Capstone Project - 3

Health Insurance Cross sell Prediction

Team members

Balaji B. Jadhav
Anant M. Patil
Nishigandha Ingale

Content

- Introduction
- Problem definition
- Hypothesis
- EDA on given data
- Statistical Data Analysis
- Model implementation
- Model validation and selection
- Conclusion
- References

Introduction

- Currently automobiles are being used to land transportation and living, and the scope of use and equipment is expanding.
- This rapid increase in automobiles has caused automobile insurance to emerge as an essential business target for insurance companies.
- Therefore, if the car insurance sales are predicted and sold using the information of existing health insurance customers.

Introduction

- It can generate continuous profits in the insurance company's operating performance.
- Aim of the project is to put a model on given data for health insurance cross sell prediction.

Problem definition

- Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company.
- It can then accordingly plan its communication strategy to reach out to those customers and optimise its business m odel and revenue.

Hypothesis Statement

- From data and problem statement we would like to put hypothesis on 5 features as.
- Vehicle previously insured will affect Response variable.(If the vehicle is already insured then no will buy insurance twice.)
- Age will not affect Response.(All the vehicle owners no matter what age will buy insurance.)
- Driving license will not affect Response.(In general it will not affect)
- There will be positive co relation in Vintage and Response.(Old customers with company will buy insurance.)
- Vehicle damage will affect Response.(If the vehicle is damaged previously owner will definitely buy insurance.)

EDA

Digging into data we understand that

There are total 11 Features such as

- 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 have DL.
- Region code - Unique code for the region of the customer.
- Previously Insured – 1: customer has already insurance, 0: customer does not have insurance.

EDA

- Vehicle age – Age of the vehicle.
- Vehicle damage – 1: customer got his/ her car damage in the past 0: customer didn't get his/ her car damage in the past
- Annual premium – The amount customer needs to pay as a premium in the year.
- Policy sales channel – Anonymized code for the channel of outreaching the customer. i.e. Different agents, over mail etc.
- Vintage : Number of Days, Customer has been associated with the comp any

Response variable is :

- Response : 1 : Customer is interested, 0 : Customer is not interested

EDA

- We can see there is no null values in the given data set.

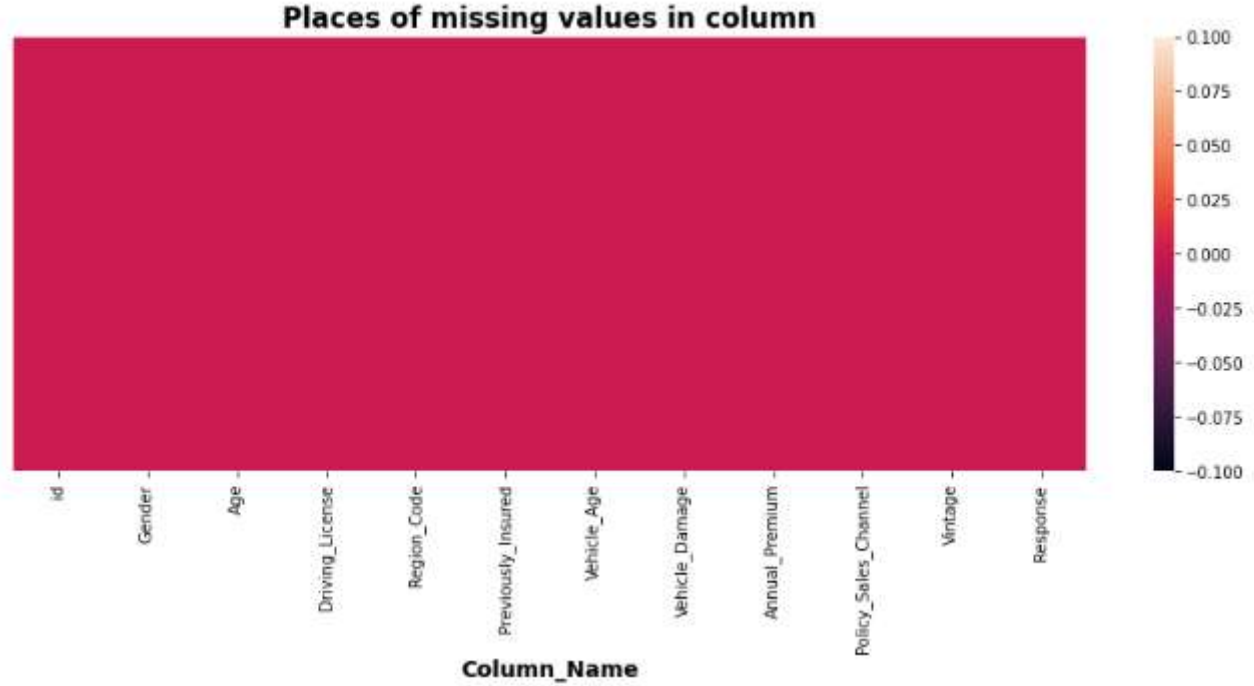

Fig 1: Missing values in data set.

EDA

- Pi plot shows percentage of response(Fig 2) and Density plot of age and response(Fig 3).

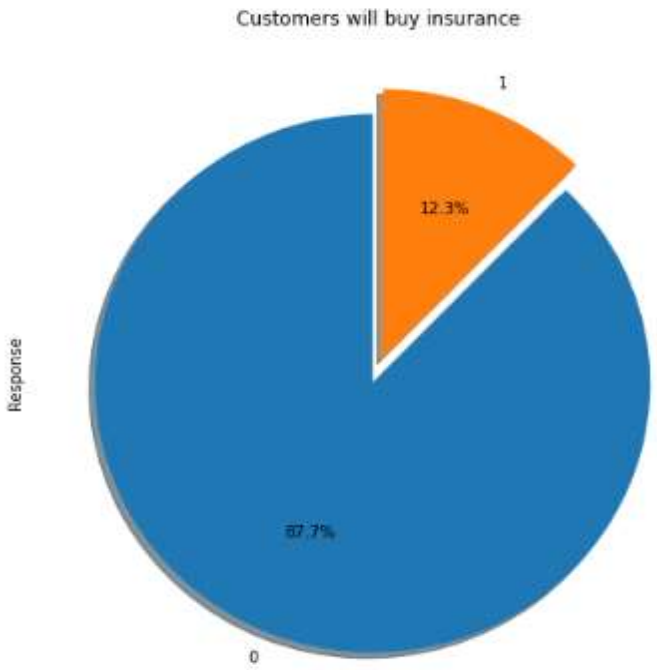

Fig 2: Pi plot of Response 0 and 1.

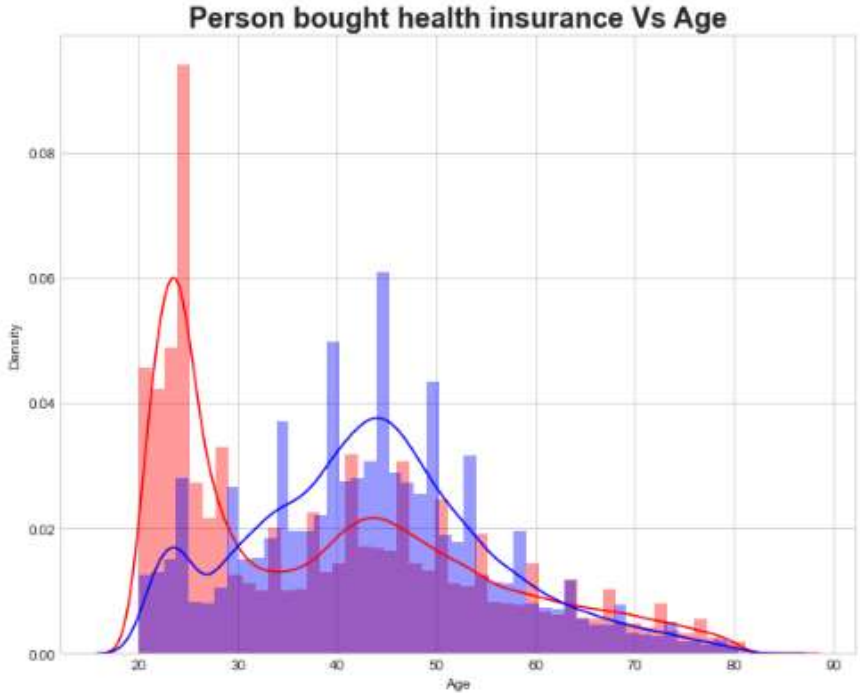

Fig 3: Plot of Age vs Response.

EDA

- In the age of 35 to 50 Response count is more.

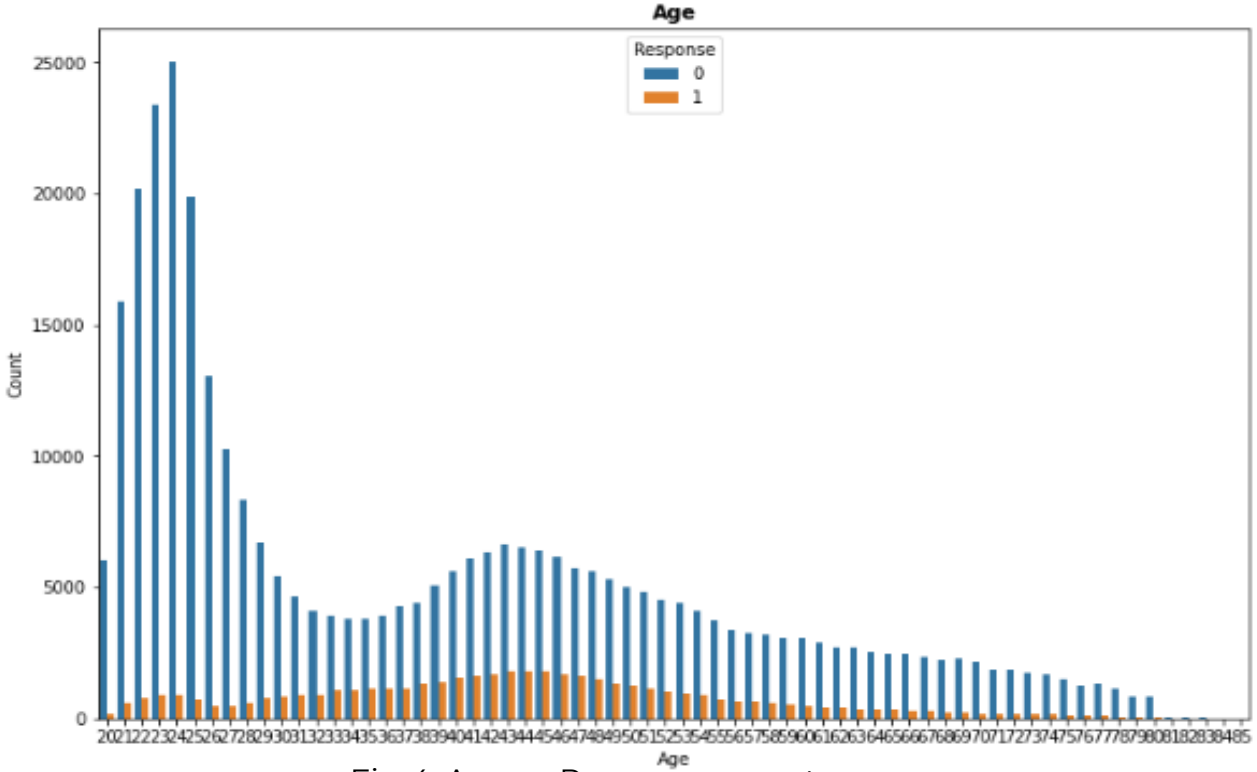

Fig 4: Age vs Response count.

EDA

- Region code 29 has more count (0 and 1) for Response.

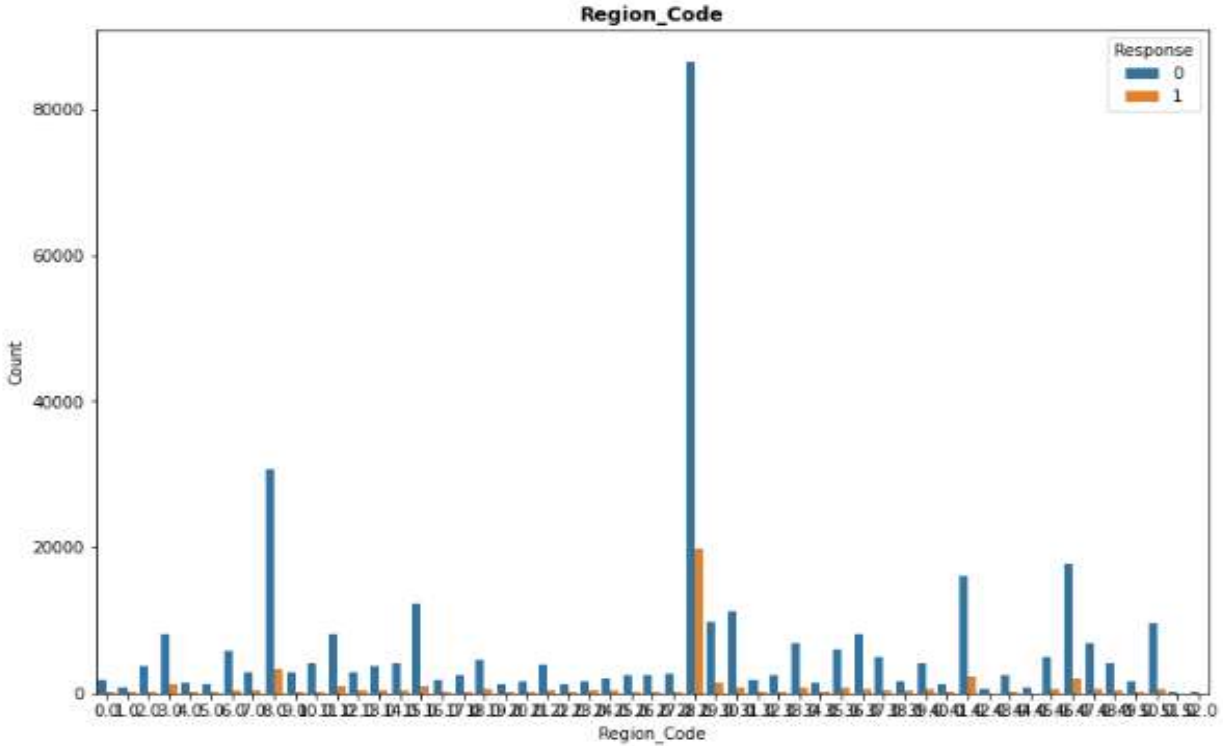

Fig 5: Region vs Response count.

EDA

- Gender has no specific impact on Response same condition for Driving license.

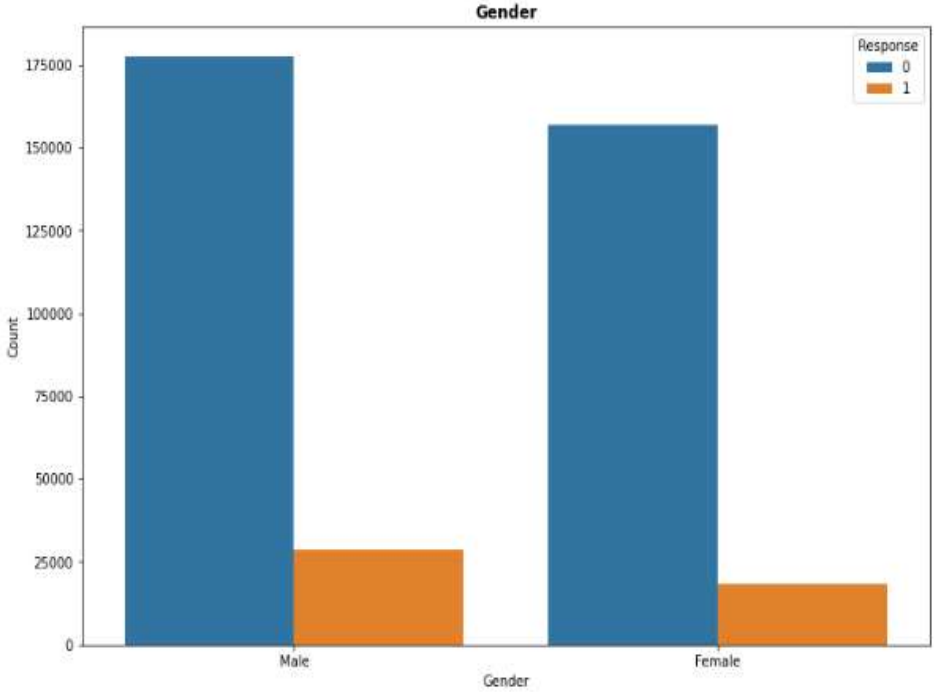

Fig 6: Region vs Response count.

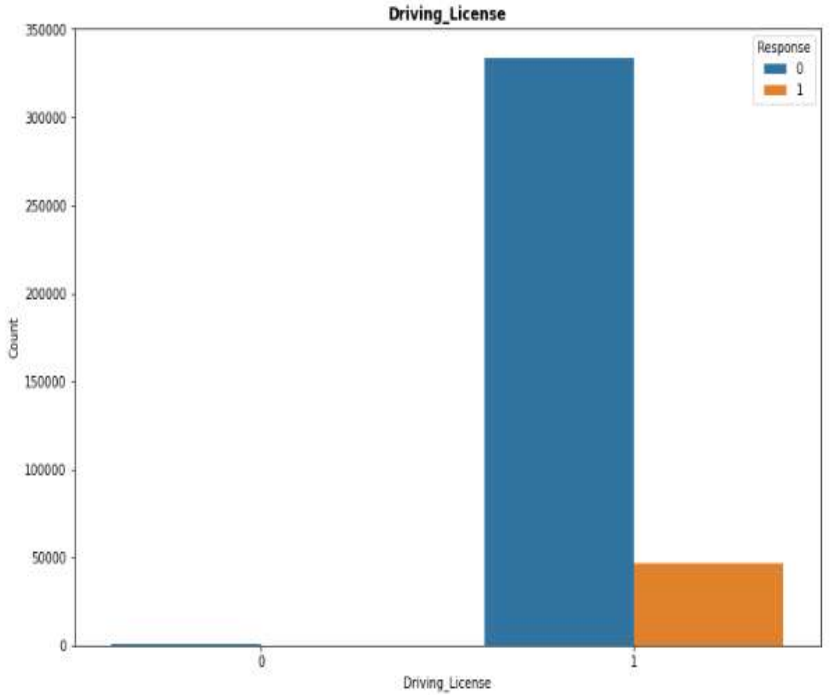

Fig 7: Region vs Response count.

EDA

- Previously insured has more impact on Response and Vehicle age between 1-2 years are more interested in insurance.

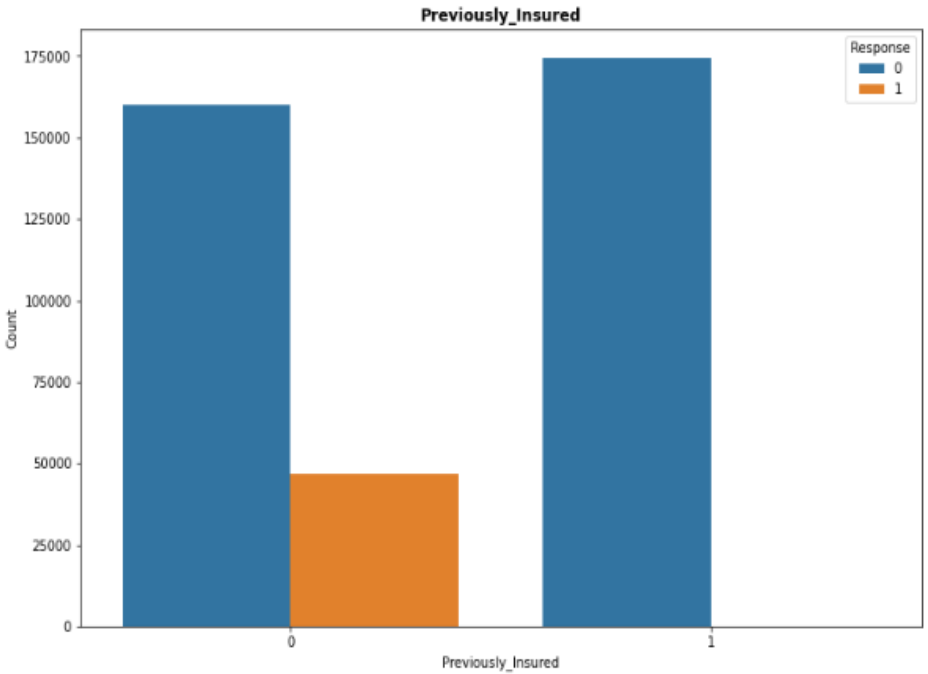

Fig 8: Previously insured vs Response count.

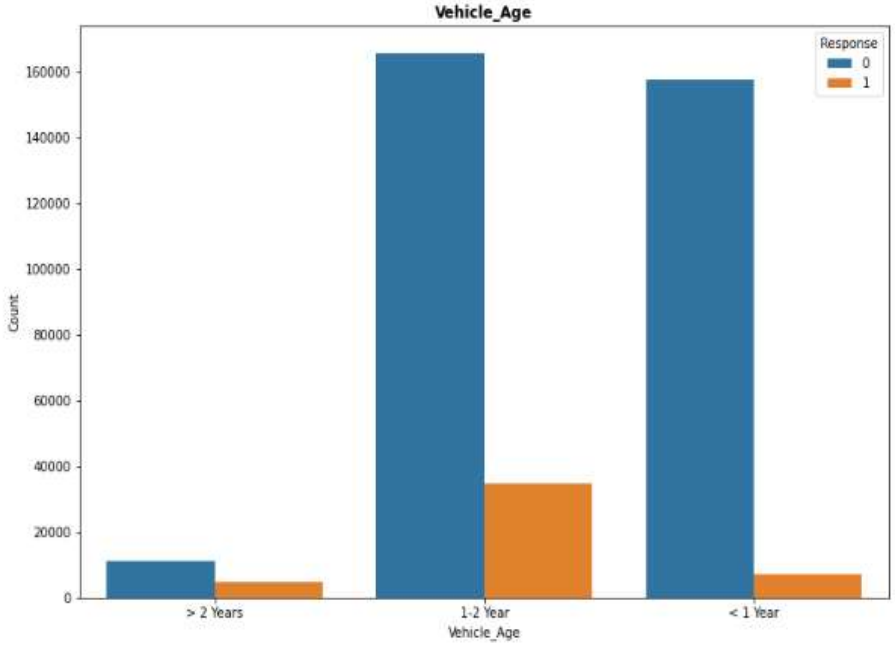

Fig 9: Vehicle age vs Response count.

EDA

- If Vehicle is damaged then response is 1. There is no much difference in between vehicle damage yes and no.

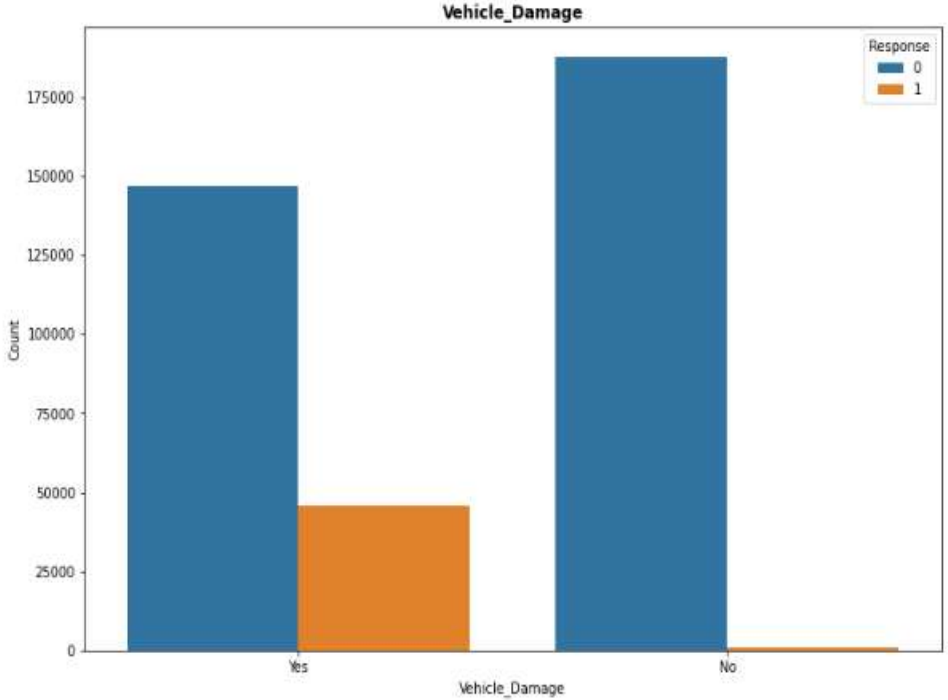

Fig 10: Vehicle damage vs Response count.

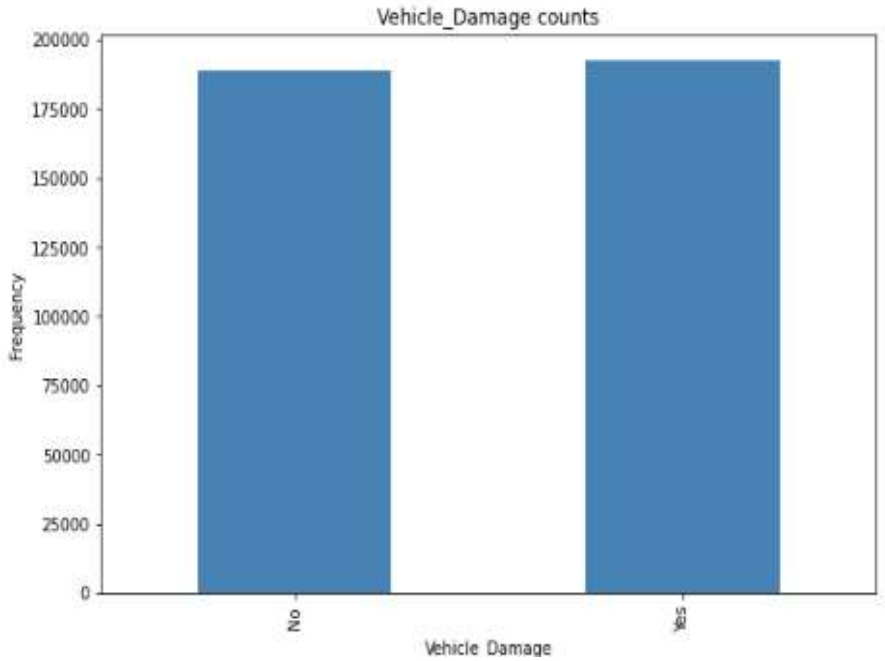

Fig 11: Vehicle damage vs Frequency.

EDA

- Age is positively co related with Annual premium.

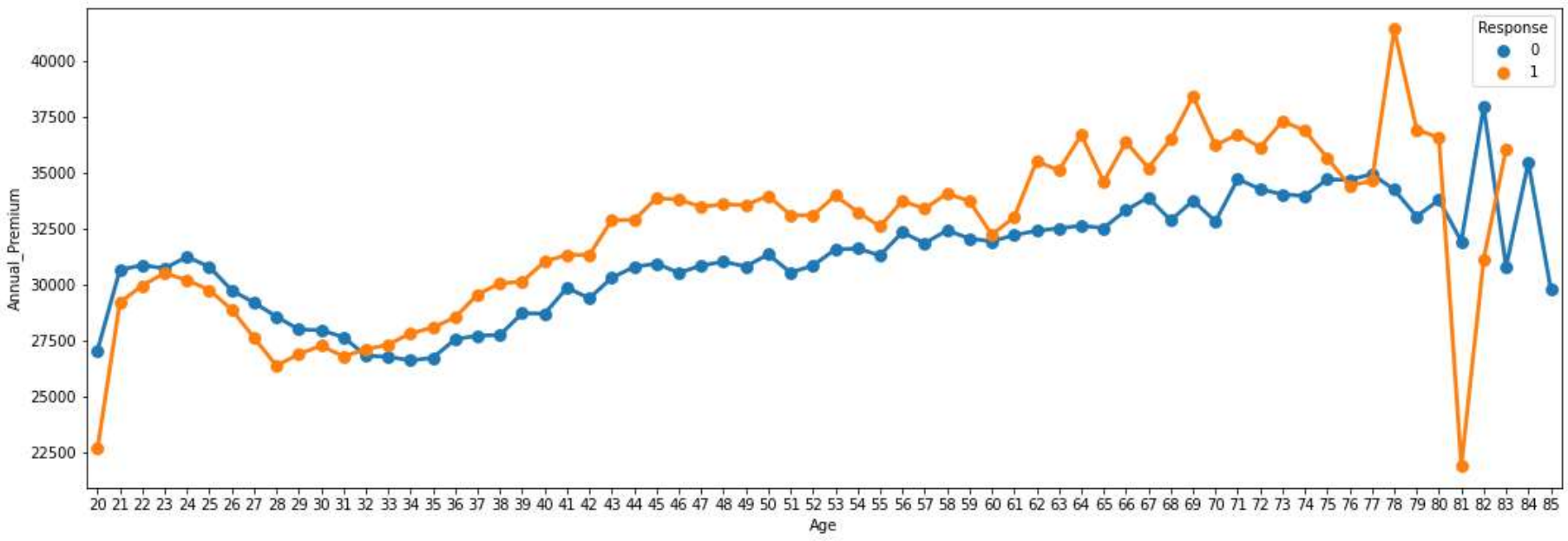

Fig 12: Age vs Annual premium.

EDA

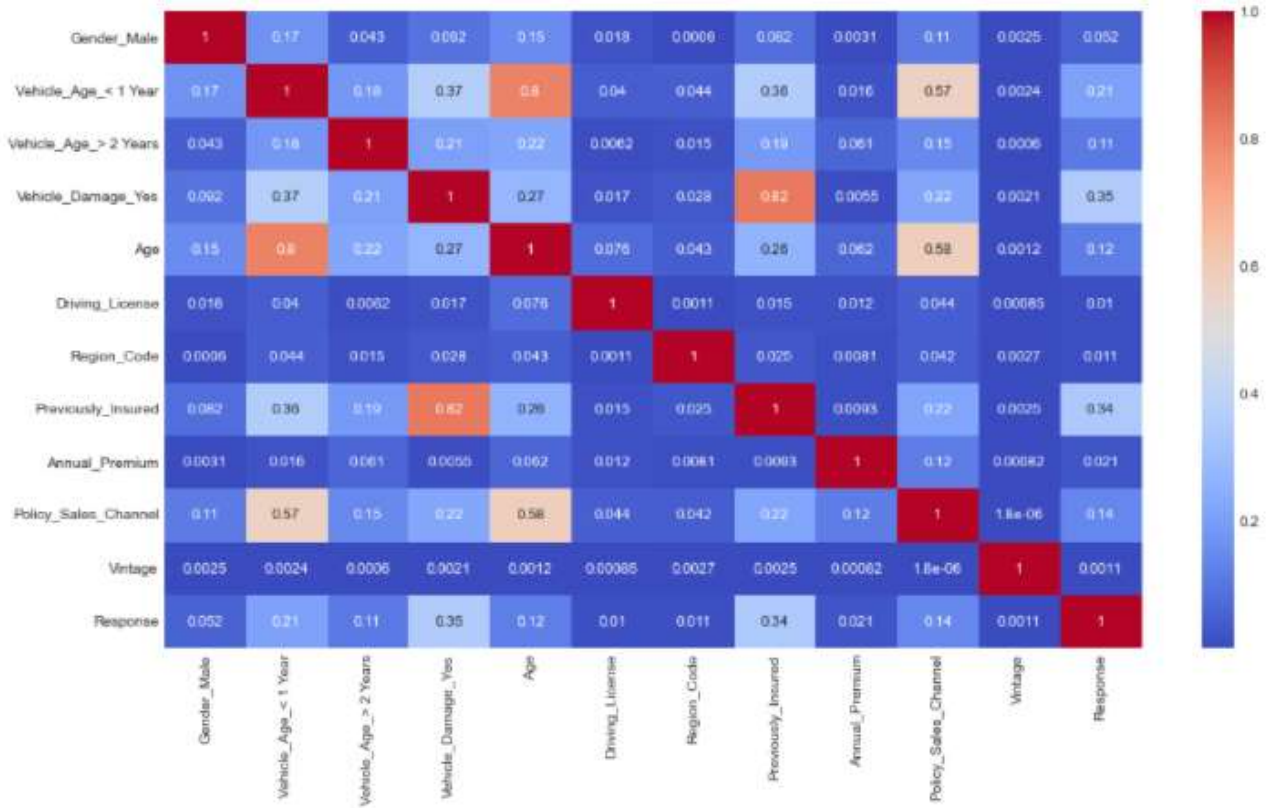

Fig 13: Heat map of variables

Statistical Data Analysis

- Region code is normally distributed and has no outliers.

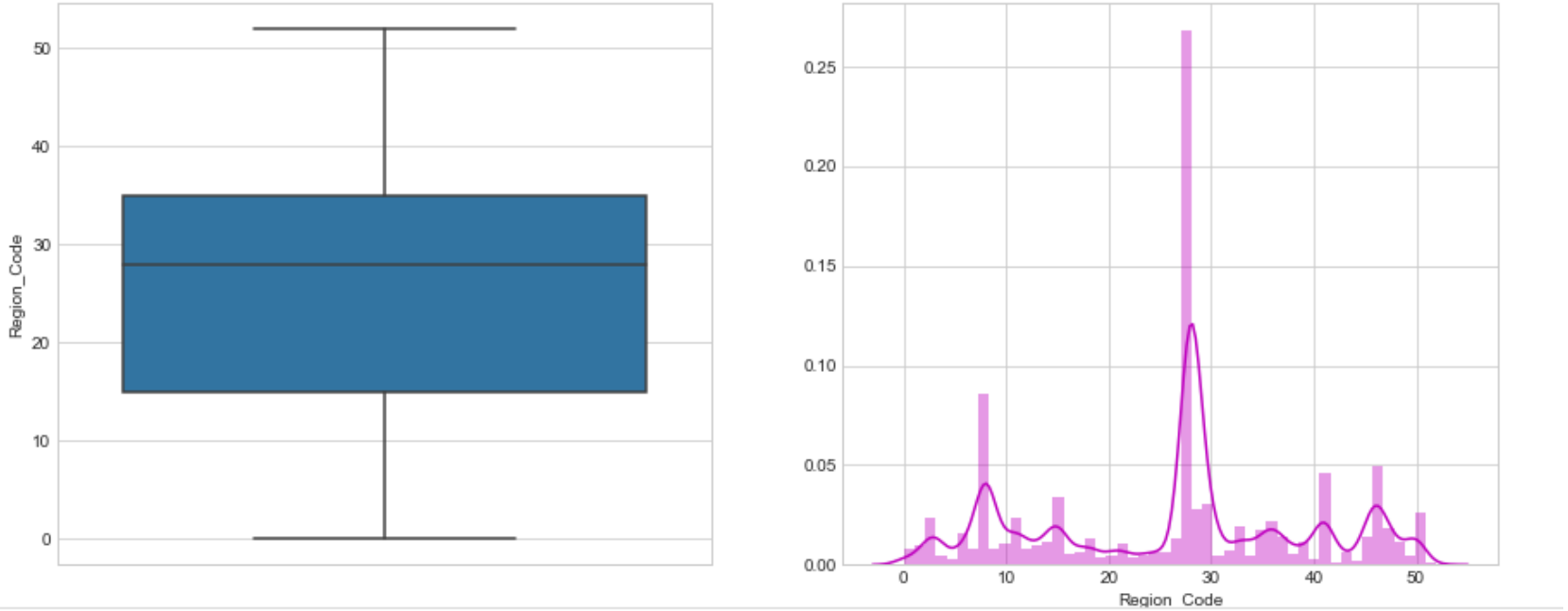

Fig 14: Distribution of Region code.

Statistical Data Analysis

- Annual premium is positively skewed and has outliers.

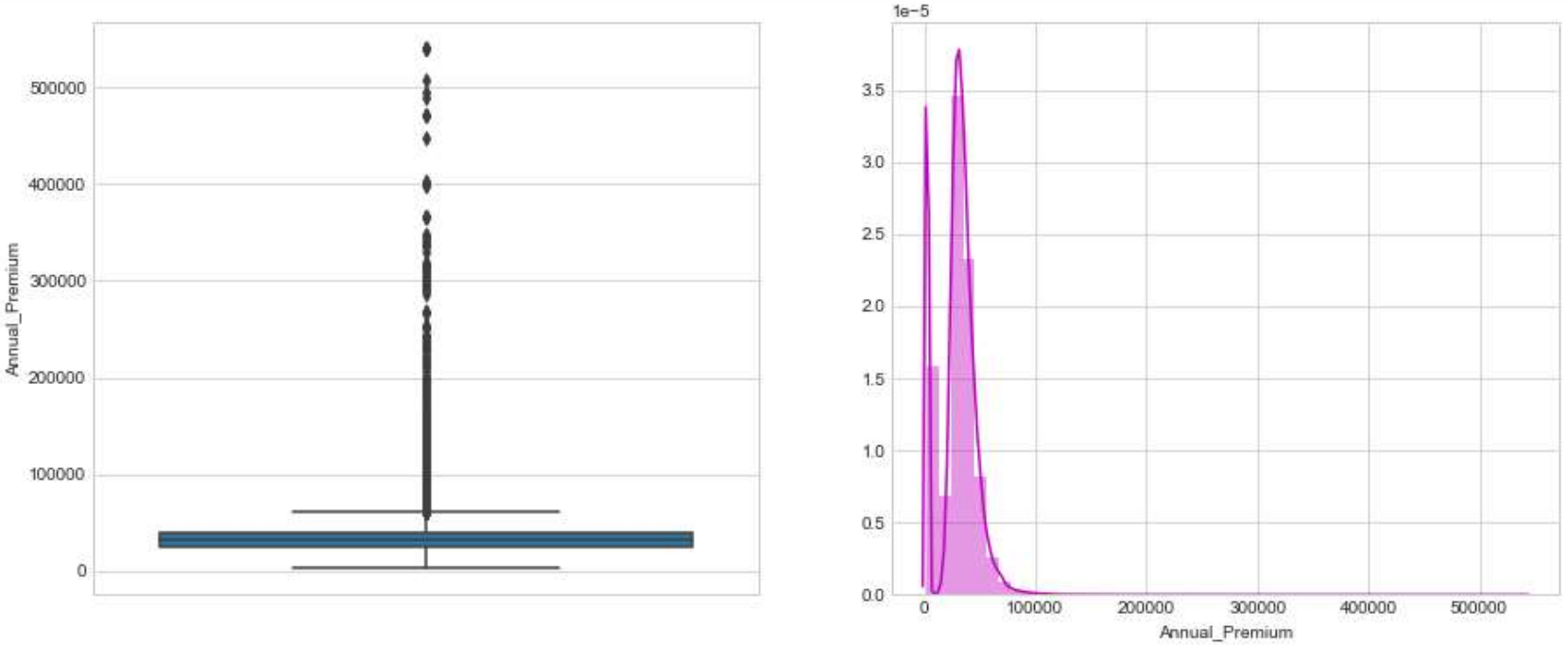

Fig 15: Distribution of Annual premium.

Statistical Data Analysis

- Policy sales channel has 3 kurtosis and has no outliers.

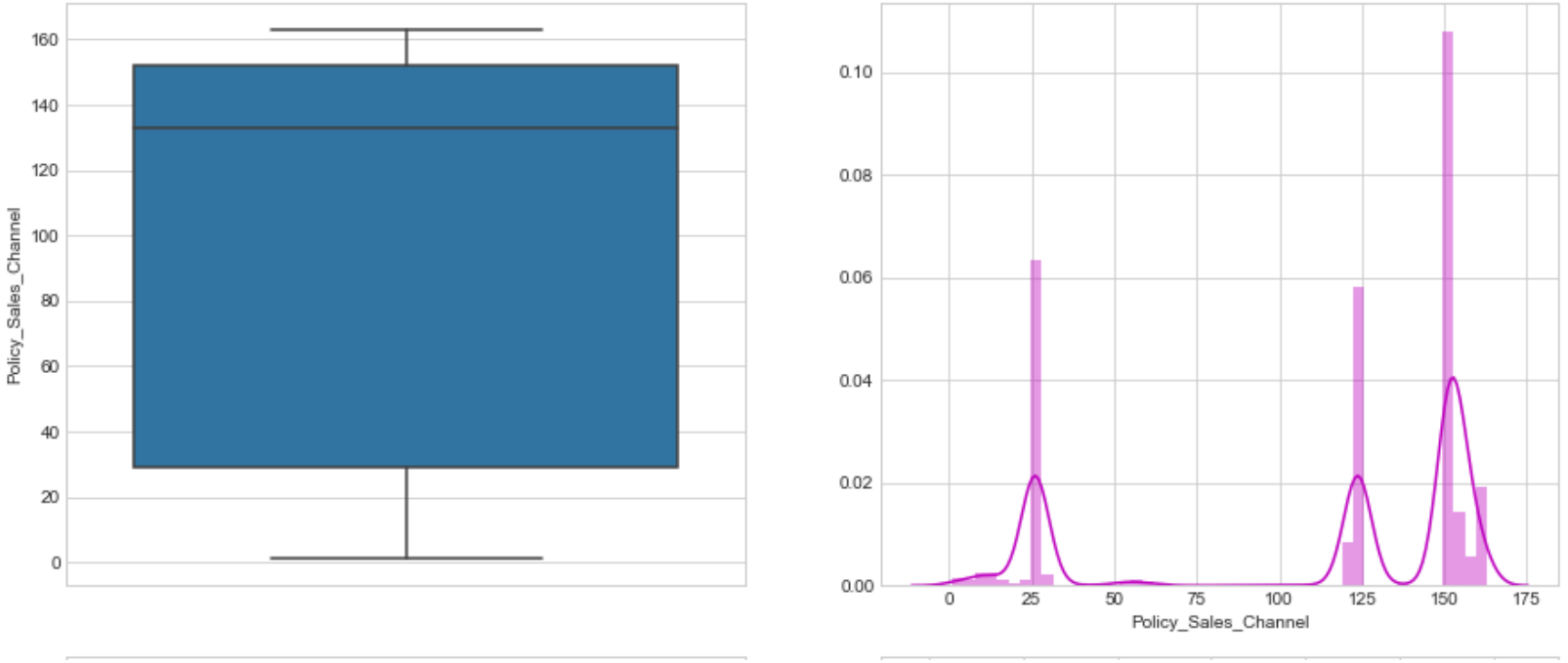

Fig 16: Distribution of Policy sales channel.

Statistical Data Analysis

- Distribution of Vintage customers which has no outliers.

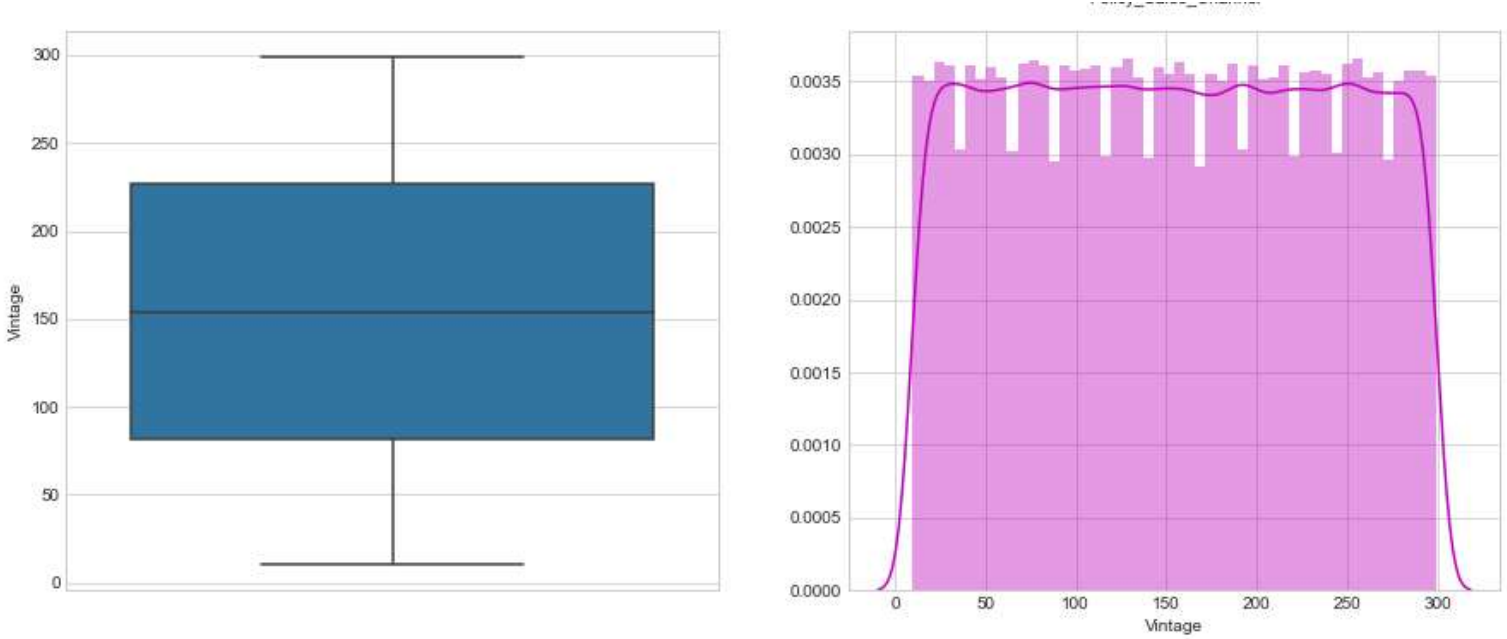

Fig 17: Distribution of Vintage customers.

Statistical Data Analysis

- Age is positively skewed and has no outliers.

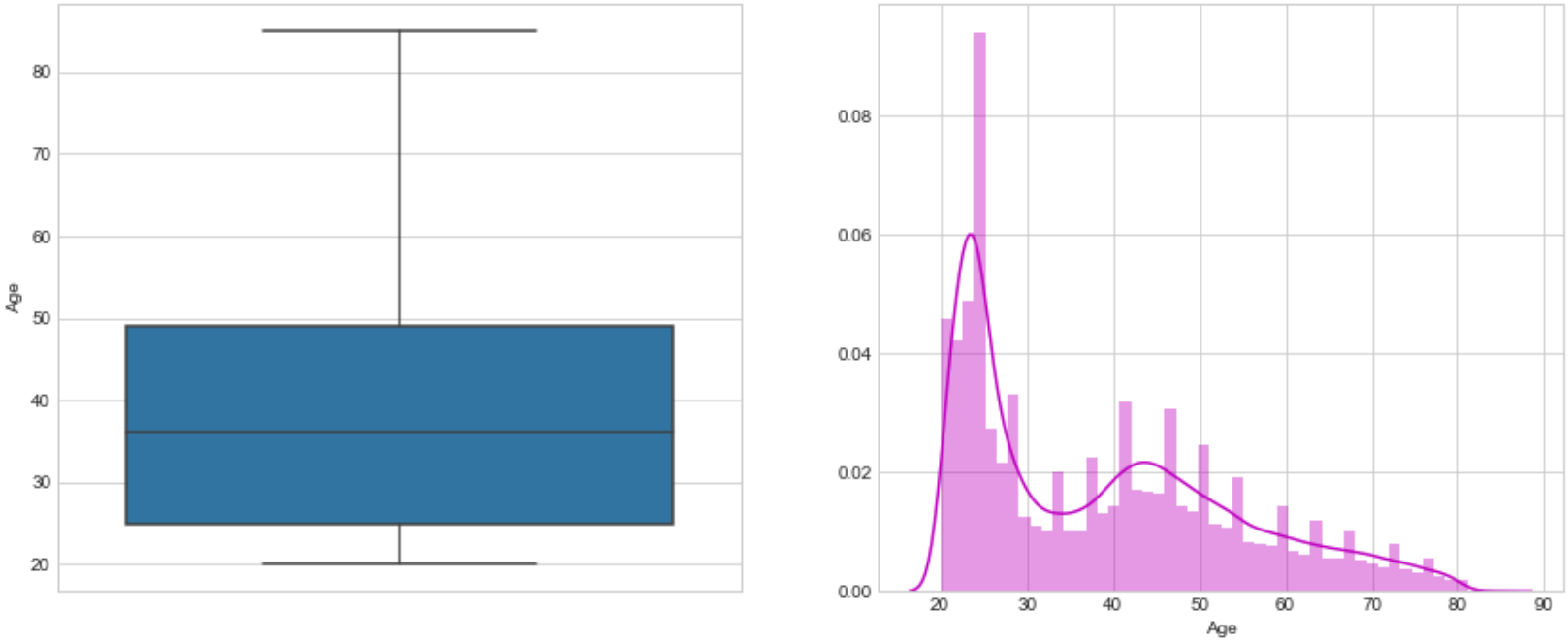

Fig 18: Distribution of Age.

Statistical Data Analysis

- Treated Annual premium with removal of outliers .

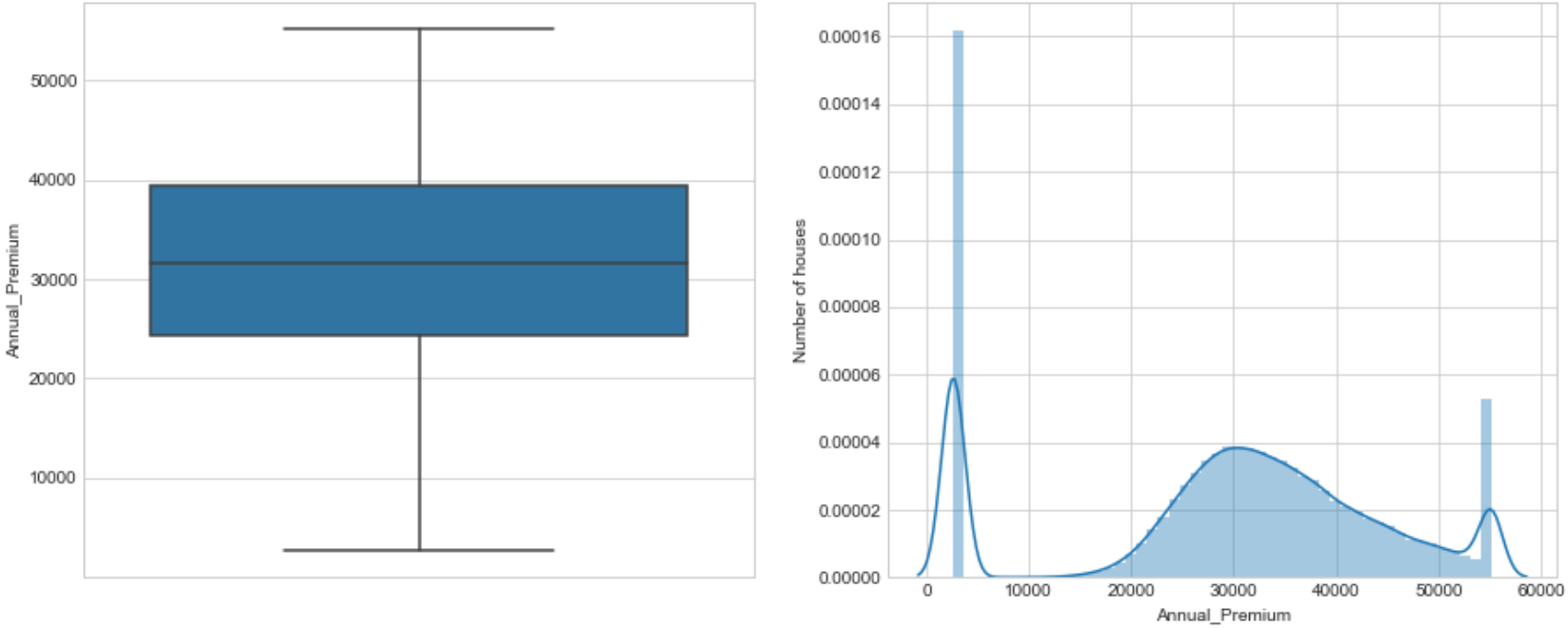

Fig 19: Treated annual premium

Statistical Data Analysis

- Treated Age with removal of outliers .

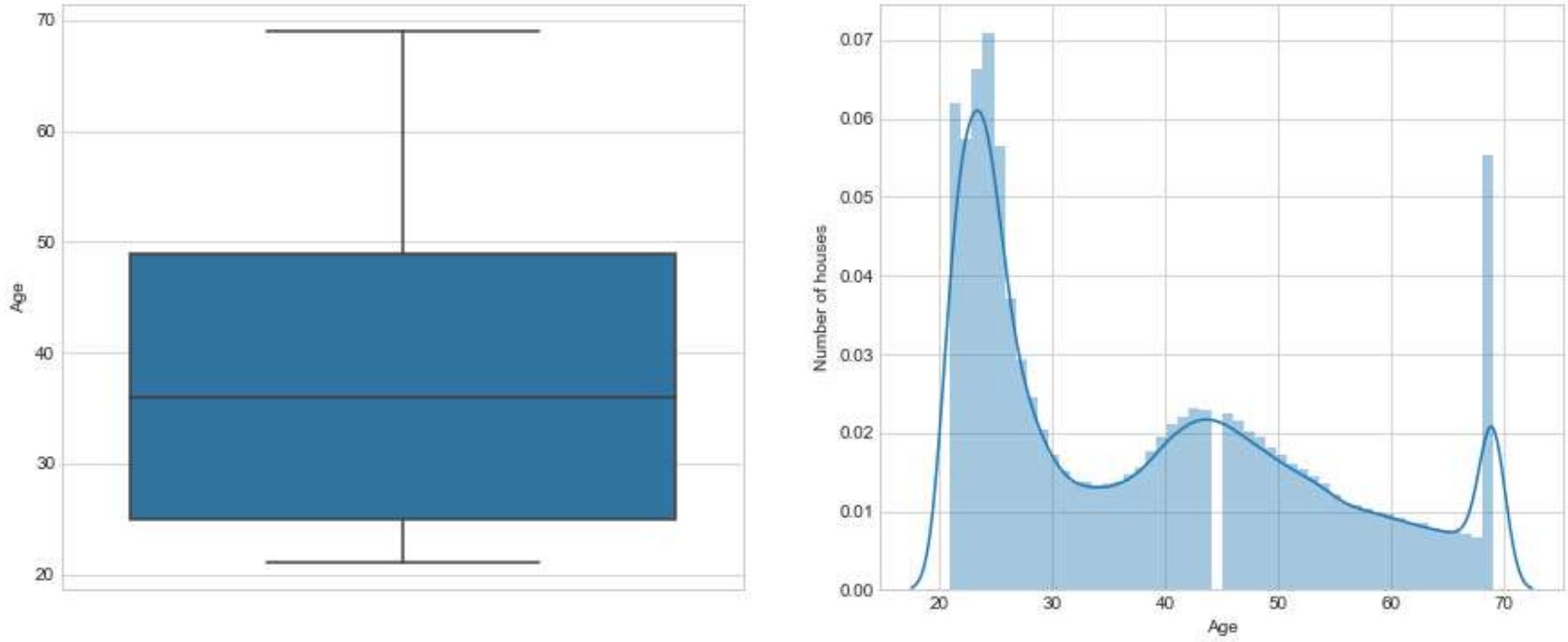

Fig 20: Treated Age .

Statistical Data Analysis

- Age and Response are positively co related.(Fig 22). And Driving license and Response are positively co related.(Fig 23)

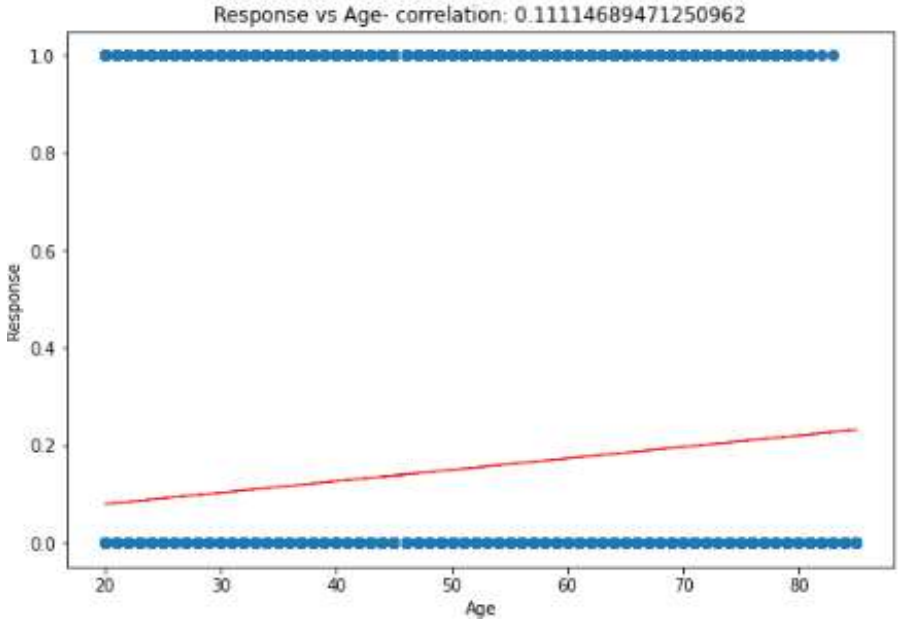

Fig 21: Co-relation plot of Age and Response.

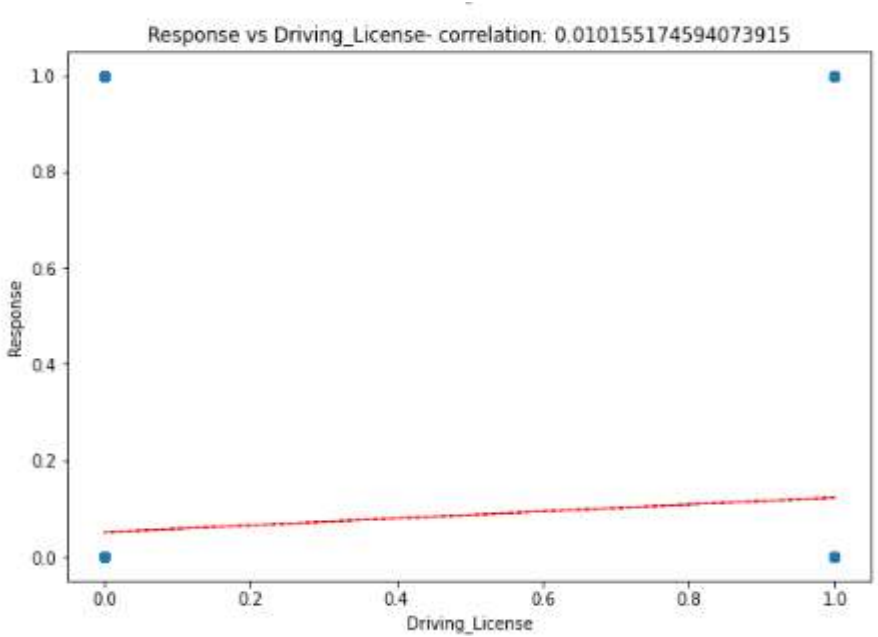

Fig 22: Co-relation plot of Driving license and Response.

Statistical Data Analysis

- Region code and Response have no co relation.(Fig 24) and Previously insured and Response are negatively co related(Fig 25)

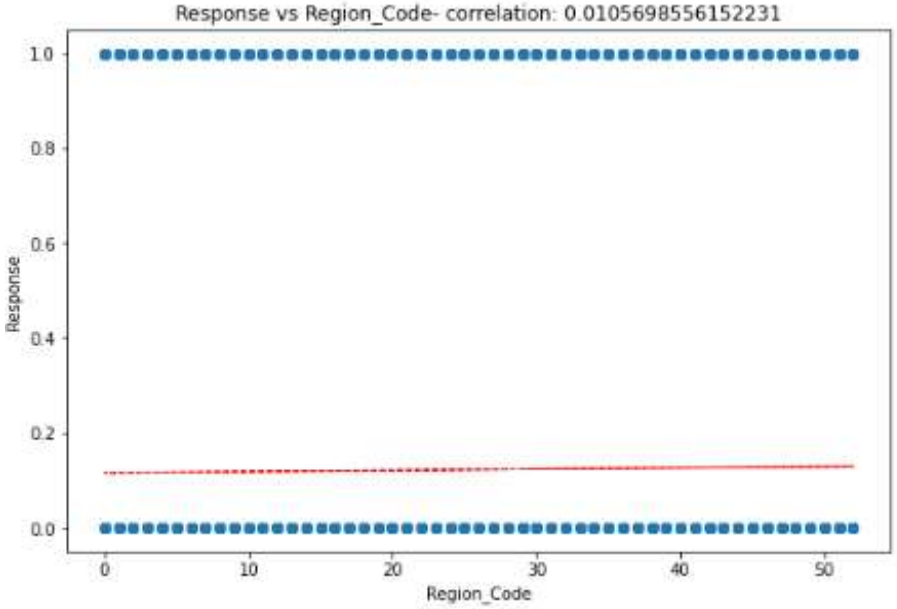

Fig 23: Co-relation plot of Region code and Response.

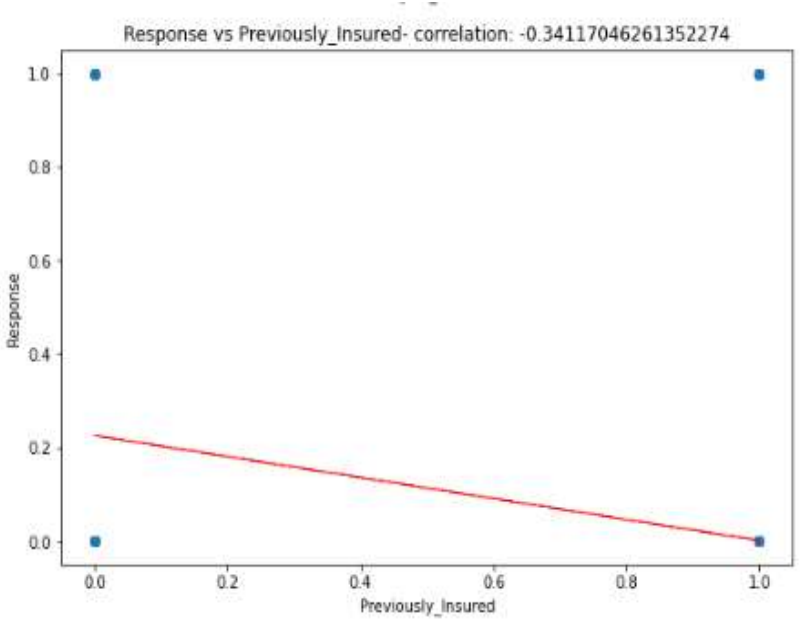

Fig 24: Previously insured and Response.

Statistical Data Analysis

- Annual premium and Response are positively co related.(Fig 26) and Policy sales channel and Response are negatively co related.(Fig 27)

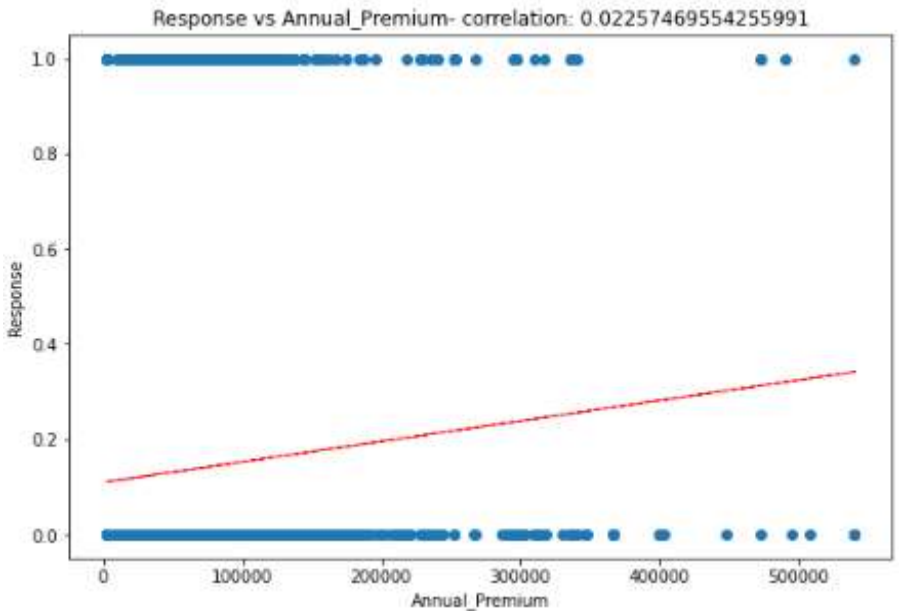

Fig 25: Co-relation plot of Annual premium and Response.

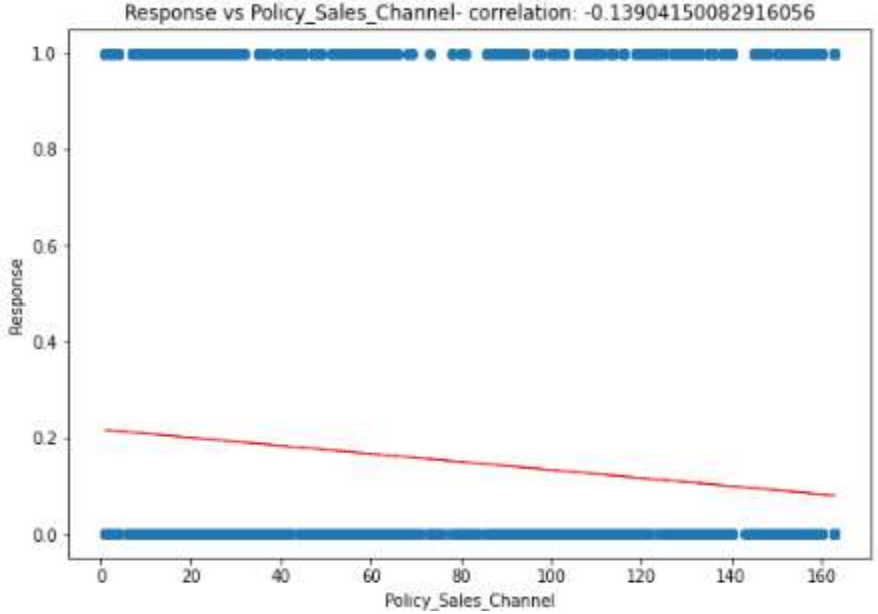

Fig 26: Co-relation plot Policy sales channel and Response.

Statistical Data Analysis

- Vintage and Response has no co relation.(Fig 28).

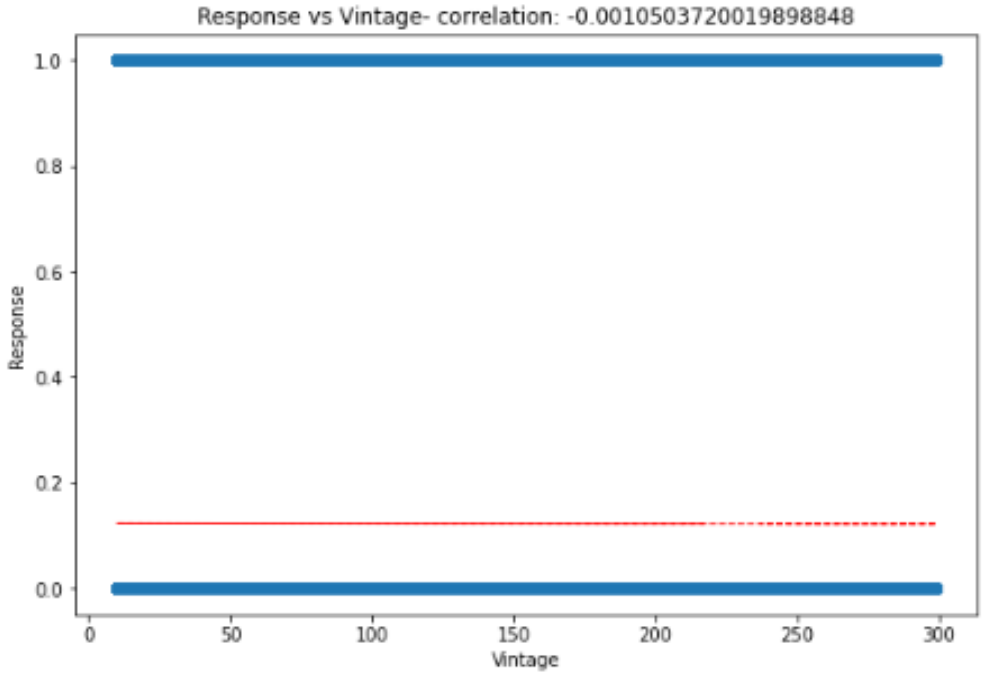

Fig 27: Co-relation plot of Vintage and Response.

Hypothesis Testing

- With EDA we can justify our hypothesis.
- Vehicle previously insured will affect Response.(Fail to reject.)
- Age will not affect Response.(Reject)
- Driving license will not affect Response.(Reject)
- There will be positive co relation in Vintage and Response.(Reject)
- Vehicle damage will affect Response.(Fail to reject)

Model implementation

- After implementing various models on the given data such as Logistic Regression, Decision Tree Classifier, XG boost, Naïve Bays Classifier, SGD Classifier, Cat Boost Classifier, KNN Classifier, Gradient Boosting Classifier, AdaBoost Classifier, Random Forest Classifier.

Model	Train	Test	Model	Train	Test
Logistic Regression	0.76	0.59	Cat Boost Classifier	0.5	0.87
Decision Tree Classifier	0.99	0.80	KNN Classifier	0.86	0.73
XG boost	0.88	0.81	Gradient Boosting Classifier	0.82	0.72
Naïve Bays Classifier	0.76	0.58	AdaBoost Classifier	0.81	0.70
SGD Classifier	0.76	0.59	Random Forest Classifier	0.99	0.81

Table 1 Result table of models.

Model implementation

- We get maximum accuracy with Decision Tree Classifier, Random Forest Classifier, XG boost but in case of Decision Tree and Random forest accuracy is decreased for testing data it is the case of Over fitting. In case of XG Boost accuracy decreases but with less percentage here there is no problem of overfitting.

Model	Train	Test
XG Boost	0.88	0.81
Decision Tree Classifier	0.99	0.80
Random Forest Classifier	0.99	0.81

Table 2 Result table of Complex models.

Model implementation

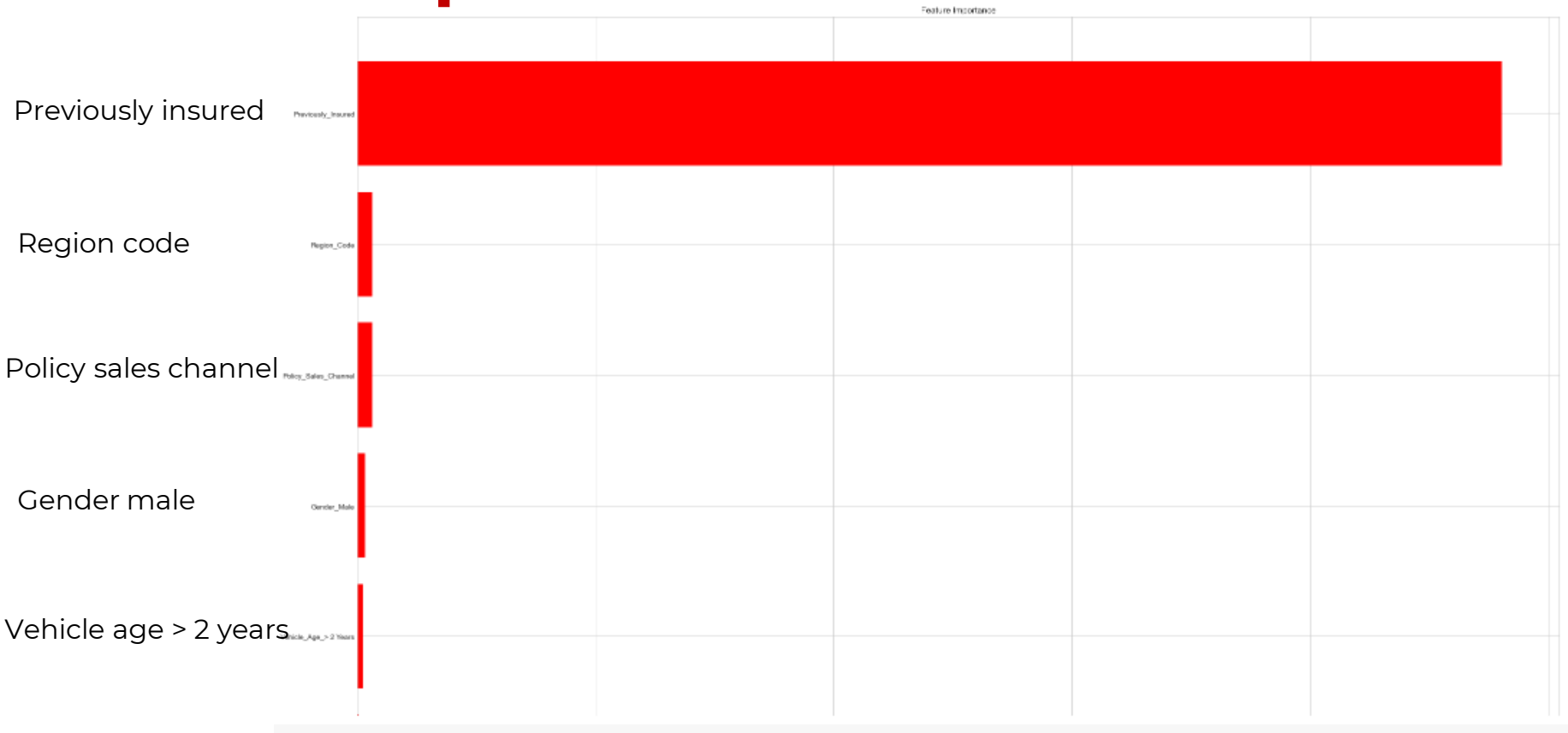

Fig 28: Feature importance graph From XG Boost.

Model implementation

Fig 29: Heat map of confusion matrix.

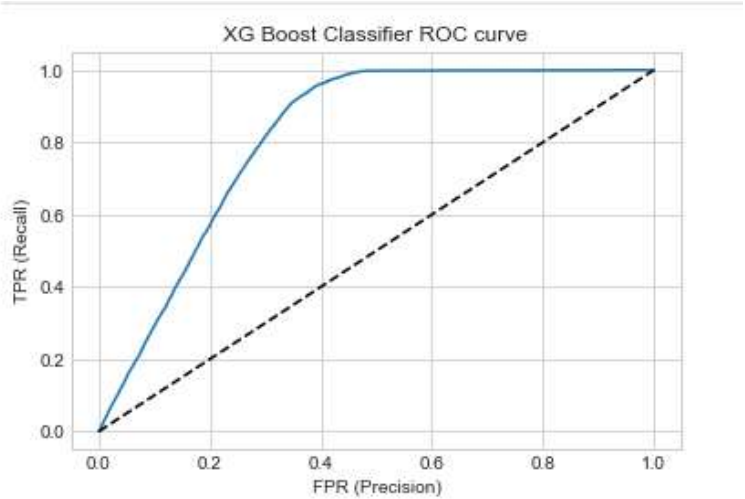

Fig 30: XG Boost Classifier ROC Curve

	Prediksi 1	Prediksi 0
Aktual 1	3190	6186
Aktual 0	7953	58893

Table 3: Values of Confusion matrix for XG Boost.

Conclusion

- There will be more profit if company sells both health and vehicle insurance.
- Previously insured is important feature for cross sell.
- After implementation of various models we got best results from Decision tree classifier, Random forest classifier and XG boost classifier.
- Decision tree and Random forest are over fitting so finally we decide to go with XG boost.
- XG boost with train accuracy as 0.88 and test accuracy of 0.81

References

[1] “A Study on a car Insurance purchase Prediction Using Two-Class Logistic Regression and Two-Class Boosted Decision Tree” by Su Hyun AN1 , Seong Hee YEO2 , Minsoo KANG3

[2]”A study on the meaning of automobile in the no insurance automobile injury insurance” by Choi, B. G

[3]”Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction” by Wagner A. Kamakura

Thank You...