Capstone Project- Classification

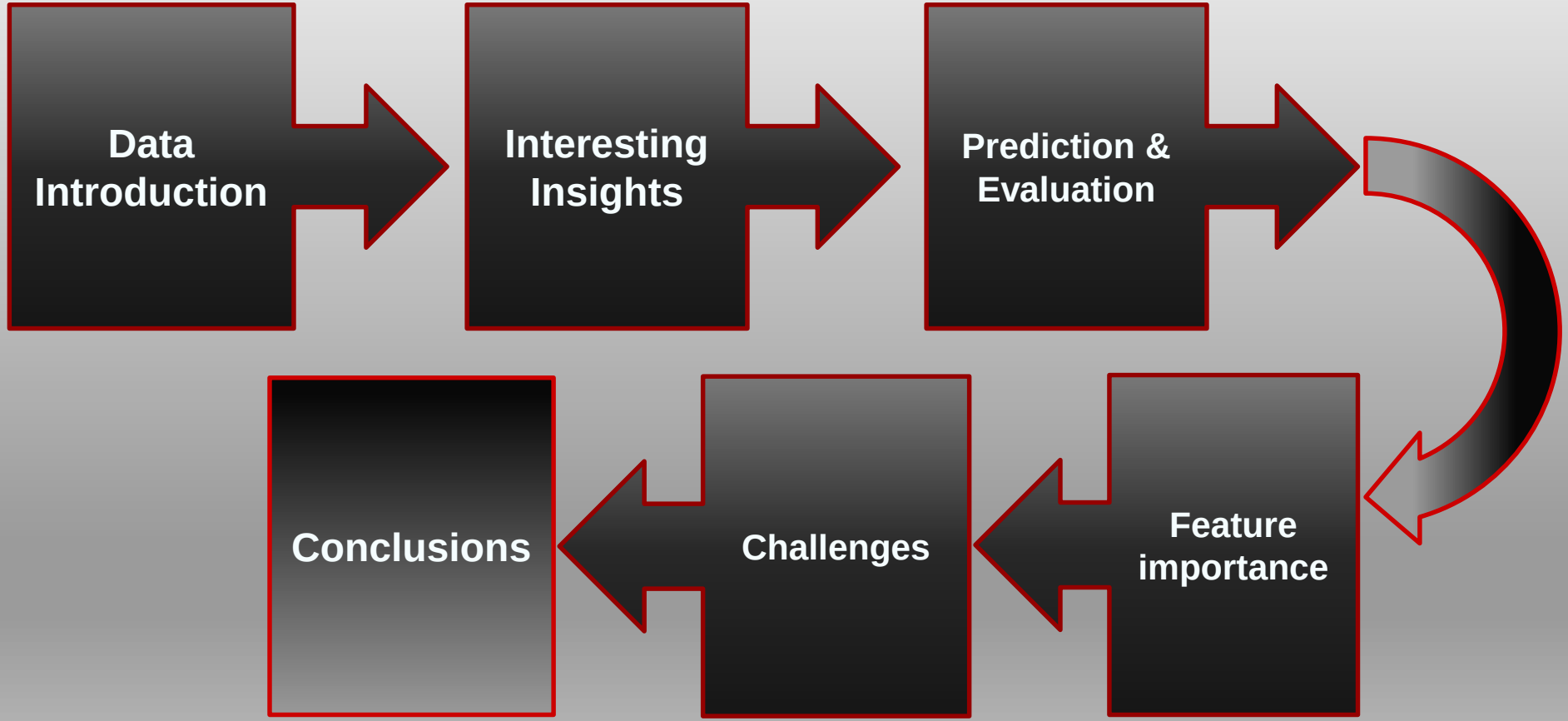
Insurance cross sell prediction

Presented By :-
Mohd Sharik

Problem Statement

Our client is 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.

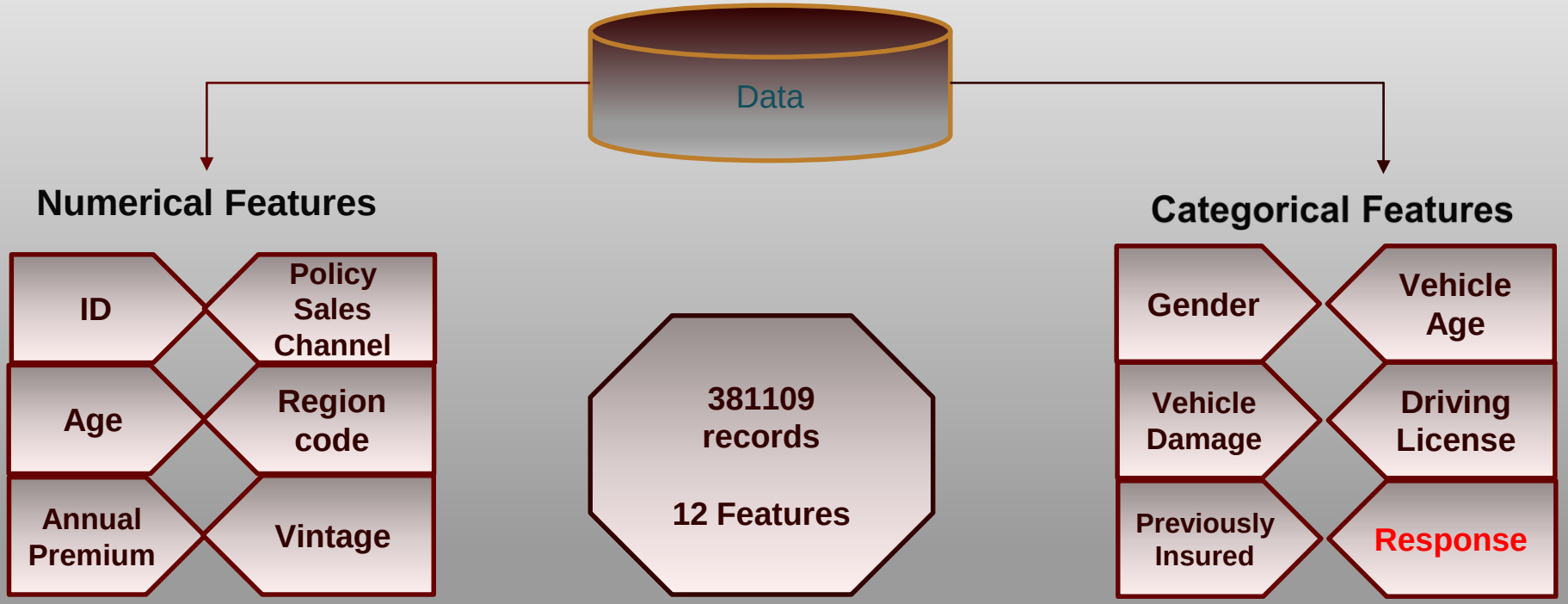
Overview


Data Introduction
Interesting Insights
Prediction & Evaluation
Feature importance
Challenges
Conclusions

Data Introduction

Column	Description
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 one.
Vehicle Age	Age of the Vehicle
Vehicle Damage	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 i.e. 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

Data Introduction



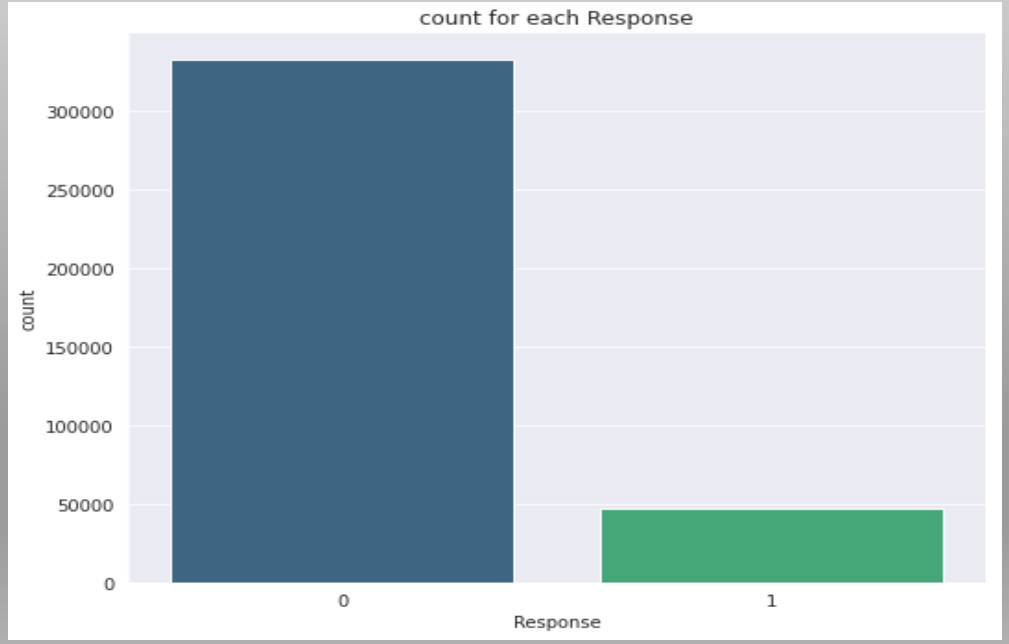
Interesting insights:

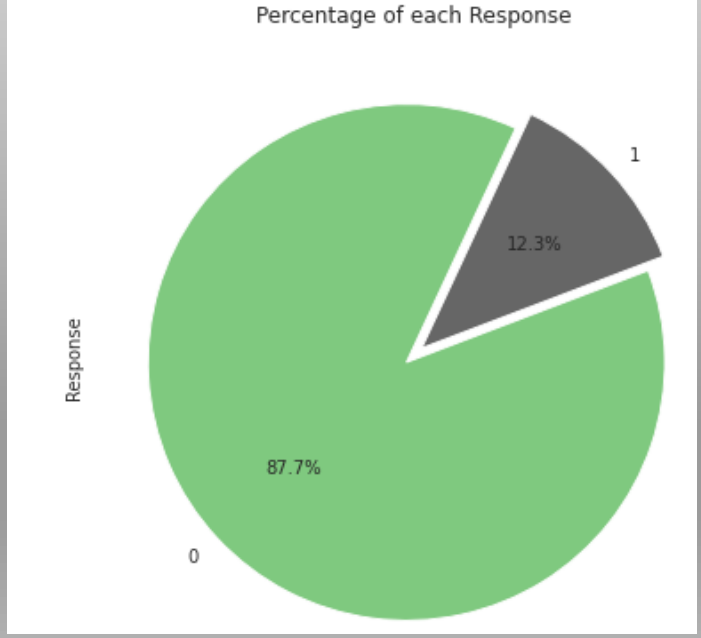
Draw insights and take Business decisions

Overall conversion rate is 12.3%

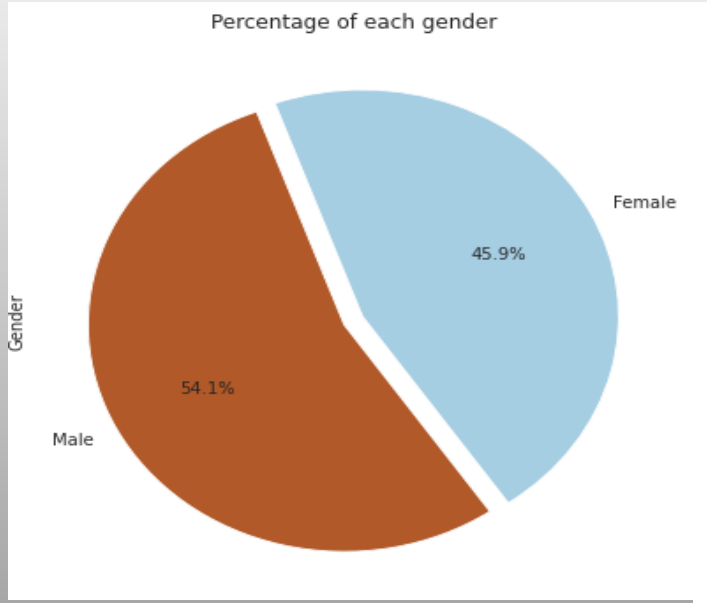
381109 people are contacted/approached and 46582 are interested in vehicle insurance.

Response	Counts
0	333107
1	46582

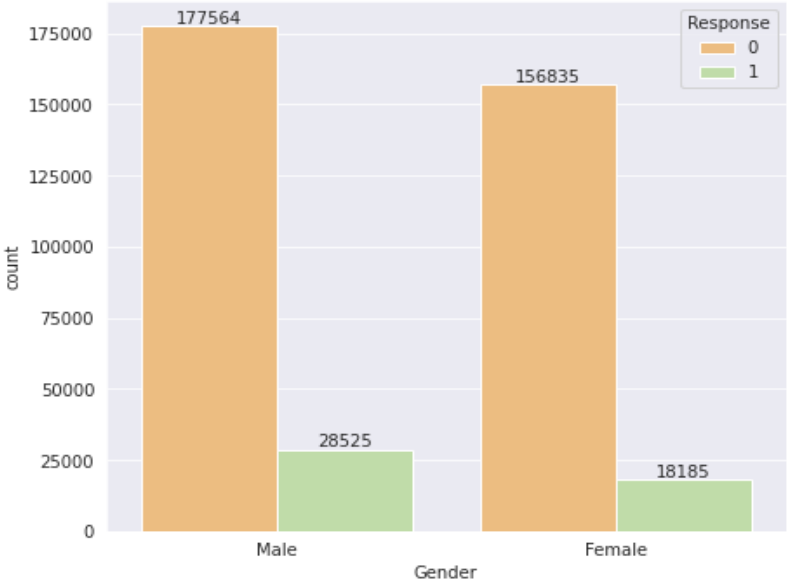




Men's give higher conversion

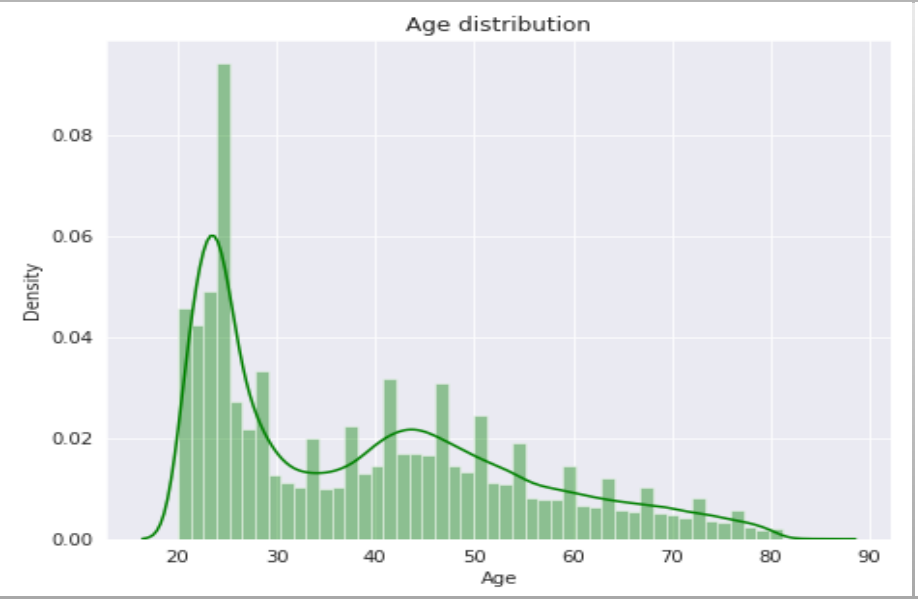


Gender	Percent conversion
Male	13.84%
Female	10.39%



Observing the given data shows that interest rate is comparatively high for Male gender

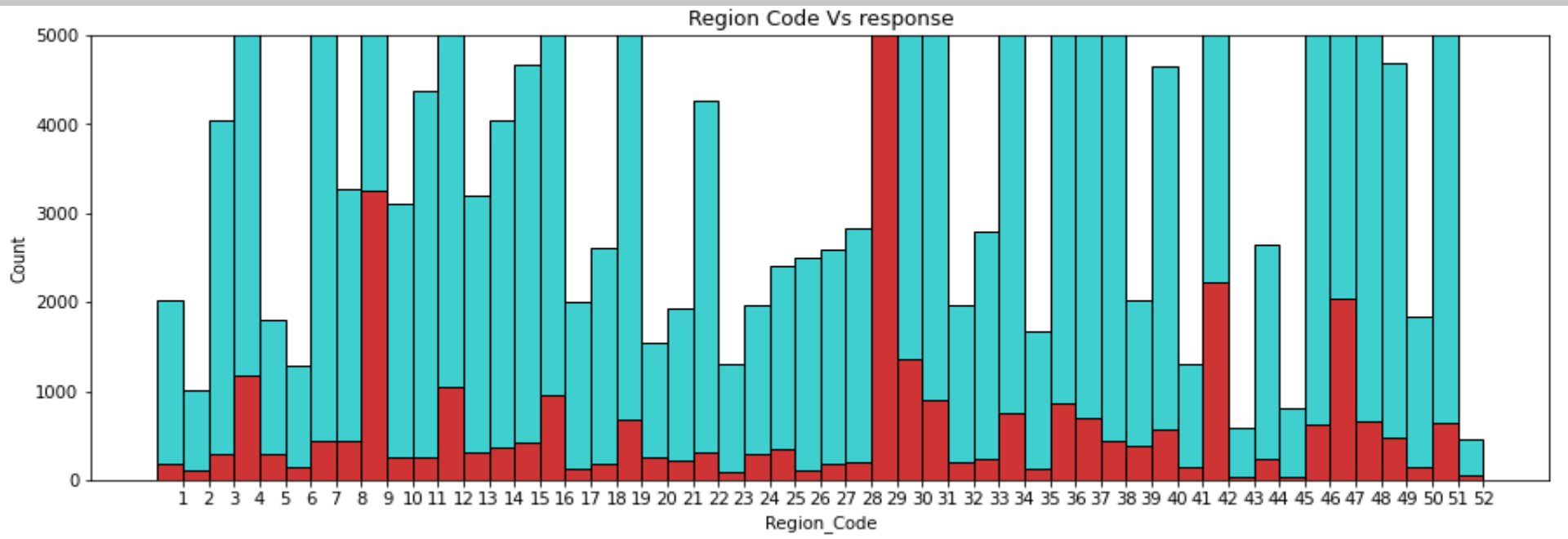
Target more people having age range 30-55 years



It has been observed that more customers are targeted having age range 23-30 years but we should target customers having age 30-55 years.

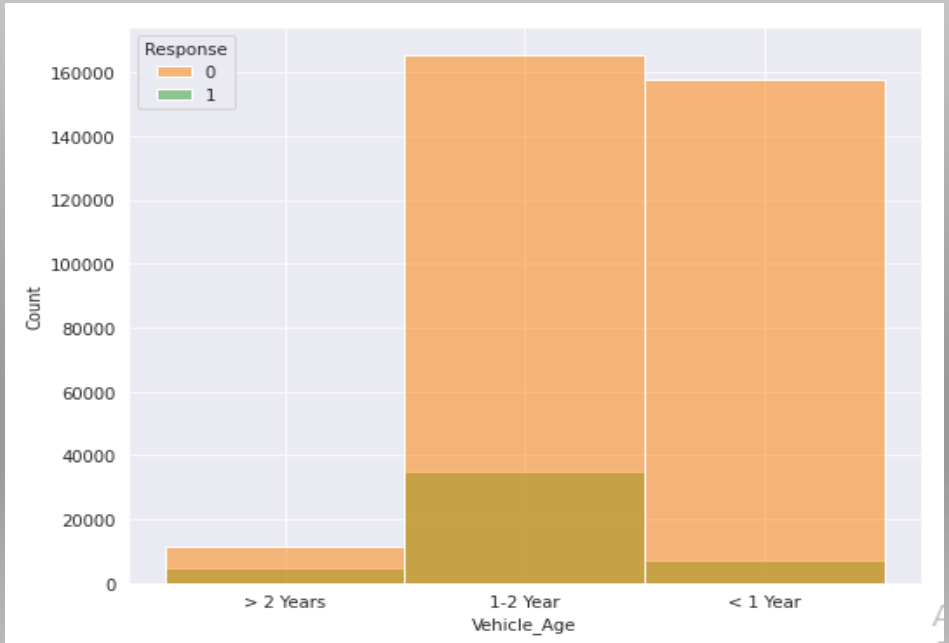
Region 29 has given us highest conversions

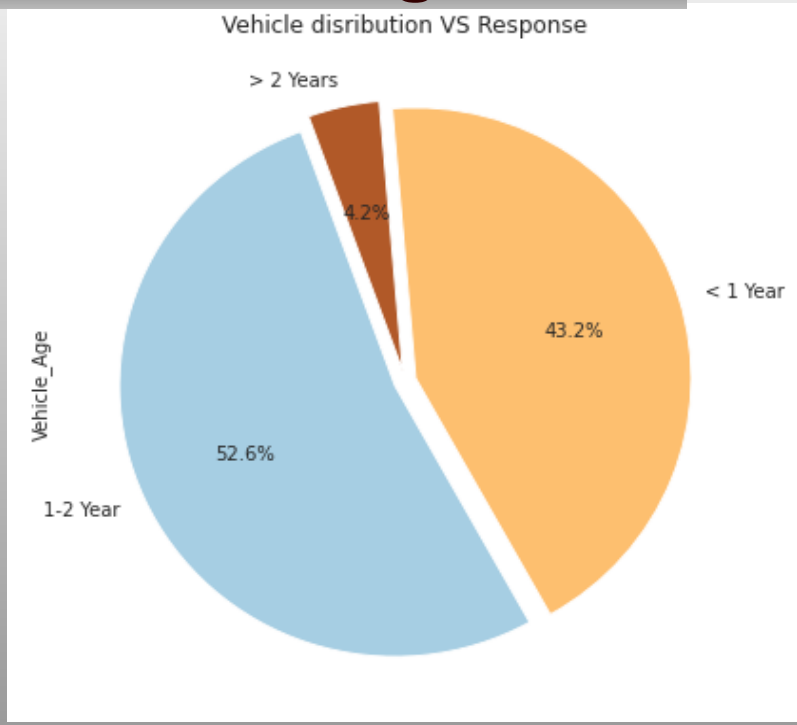
We have seen that people are more interested which are from Region having code 29 followed by region 9 and Region 42



Poor conversion from “< 1year” vehicle age

We analyzed that although 43% of records are coming from vehicle age “< 1year” but the interest rate is not very satisfying

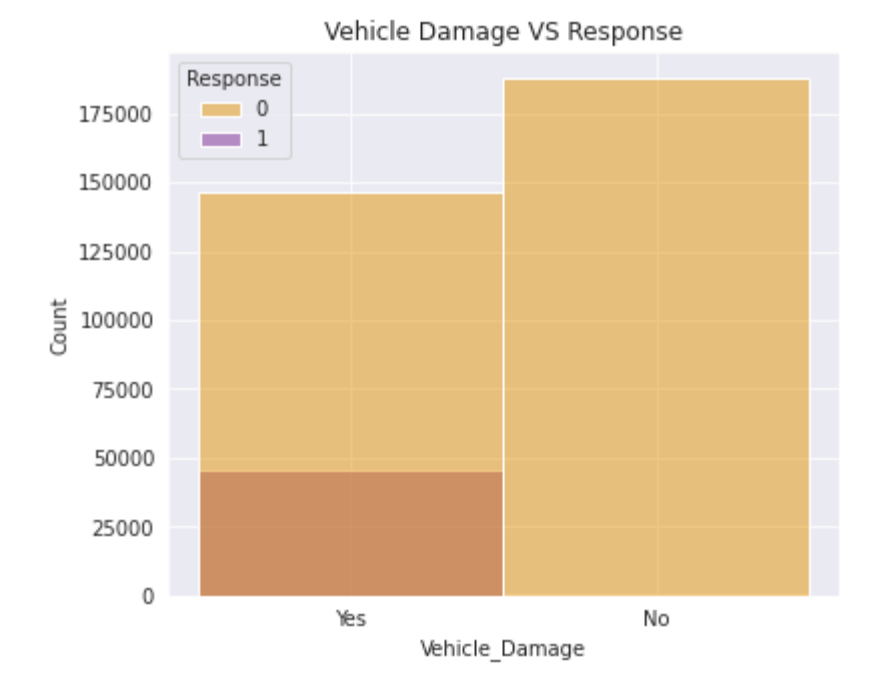




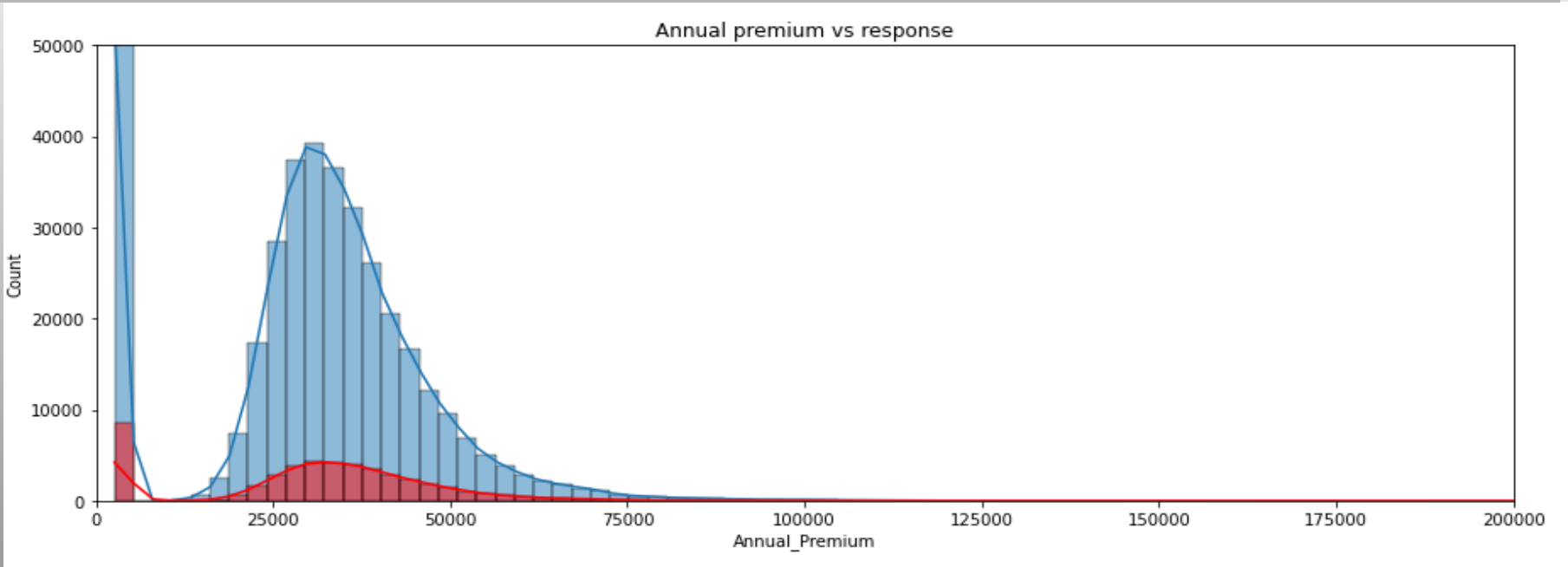
Attract people having no damage vehicle

	Vehicle_Damage
Yes	192413
No	188696

As per the data we have seen that if a customer has no vehicle damage then they are not interested. Hence organization should make some strategy to attract no-damage vehicle



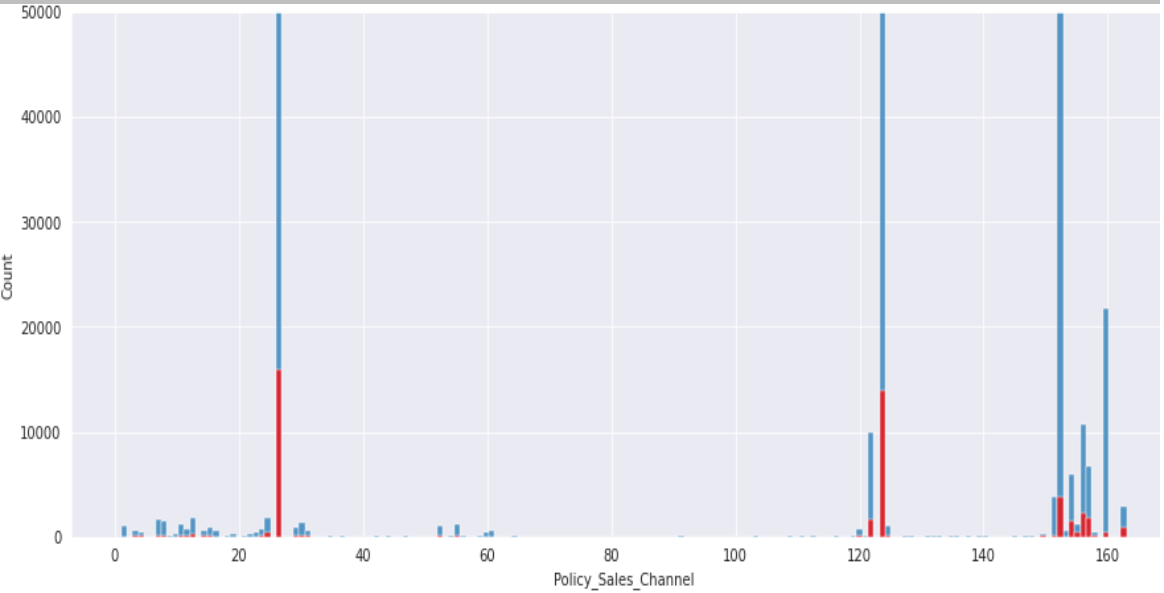
Less records having > 75000 premium and hence less +ve responses



We Observed that there are very less people who were offered higher premium insurance and hence no. of positive responses also less. Same can be seen in the adjacent plot.

Top 10 Sales channels

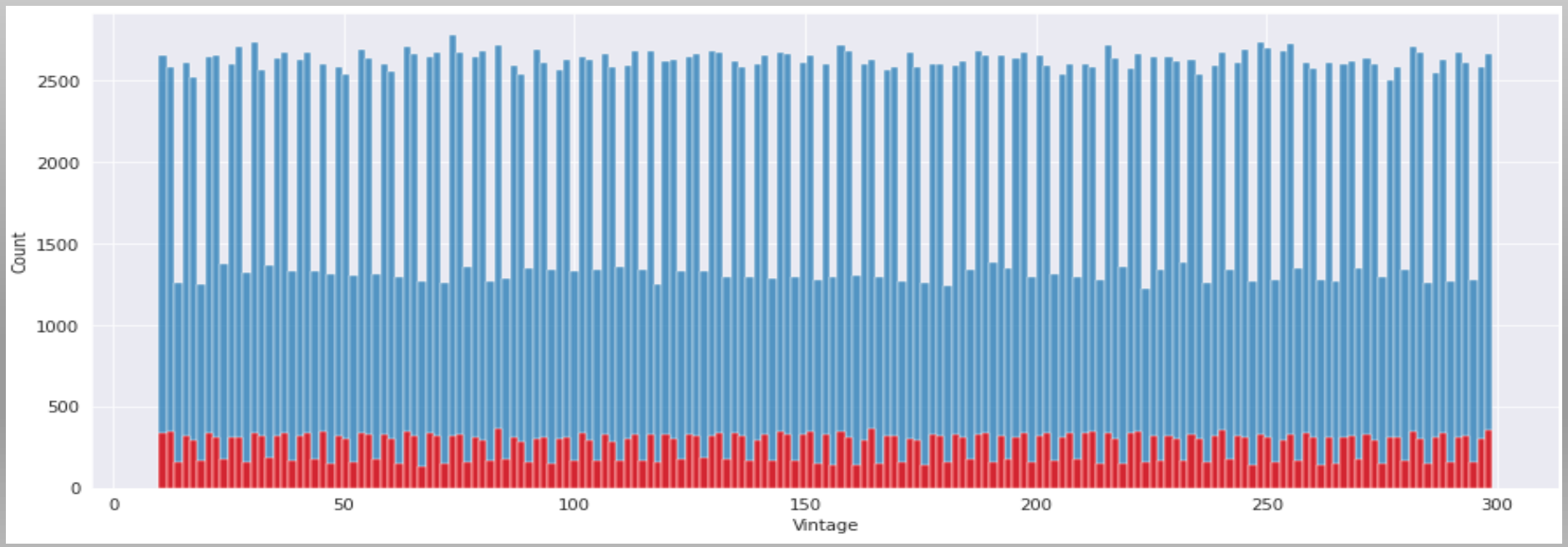
It has been observed that channel no. 26 comes out to be best channel as it gave us 15891 interested people count followed by channel 124 which gave us 13996 conversion.



	Policy_Sales_Channel	Counts	Response_1
0	152	134784	3858
1	26	79700	15891
2	124	73995	13996
3	160	21779	475
4	156	10661	2297
5	122	9930	1720
6	157	6684	1794
7	154	5993	1474
8	151	3885	122
9	163	2893	880

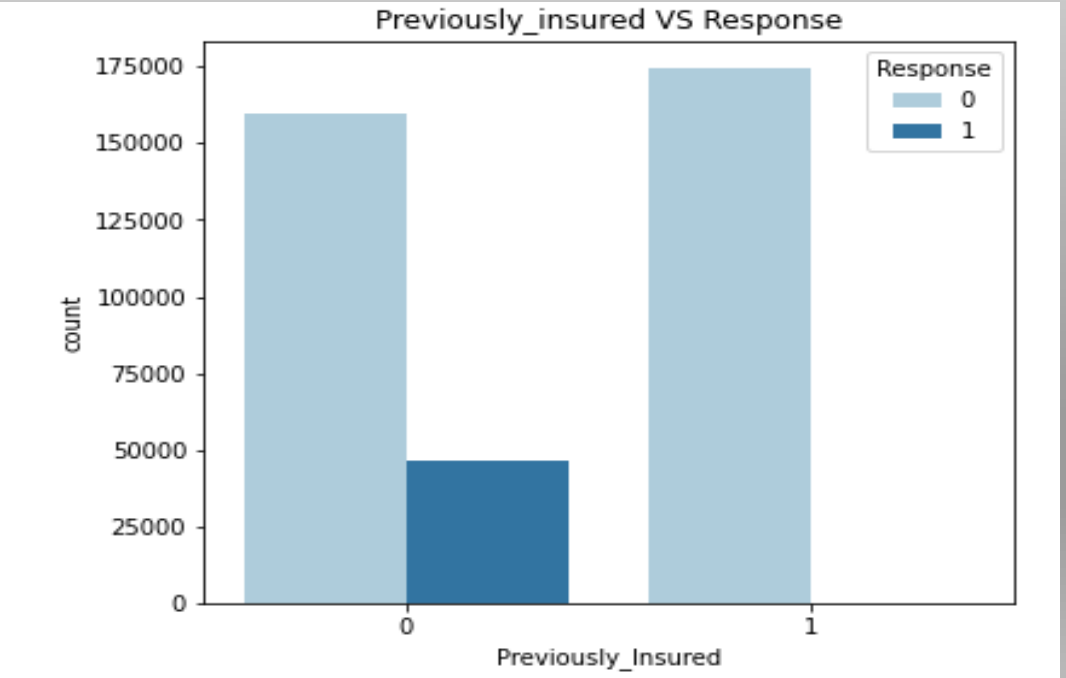
Vintage is directly proportional to insurance sell

We have seen that vintage is directly proportional to interested customers. Which means more we target to particular range we get more interested customers but the conversion is not really impressing.



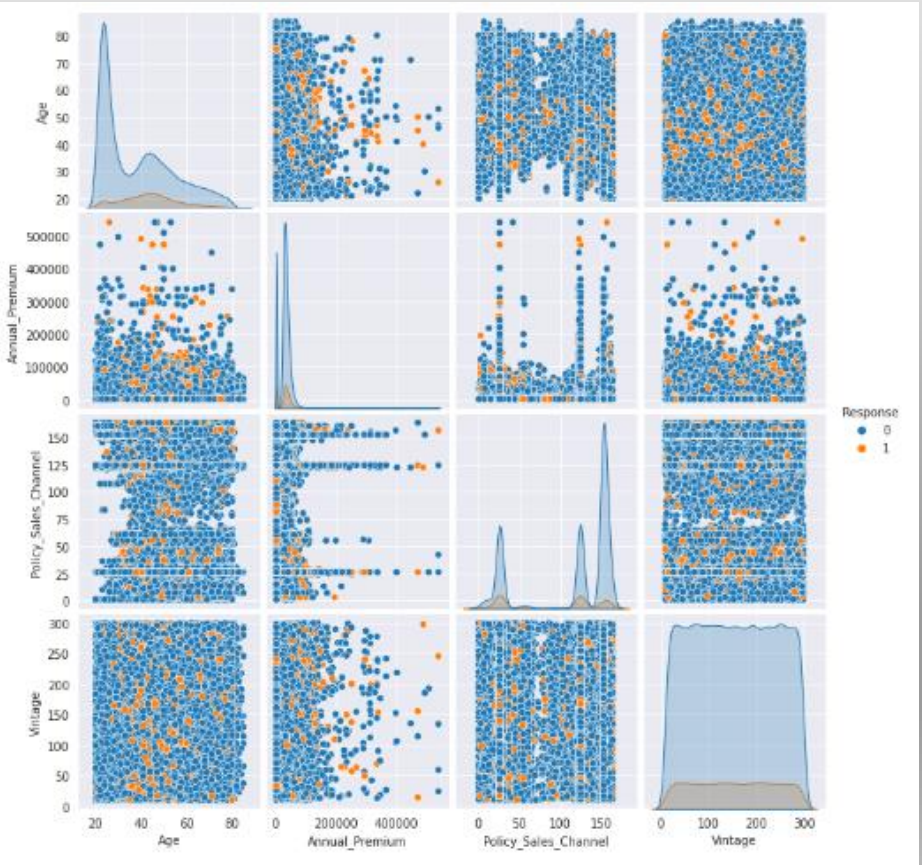
Not insured means good opportunity

If a customer is not previously insured then such customers are more likely to buy.



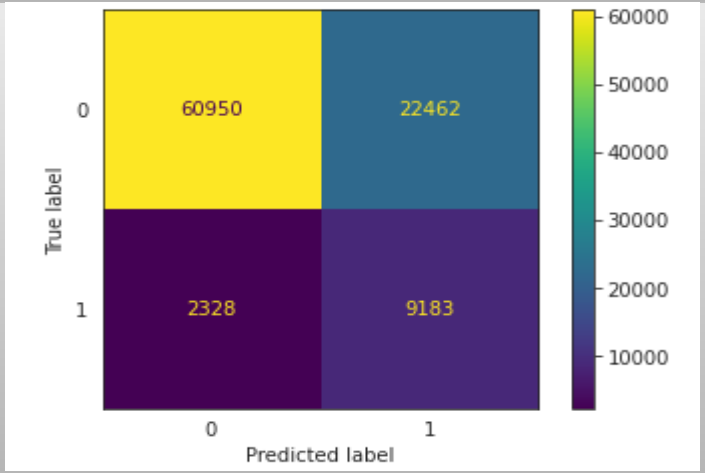
	Previously_Insured
0	206481
1	174628

Relation between features

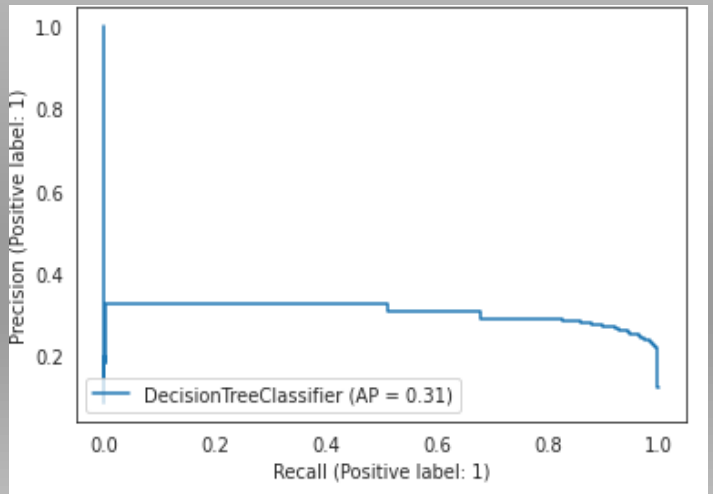


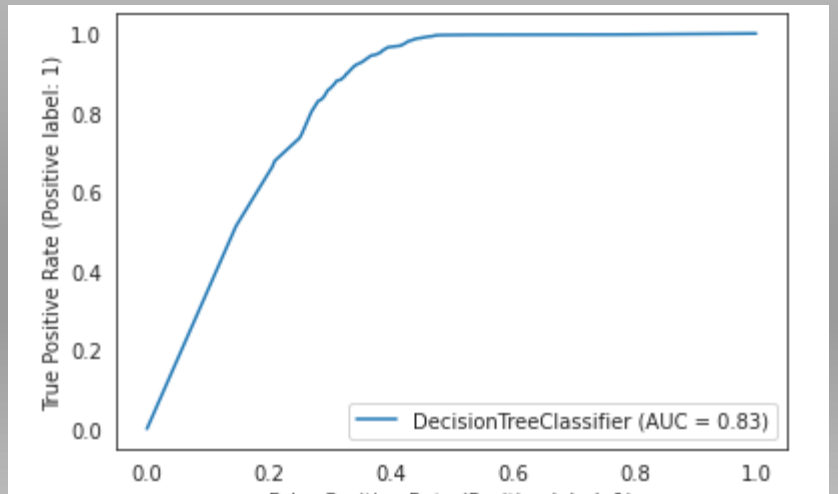

Precision score
Recall
Predictions &
MODEL EVALUATION
AUC
Confusion matrix
F1 Score

Decision Tree



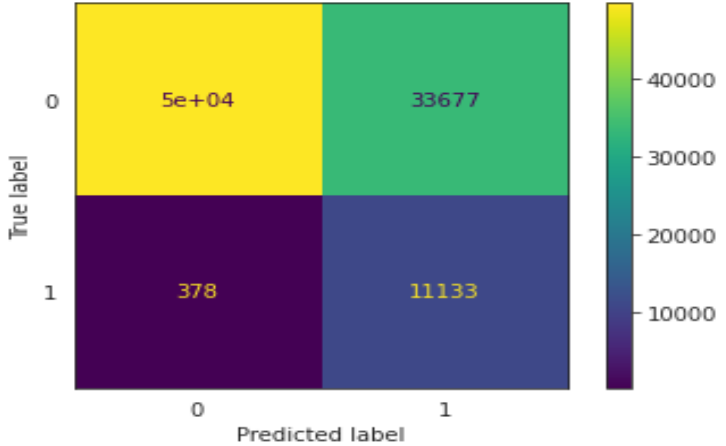
	precision	recall	f1-score	support
0	0.96	0.73	0.83	83412
1	0.29	0.80	0.43	11511
accuracy			0.74	94923
macro avg	0.63	0.76	0.63	94923
weighted avg	0.88	0.74	0.78	94923

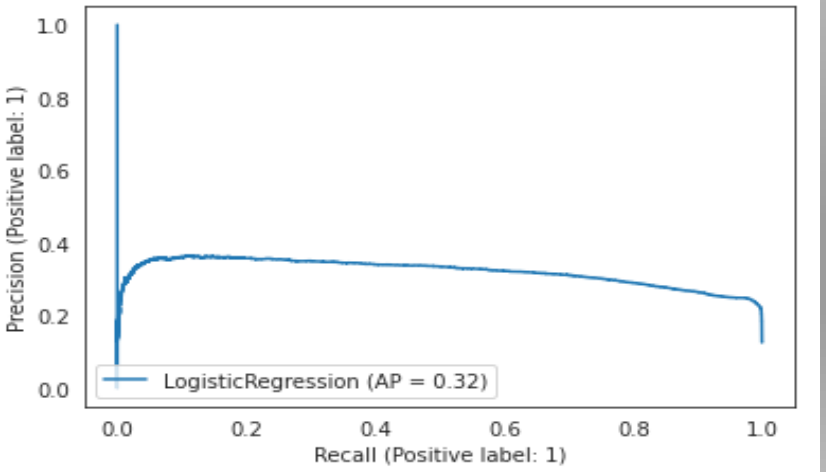


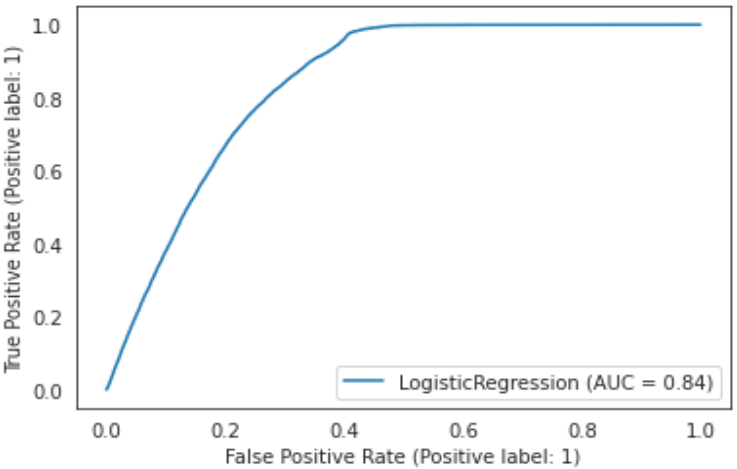


Logistic Regression

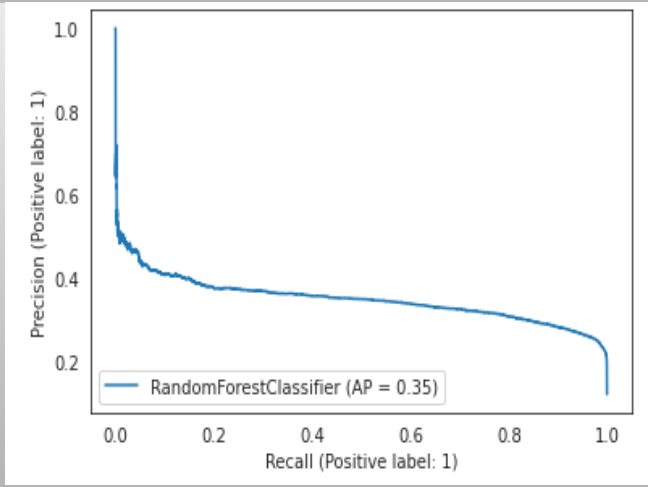
	precision	recall	f1-score	support
0	0.99	0.60	0.74	83412
1	0.25	0.97	0.40	11511
accuracy			0.64	94923
macro avg	0.62	0.78	0.57	94923
weighted avg	0.90	0.64	0.70	94923

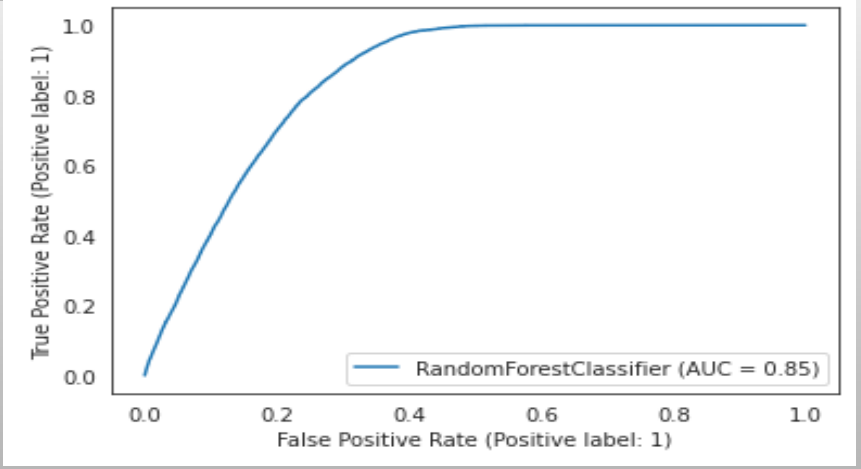




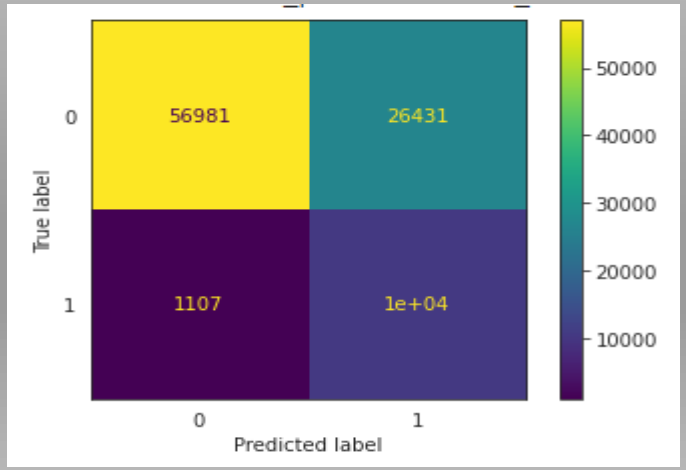


Random Forest






```
([[56981, 26431],
  [ 1107, 10404]])
```

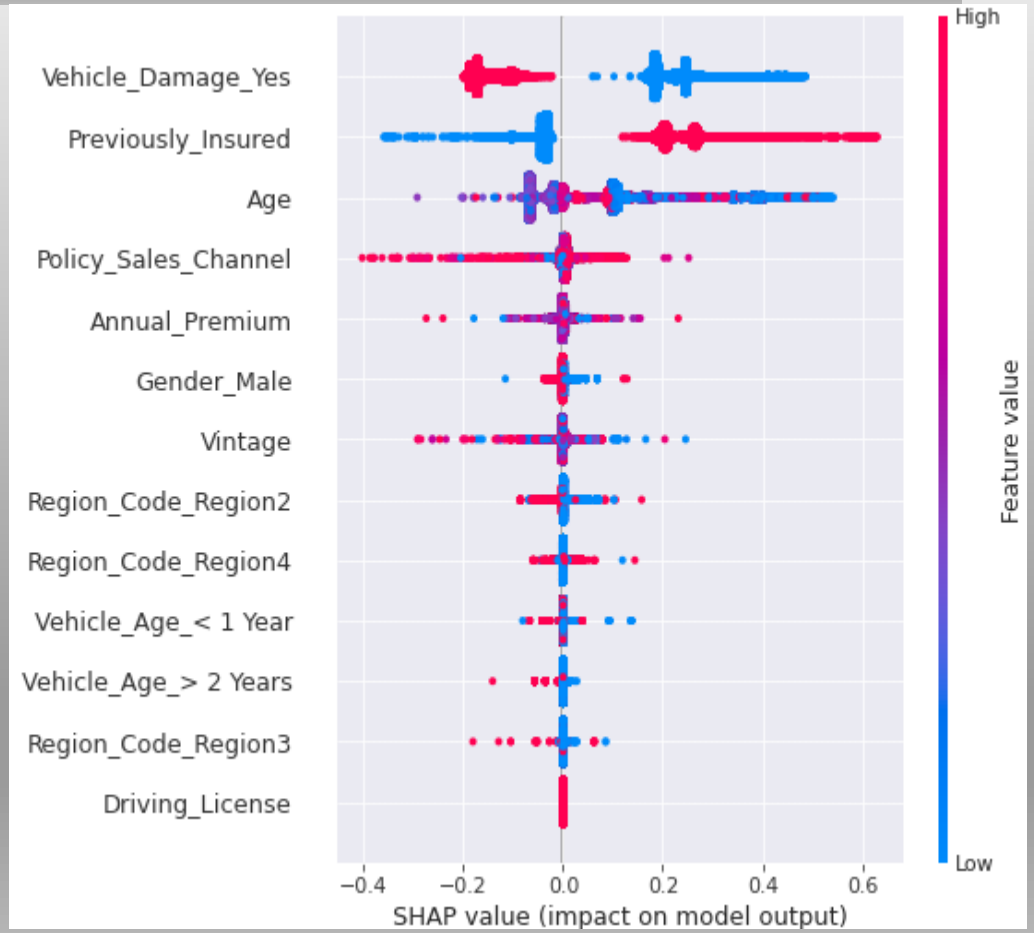


	precision	recall	f1-score	support
0	0.98	0.67	0.79	83412
1	0.28	0.92	0.43	11511
accuracy			0.70	94923
macro avg	0.63	0.79	0.61	94923
weighted avg	0.90	0.70	0.75	94923

Feature Importance

Global Feature importance based on SHAP value

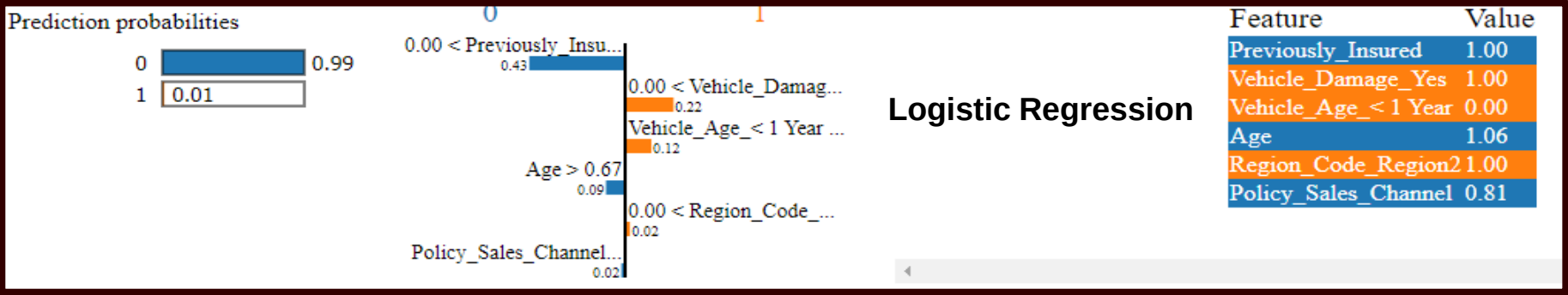
In this plot we can see which features are contributing more to the prediction whether a customer will buy insurance or not



Local interpretability

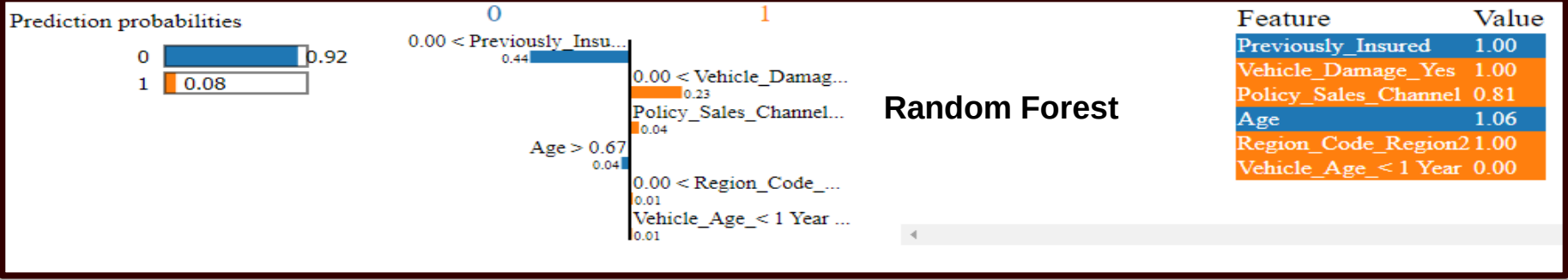
Decision Tree

Feature	Value
Previously_Insured	1.00
Vehicle_Damage_Yes	1.00
Policy_Sales_Channel	0.81
Age	1.06
Region_Code_Region4	0.00
Region_Code_Region2	1.00



Logistic Regression

Feature	Value
Previously_Insured	1.00
Vehicle_Damage_Yes	1.00
Vehicle_Age_< 1 Year	0.00
Age	1.06
Region_Code_Region2	1.00
Policy_Sales_Channel	0.81



Random Forest

Feature	Value
Previously_Insured	1.00
Vehicle_Damage_Yes	1.00
Policy_Sales_Channel	0.81
Age	1.06
Region_Code_Region2	1.00
Vehicle_Age_< 1 Year	0.00

Challenges

- *As the number of records is very high hence it was very time consuming to run a model even in some code/model it take a lot of time to run the code.*
- *We observed that the dataset that we are given is imbalance dataset and thus we had to use some synthetic technique to create a balanced dataset.*
- *Evaluation metric selection was challenge as choosing the incorrect metric can lead poor results.*
- *Due to very large dataset we were not able to run complex models even when we tried it took approx. six hours but still there was no output and this might be the case because the system configuration is not that much efficient.*
- *AS the dataset has mixed datatypes hence we faced challenges while we have transformed data to feed our model.*

Conclusions

- *Men's are more interested compare to women.*
- *We can target more people having age range 30-55 years.*
- **Region 29 has given us highest is giving us the most possible customers.**
- **Less people are interested having "< 1year" vehicle age while the count of ">1year" is comparatively good. *Hence organization should make some strategy to attract no-damage vehicle.***
- **Less records having > 75000 premium and hence less +ve responses.**
- ***We saw that channel no. 26 comes out to be best channel as it gave us 15891 interested customers followed by channel 124 which gave us 13996 conversion.***

Conclusions

- *If a customer is not previously insured then such customers are more likely to buy.*
- *If we look from the perspective of explainability so we can choose decision tree but Random forest results are slightly better however the model is bit complex.*
- *Vehicle damage is the most important feature that drive predictions followed by previously insured.*

THANK YOU

Presented By :-
Mohd Sharik