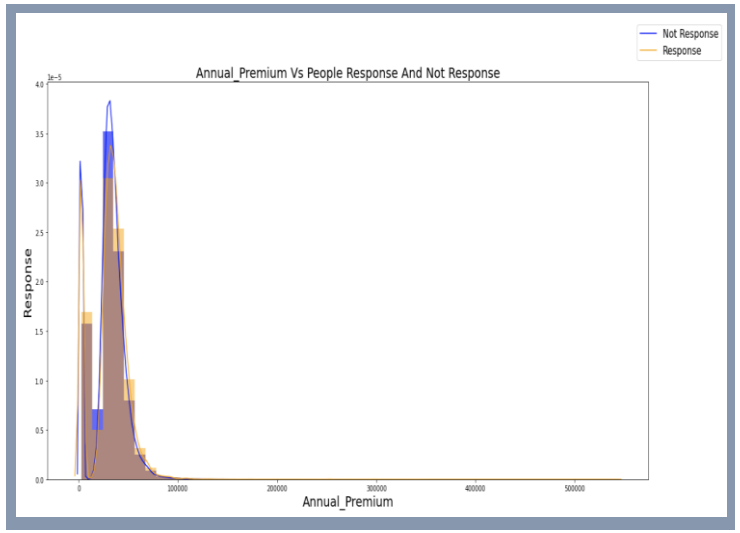
From the result of heatmap, we found that “Age” and “Vehicle_Damage” and “Vehicle_Age” should also have a high probability to be correlated with “response” somehow. So, we decided to explore more about their data distribution pattern here.

Exploratory Data Analysis (EDA)



Response v.s. Gender

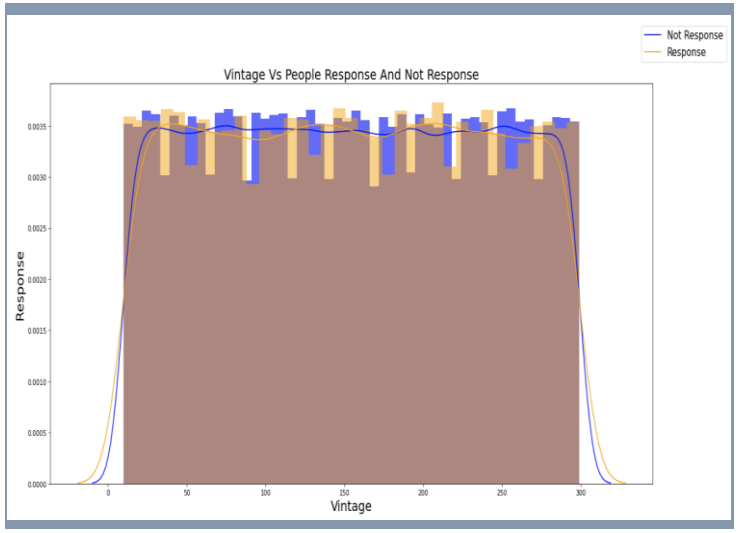
Male customers are more willing to response our car insurance than female customers.



Response v.s. Annual Premium

Annual_Premium has a long-tailed distribution, with a lot of observations on the low value end, and few on the high value end.

We did not find clear clues to explain the distribution of Annual_Premium.



Response v.s. Vintage

We did not find clear clues to explain the distribution of Vintage.

Other than those features suggesting significant correlations, we also explore the association patters for other features, including “Gender” and “Annual Premium” and “Vintage”. We found Male is more likely to response than female, but the difference is not that significant. In addition, unfortunately, we did not find any clue showing that annual premium and vintage would influence the customers’ interest in our car insurance.

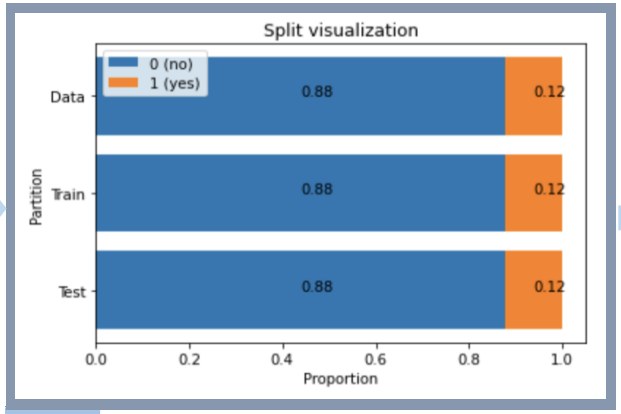
Exploratory Data Analysis (EDA)

Dropped Columns
ID
Driving_License
Region_Code
Policy_Sales_Channel

Drop variables

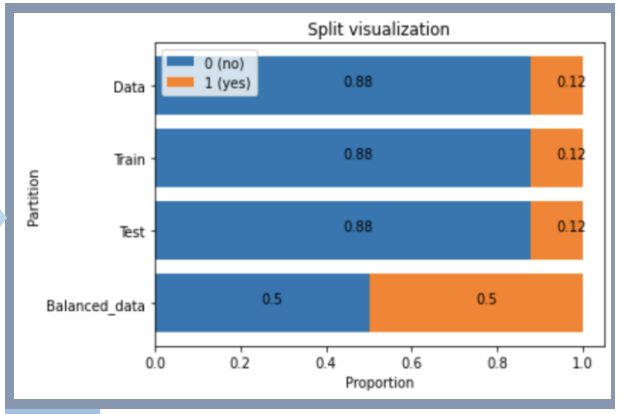
Gender_Male	Vehicle_Age_< 1 Year	Vehicle_Age_> 2 Years	Vehicle_Damage_Yes
1	0	1	1
1	0	0	0
1	0	1	1
1	1	0	0
0	1	0	0

Encode categorical variables



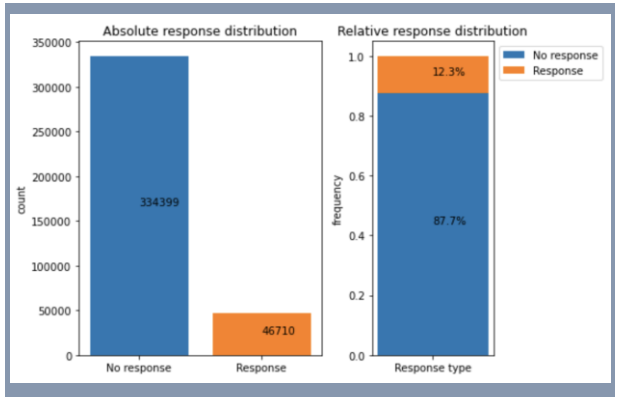
Data Splitting

Data was split into train and test set for modeling by the ratio 7:3.

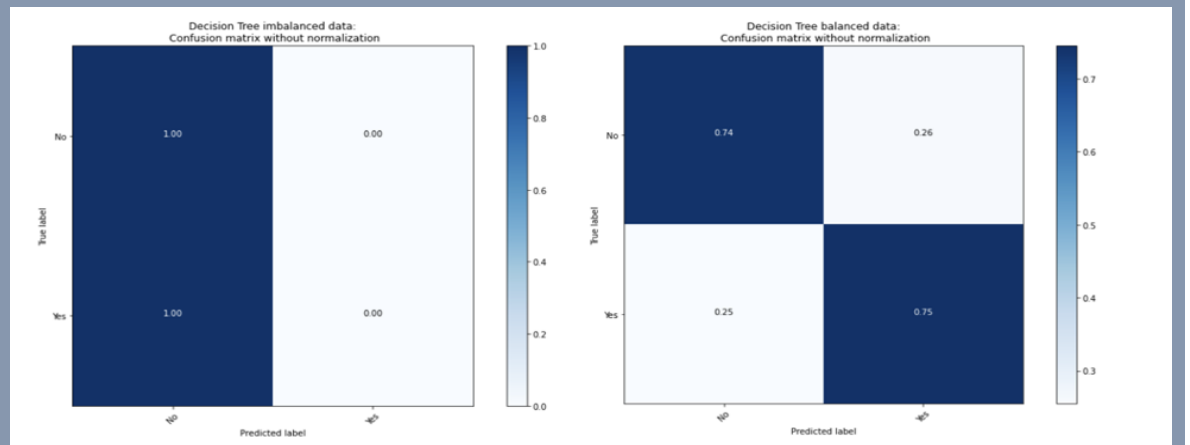


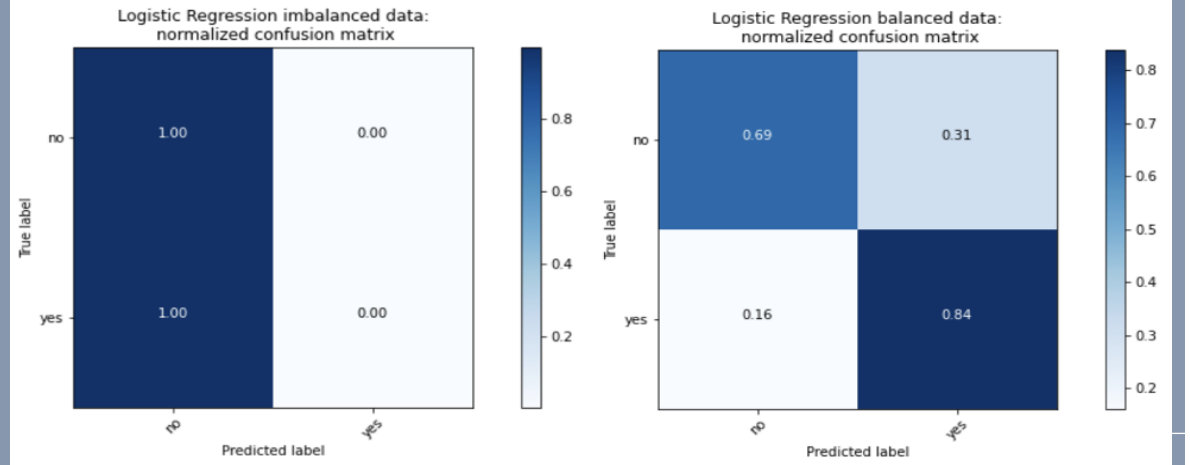
Data Balancing (SMOTE)

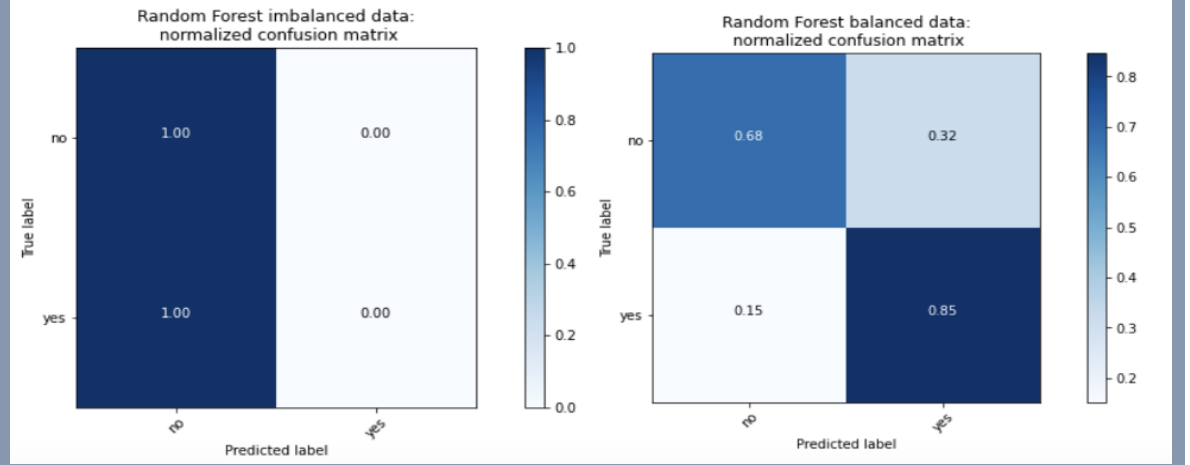
Dropped variables	Explanation for Dropping Reasons
ID	No explanatory power
Driving_License	It's quite obvious assumption that people who don't have driving license would not be interested in vehicle insurance, Also, we can see from the data that almost all of respondent has a driving license.
Region_code & **Policy_Sales_Channel**	If using one hot encoding, it will add 208 new columns to dataset. Our group also considered some solutions (e.g divide 'Region_code' values into groups like 'northern' or 'southern'). However, due to limited background information about the company and these features, we don't have any evidence to combine these numbers with "meaningful" staffs. Dropping them is the optimal choice.



The original data was highly imbalanced (88% not interested, 12% interested). Thus, we decided to balance the data using SMOTE







Decision Tree

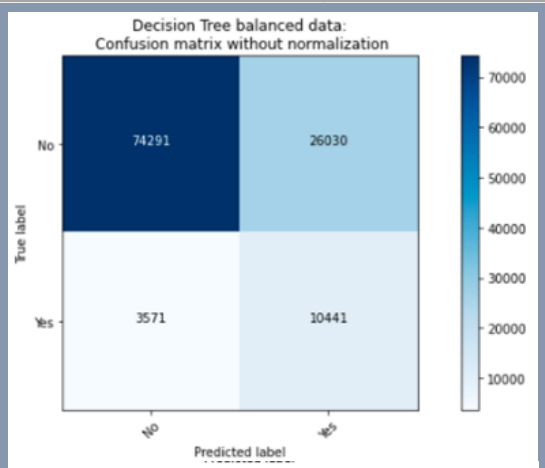
Logistic Regression

Random Forest

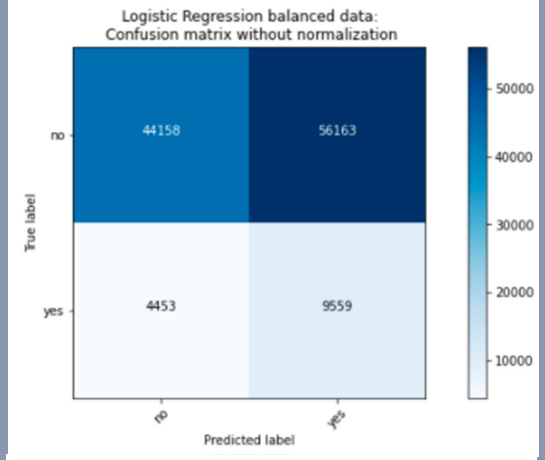
Confusion Matrix

After balancing the data, the models have a better classification performance.

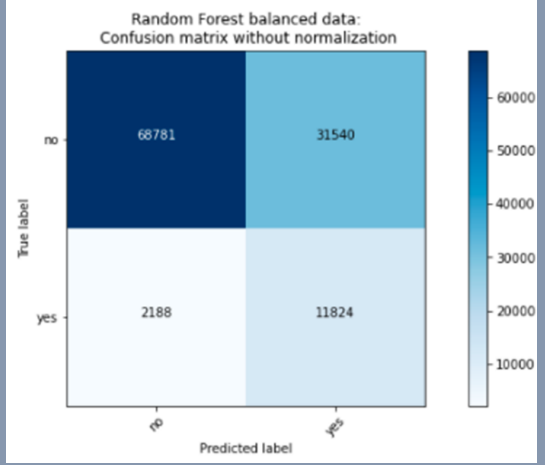
- With imbalanced data, the correctness of "Interested" predictions is 0 in all models, so it is not that useful in practice.
- With balanced data, even though the correct prediction of "Not interested" predictions decreases a little bit, models can predict around 80% "Interested" correctly
- Only taking balanced data into account in the project later



Decision Tree



Logistic Regression



Random Forest

Cost-Benefit analysis

Assumptions:

- Marketing cost per customer =200$
- Revenue gained from a vehicle insurance sold = 2000$
- Customers who are correctly predicted to be interest, will actually buy insurance

Therefore:

Profit of correctly predicted = 2000-200=800$
The cost of incorrectly predicted (predicted yes but actually no) =200$

Results:

		Actual Class	
		p	n
Predicted Class	Y	$ 800	$(200)
	N	$ -	$ -

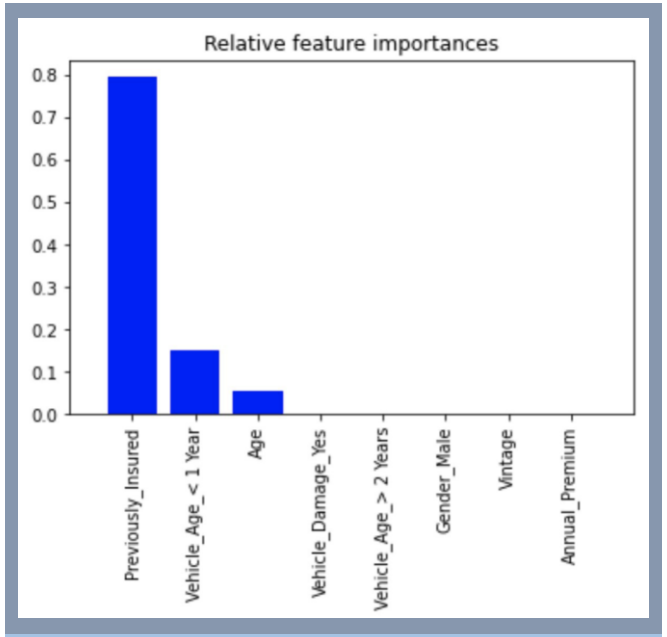
Expected benefit per customer:	
Decision tree	$ 27.52
Logistics Regression	$ (31.36)
Random forest	$ 27.56

- ❑ The performance of Decision Tree and Random Forest are pretty similar.
- ❑ But the Logistics Regression seems like not a good choice.

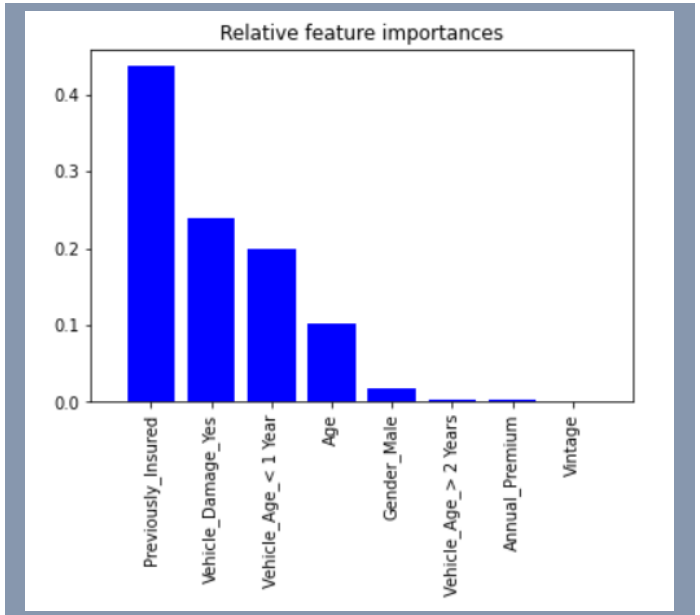
Feature importance (Balanced Data)

	variable	coefficient
7	Vehicle_Damage_Yes	2.520000
2	Annual_Premium	0.000000
3	Vintage	0.000000
0	Age	-0.020000
6	Vehicle_Age_> 2 Years	-0.070000
4	Gender_Male	-0.850000
5	Vehicle_Age_< 1 Year	-2.090000
1	Previously_Insured	-2.860000

Logistic Regression



Decision Tree



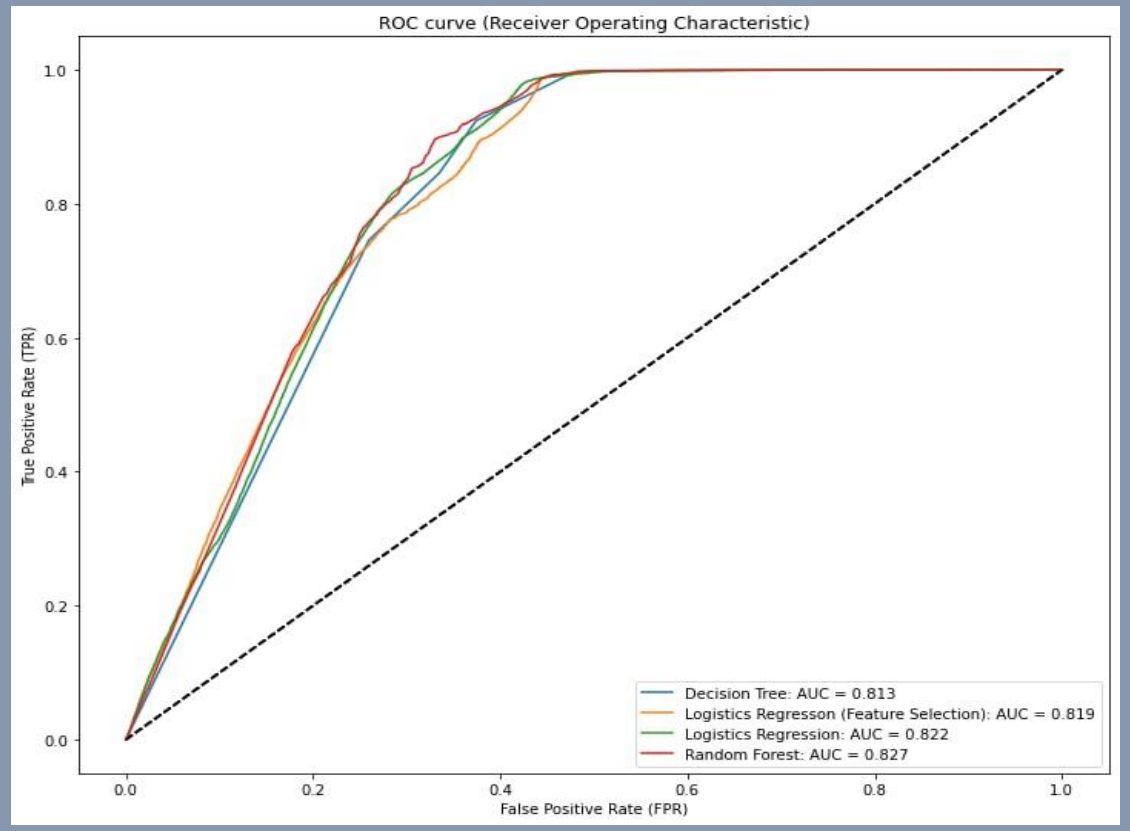
Random Forest

The relative importance of features varies, but the difference is minor:

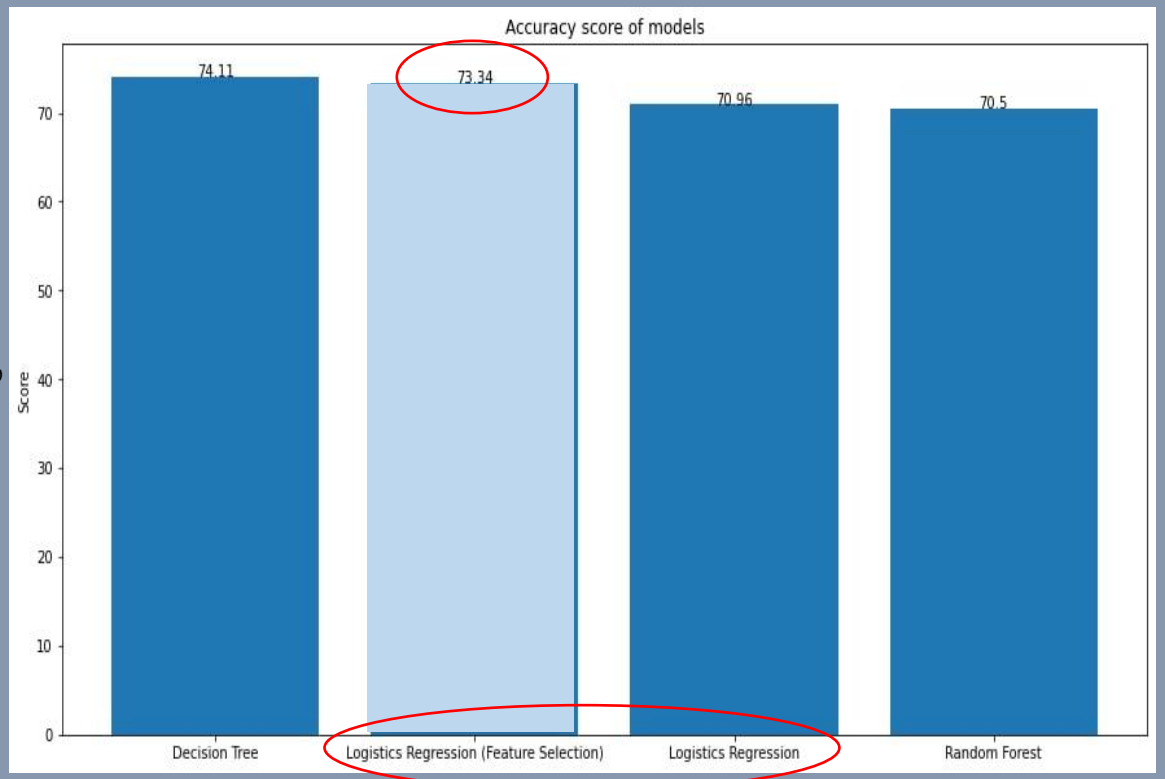
- For both Logistic Regression and Random Forest: "Previously_insured" , "Vehicle_Damage", and “Vehicle_Age” are the top 3 significant features.
- For Decision Tree: “Previously_insured" and “Vehicle_Age” are the top 2 significant features, but ”Age” is the 3rd significant feature.

Conclusions:

- We can basically conclude that the 3 main factors on customer's response are: "Vehicle_Damage", "Previously_insured" and "Vehicle_Age", which are coherent with our previous EDA findings.
- The attributes "Annual_Premium" and "Vintage" are less significant in our response prediction.



"Feature Selection"



After seeing the result of feature importance shown by all three models, we found that "Annual_Premium" and "Vintage" are two less significant features for the predictive models.

We decided to re-run the logistic regression according to the result, to see whether dropping those two features could improve the behaviour of the logistic regression.

The accuracy of our logistic regression model is improved.

Model Selection

We evaluate our model performance from three indicators: Accuracy, AUC score and Gini index.

Although the Random Forest has highest AUC & Gini, the difference between its scores and other's is not too large.

So, **Decision Tree** is still the optimal option by its highest accuracy rate and simplicity in result interpretation.

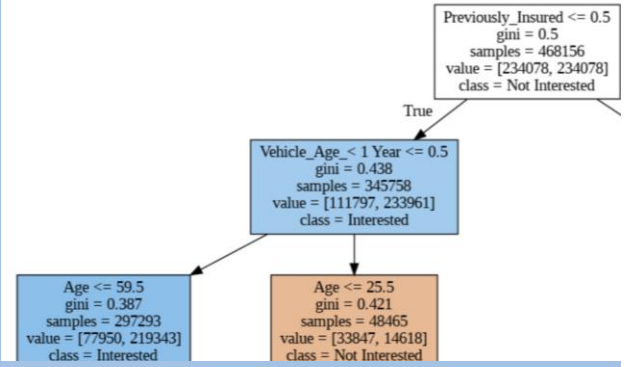
Models	Accuracy	AUC	Gini
Decision Tree	74.11%	0.813	0.626
Logistics Regression (Feature Selection)	73.34%	0.819	0.638
Logistics Regression	70.96%	0.822	0.644
Random Forest	70.50%	0.827	0.654

Recommendations:

The recommendations could be expanded to other health insurance companies who also want to expand their car insurance business.

Marketing of vehicle insurance should target customers with the following three attributes (from result of Decision Tree Model):

- ❑ Have not been previously insured,
- ❑ Have vehicle age more than 1 year and,
- ❑ Age is lower than 60



Based on above results, the company can:

- ❑ Utilize social media (effectively attracts attention from young consumers)
- ❑ Cooperate with vehicle repair shops (as they can assess conditions of vehicle, then help promote new products on right customer group)

Implications:

The Value and Limitation of our Works.

- ❑ The primary value proposition of the findings is to maximize marketing return of investment by precisely identifying interested customers.

- ❑ Accurate prediction and high hit rate are extremely important for an insurance company which utilizing premiums to compensate the claims – that's one reason why we recommend decision tree model.

- ❑ The limitation of our analysis is that we ignore the inexplainable variables based on limited background about the company, such as variables Region_code and Policy_Sales_Channel, but the model can be updated easily after getting specific instruction or additional information on variables.