



HEALTH INSURANCE CROSS SELL PREDICTION

AUFA BIAHDILLAH



CONTENT



01

DATA UNDERSTANDING & GOALS

02

DATA CLEANING & DATA MANIPULATION

03

EXPLORATORY DATA ANALYSIS

04

FEATURE ENGINEERING

05

MODEL TRAINING

06

MODEL EVALUATION

07

FEATURE IMPORTANCE AND PARTIAL DEPENDENCE PLOT

DATA UNDERSTANDING

A health insurance company has served health insurance to some customers. They want to expand their business to the vehicle insurance. Thus, by using the data they have, we have to predict whether the customers are interested in purchasing the vehicle insurance or not. The dataset has 381109 rows and 12 columns contains the information of the customers including:

- 0. id: Unique identifier for the Customer.
- 1. Age: Age of the customer.
- 2. Gender: Gender of the customer.
- 3. Driving_License: 0 for customer not having DL, 1 for customer having DL.
- 4. Region_Code: Unique code for the region of the customer.
- 5. Previously_Insured: 0 for customer not having vehicle insurance, 1 for customer having vehicle insurance.
- 6. Vehicle_Age: Age of the vehicle.
- 7. Vehicle_Damage: Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
- 8. Annual_Premium: The amount customer needs to pay as premium in the year.
- 9. Policy_Sales_Channel: Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
- 10. Vintage: Number of Days, Customer has been associated with the company.
- 11. Response: 1 for Customer is interested, 0 for Customer is not interested.

PROJECT GOALS

01

Analyze the information
obtained from the
dataset.

02

Train some models to
predict the response of
the customer towards
the vehicle insurance
whether they are
interested or not.

03

Evaluate the models and
decide which model is
the best for this study
case.

04

Give recommendations
about the future
development of the
model and future
improvement or
strategies of the
business.

DATA CLEANING



1. Missing Values

As we can see here, we do not have any missing values for this dataset.



2. Duplicate Values

After we do a checking, we obtain that there is no duplicate values in this dataset as well.

	Missing Values	Total	Percentage
id	0	0	0.0%
Gender	0	0	0.0%
Age	0	0	0.0%
Driving_License	0	0	0.0%
Region_Code	0	0	0.0%
Previously_Insured	0	0	0.0%
Vehicle_Age	0	0	0.0%
Vehicle_Damage	0	0	0.0%
Annual_Premium	0	0	0.0%
Policy_Sales_Channel	0	0	0.0%
Vintage	0	0	0.0%
Response	0	0	0.0%

DATA MANIPULATION



1.Encoding

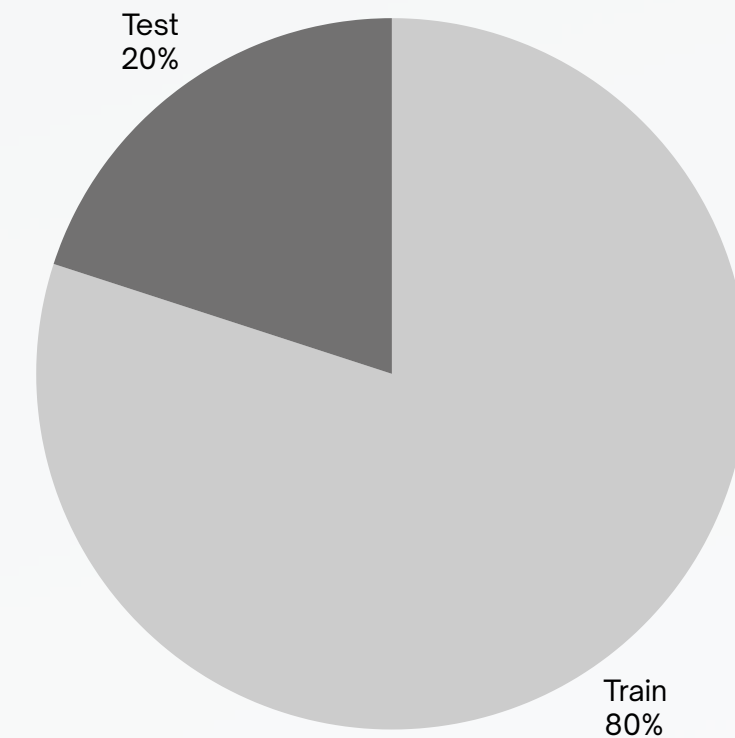
Since vehicle age can be ordered, we use label encoder. The gender and vehicle can also be encoded by 1 and 0

```
Gender:  
['Male' 'Female']  
  
Vehicle_Age:  
['> 2 Years' '1-2 Year' '< 1 Year']  
  
Vehicle_Damage:  
['Yes' 'No']
```



2.Train Test Split

We split the data into two only, 80% train dataset and 20% test dataset. We will use the train dataset for grid search cv.

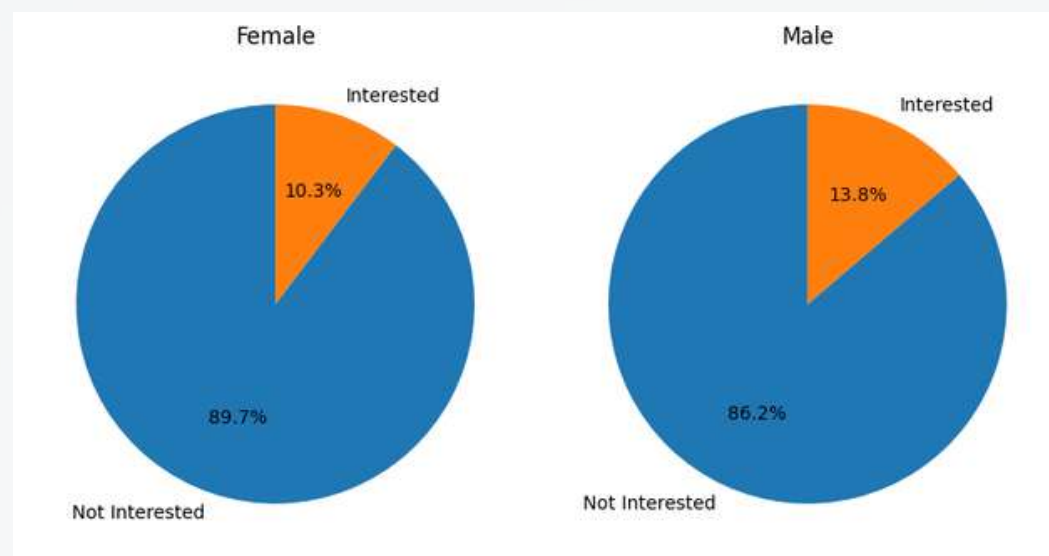
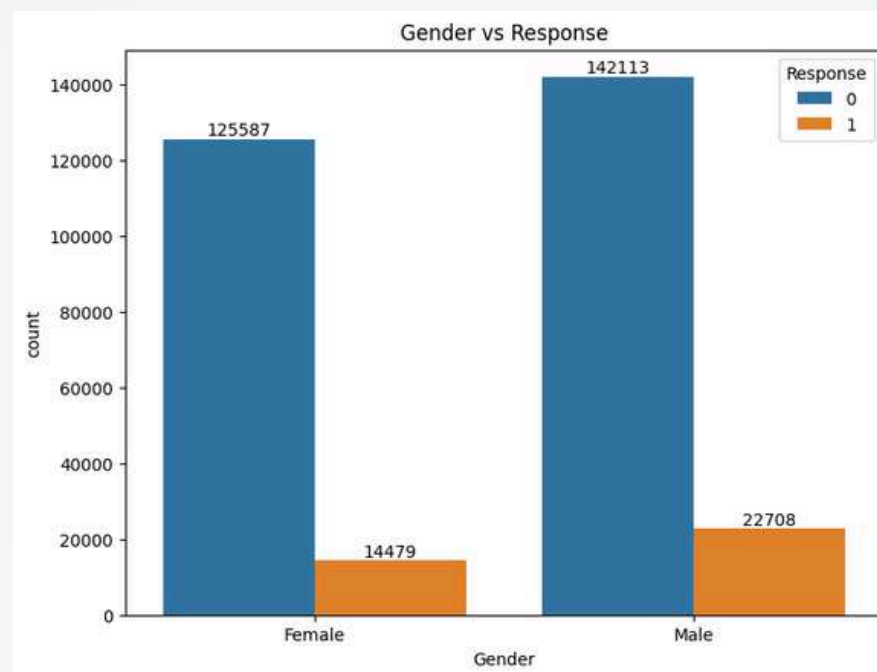


The top corners of the image feature decorative wavy lines. On the left, a series of thin, light yellow lines curve upwards and outwards. On the right, a series of thin, light yellow lines curve upwards and inwards.

EXPLORATORY DATA ANALYSIS

The central title bar is a solid black horizontal rectangle. On the left side of this bar, a series of thin, white wavy lines curve upwards and outwards. On the right side, a series of thin, white wavy lines curve upwards and inwards.

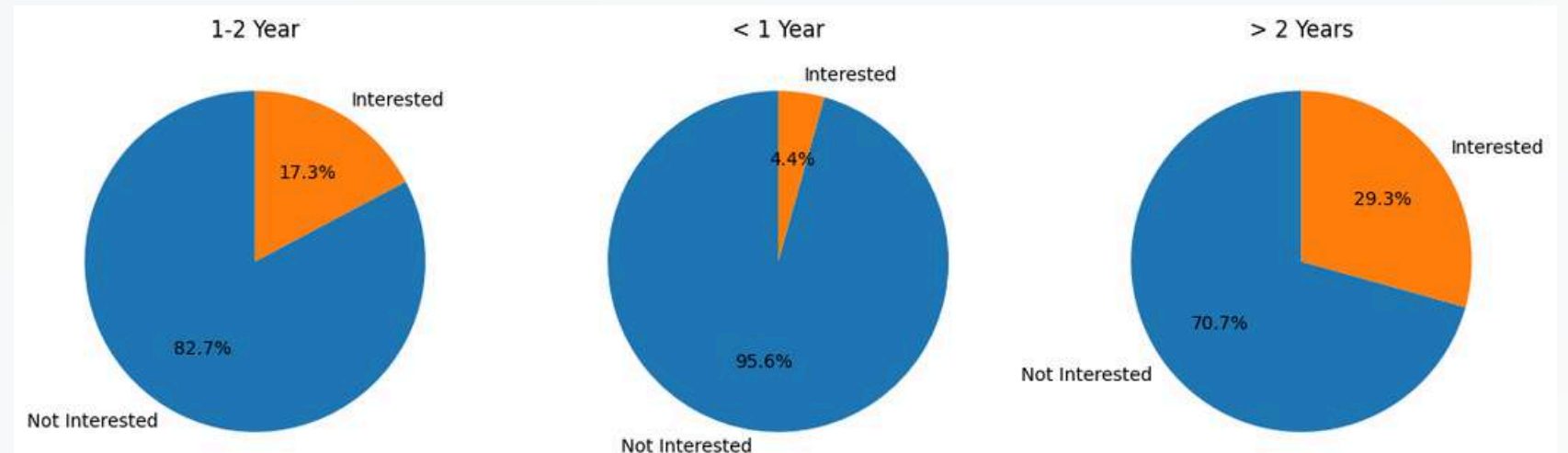
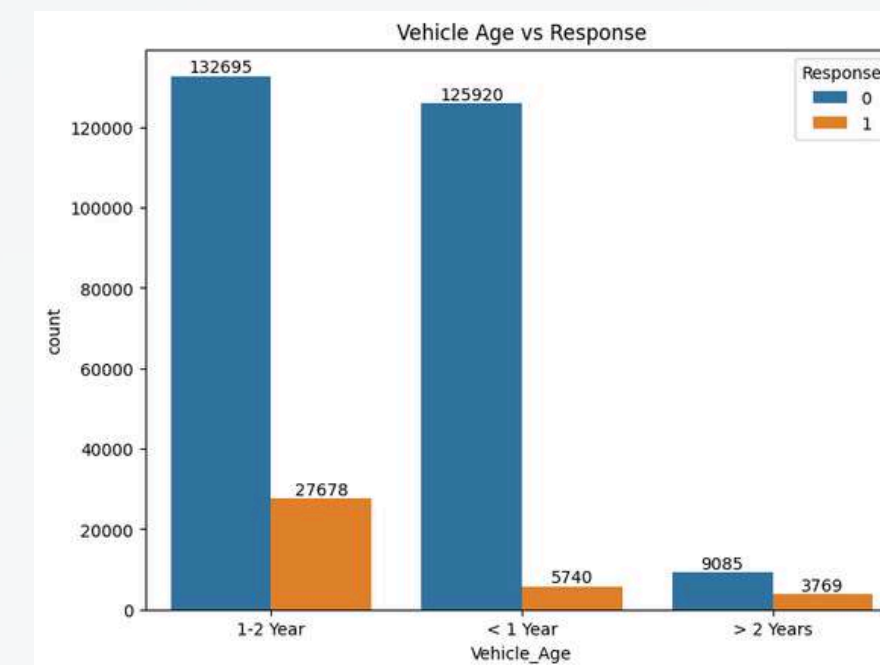
THE MALE CUSTOMER HAS HIGH PROBABILITY TO BE INTERESTED IN THE VEHICLE INSURANCE



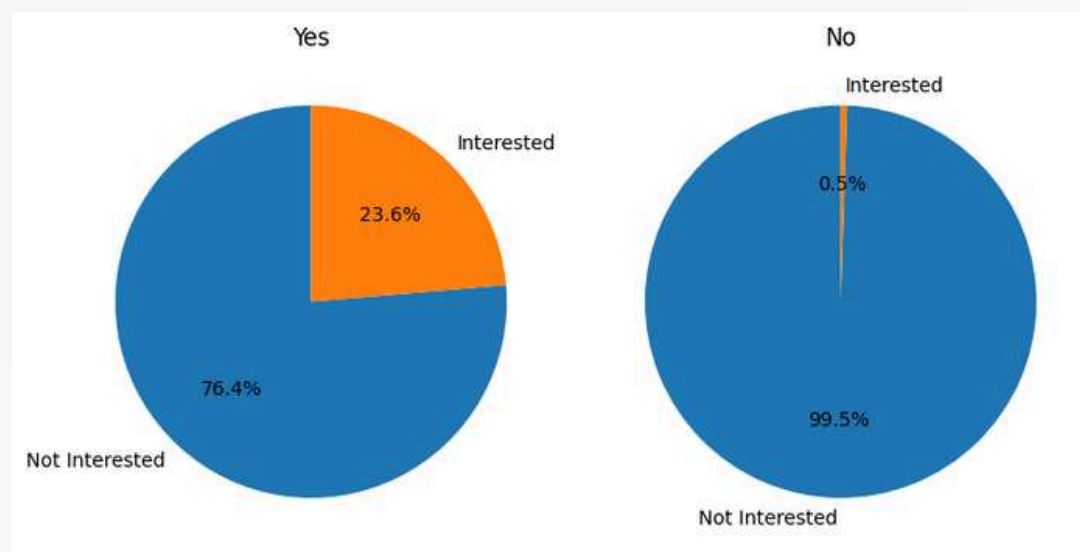
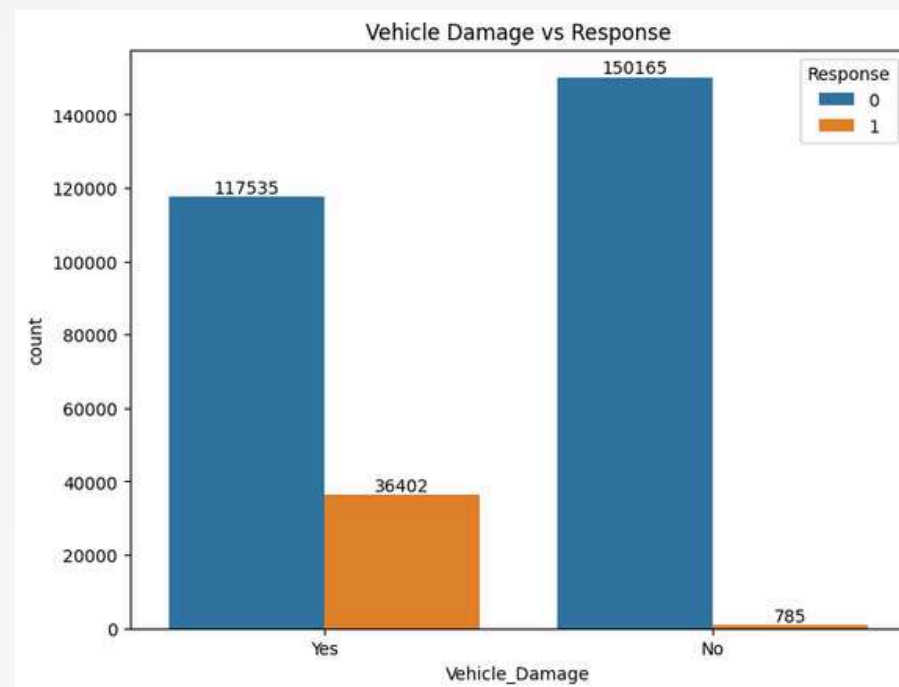
- We obtain that the male customer has higher chance of being interested in the vehicle insurance than female customer.
- Recommendation:
 1. Create a promo for masculine vehicle in order to attract male customer to subscribe our vehicle insurance.

CUSTOMER WITH OLD VEHICLE HAS HIGH RATE TO PURCHASE THE VEHICLE INSURANCE

- We see that customers whose vehicles are more than 2 years old are most likely to be interested in purchasing vehicle insurance.
- This means that many customers care about their old or even vintage vehicle
- Recommendation:
 1. Create a special offer for old or vintage vehicle insurance.
 2. Create a promo for a brand-new vehicle.



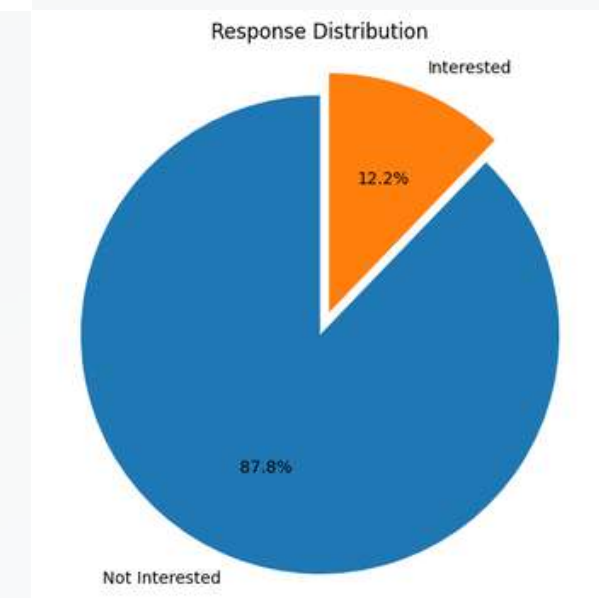
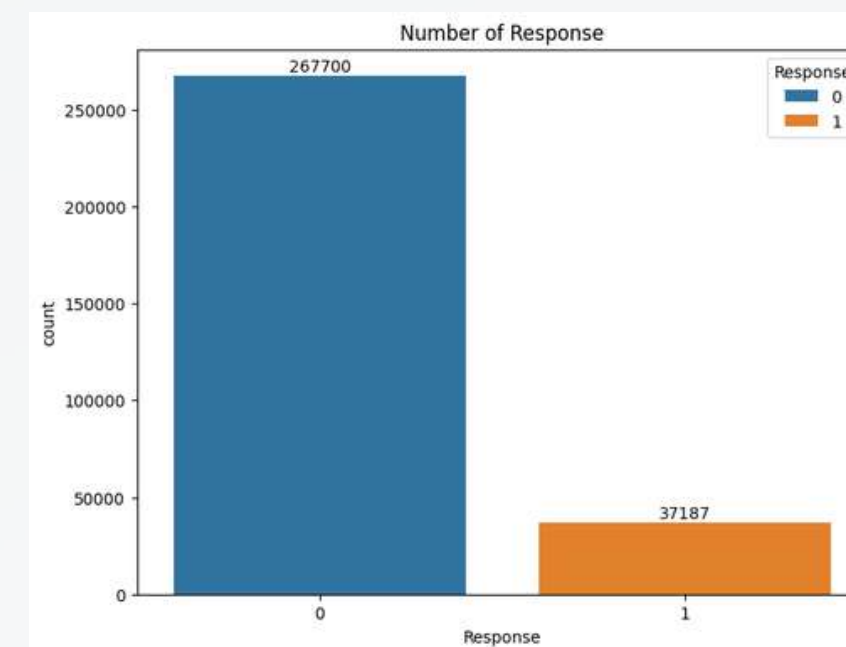
COMPANY SHOULD SEARCH FOR CUSTOMER WITH BRAND NEW VEHICLE



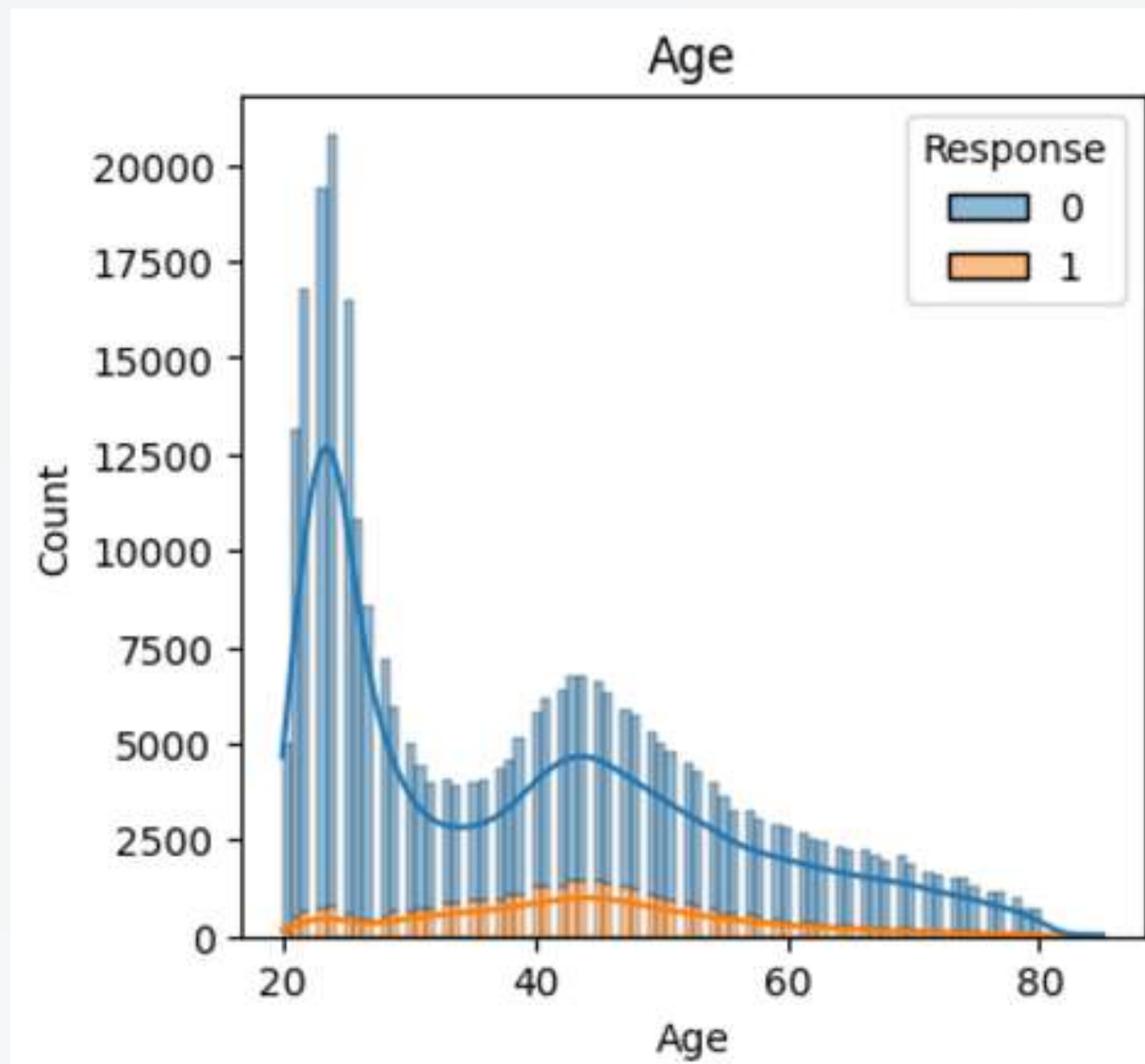
- The majority of customer who are interested in the vehicle insurance are those whose vehicle has been damaged previously.
- This may harm the company income since the previously damaged vehicle has high chance to be injured again in the future.
- Thus, company should find for customer with brand new vehicle to cover this expense (cross subsidy).
- Recommendation
 1. Create a promo for customer with brand new vehicle such as free 3 months subscription

THE RESPONSE DATA IS IMBALANCE WHICH CAN AFFECT THE PERFORMANCE OF THE MODEL

- We know that our data is imbalance, that is the number of interested data is far less than the not interested one.
- Thus, since our purpose is classification, then this may be a problem.
- Recommendation:
 1. Obtain more interested customer data or we can simply do oversampling for the model training.
 2. Wisely choose the main metric for measuring the model performance. In this case, f1 score is better since we try to avoid both predicting people who are actually not interested as interested (False Positive) to optimize our marketing and predicting people who are actually interested as not interested to avoid missing getting new customer.



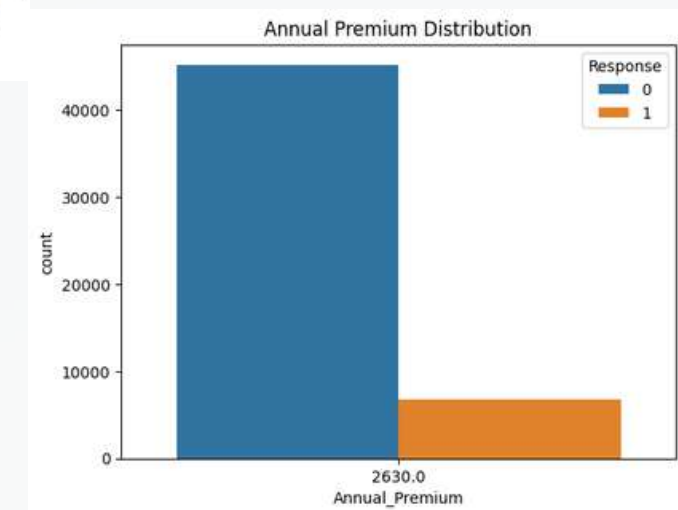
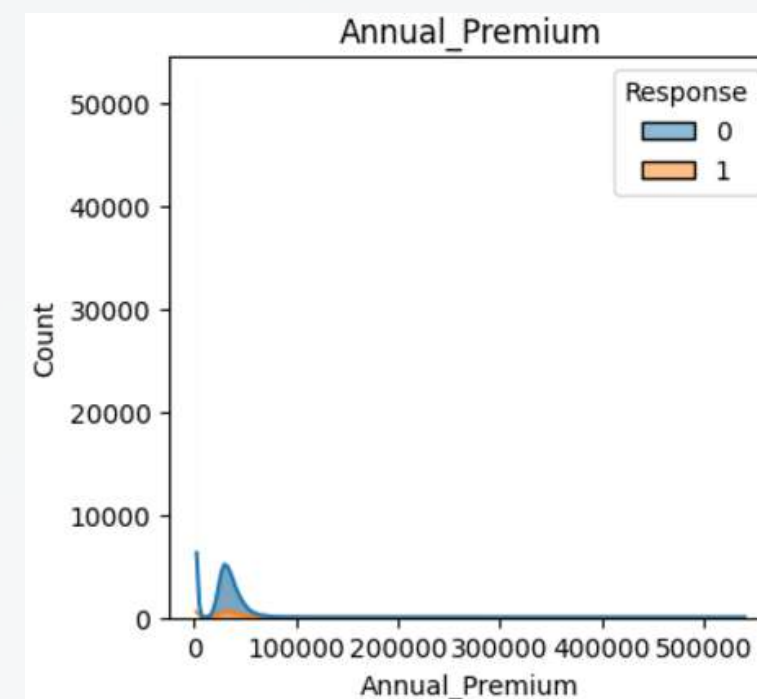
EVEN THOUGH YOUNG ADULT IS THE MAJORITY OF THE CUSTOMER, THEY ARE LESS INTERESTED IN VEHICLE INSURANCE



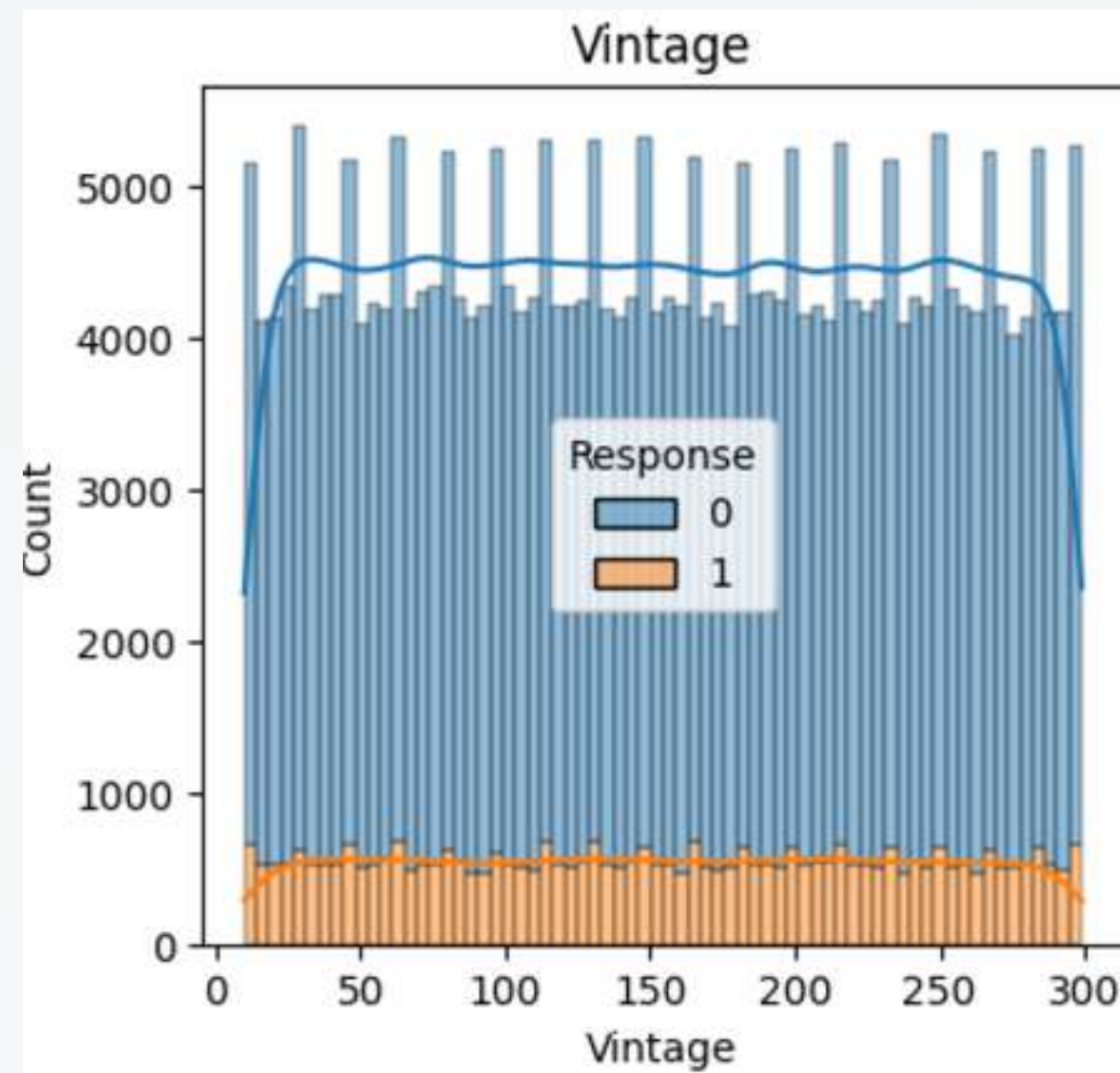
- We can see that the customers who are interested in vehicle insurance are mostly those whose age is in range (40,45).
- On the contrary, even though young adult whose age is in range (20,25) is the majority of the customer, they are mostly not interested in the vehicle insurance.
- Recommendations:
 1. We need to educate the young adult more about the advantages of vehicle insurance so that they become interested.
 2. Update the marketing strategy so it fits with the young generation's interest.

DESPITE THE HIGHEST NUMBER OF INTERESTED CUSTOMER AT ANNUAL PREMIUM = 2630, IT STILL NEEDS MARKETING IMPROVEMENT

- In the annual premium, there is a peak that is far more than others which is at annual premium = 2630.
- The number of interested customer in this value is also the highest compare to others.
- Since this value is the smallest annual premium, it can be the reason why there are a lot of people are interested in the vehicle insurance.
- However, the interested rate is still low and need marketing improvement to attract more people.



THE NUMBER OF INTERESTED CUSTOMER BASED ON THE VINTAGE IS ALMOST CONSTANT IN EACH VALUE



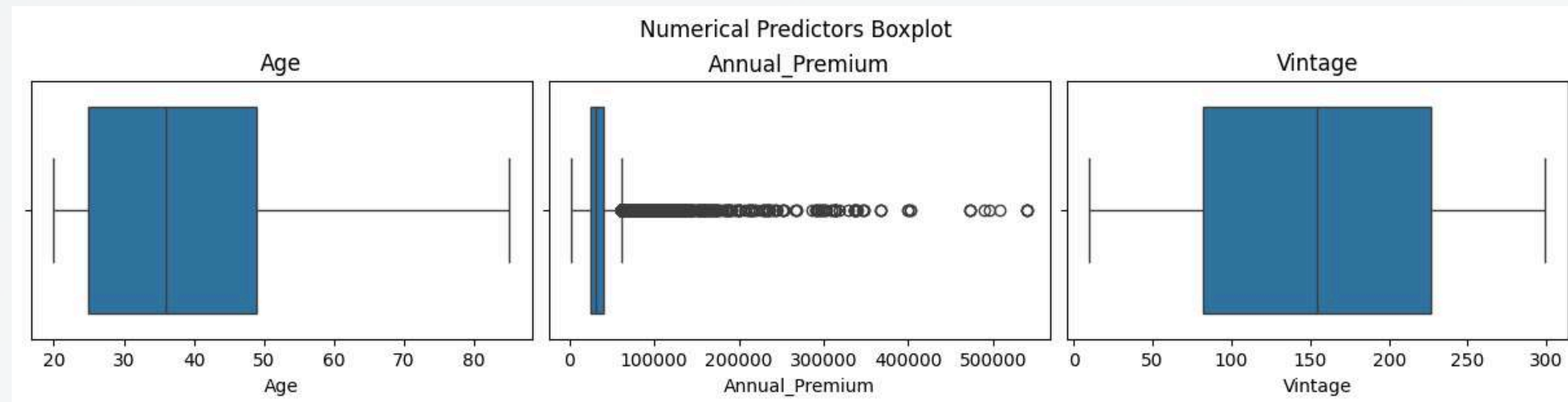
- The vintage shows relatively constant relations, that is the proportions of interested and not interested in each vintage are relatively the same.
- This may indicate that vintage does not really affect whether the customer is interested in the vehicle insurance or not.
- However, this indication requires verification which can be done by using partial dependence plot in the next section.
- Recommendation:
 1. Give a discount for loyal customer.

The top corners of the image feature decorative wavy lines. On the left, a series of thin, light yellow lines curve upwards and outwards. On the right, a series of thin, light yellow lines curve upwards and inwards.

FEATURE ENGINEERING

The central title bar is a solid black horizontal rectangle. On the left side of this bar, a series of thin, white lines curve upwards and outwards. On the right side, a series of thin, white lines curve upwards and inwards.

OUTLIERS HANDLING



The numerical column that has many outliers is only Annual_Premium. We decided to handle it using log transformation to avoid dropping too much data.

MULTICOLLINEARITY

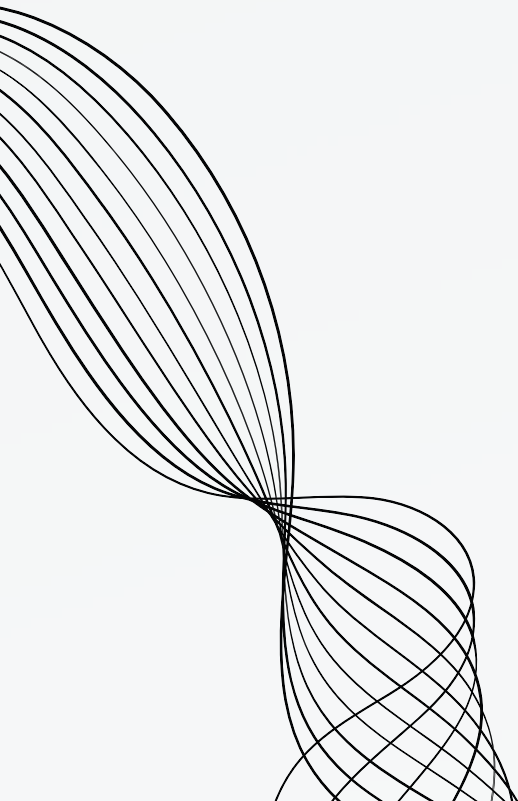
	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
Gender	1.000000	-0.145359	0.019106	-0.001186	0.082523	-0.160743	-0.093530	0.014376	0.111134	0.002760	-0.052387
Age	-0.145359	1.000000	-0.080803	0.042880	-0.254233	0.765534	0.267424	-0.009304	-0.576785	-0.002283	0.111053
Driving_License	0.019106	-0.080803	1.000000	-0.002273	0.015195	-0.037240	-0.017019	-0.006820	0.044244	-0.001412	0.010280
Region_Code	-0.001186	0.042880	-0.002273	1.000000	-0.023018	0.043748	0.027196	-0.000683	-0.042481	-0.003431	0.011651
Previously_Insured	0.082523	-0.254233	0.015195	-0.023018	1.000000	-0.380911	-0.824565	0.055391	0.220117	0.002610	-0.340417
Vehicle_Age	-0.160743	0.765534	-0.037240	0.043748	-0.380911	1.000000	0.396764	-0.038233	-0.550666	-0.002039	0.221079
Vehicle_Damage	-0.093530	0.267424	-0.017019	0.027196	-0.824565	0.396764	1.000000	-0.040960	-0.224957	-0.002402	0.353340
Annual_Premium	0.014376	-0.009304	-0.006820	-0.000683	0.055391	-0.038233	-0.040960	1.000000	-0.061248	-0.002044	0.002727
Policy_Sales_Channel	0.111134	-0.576785	0.044244	-0.042481	0.220117	-0.550666	-0.224957	-0.061248	1.000000	-0.000127	-0.137533
Vintage	0.002760	-0.002283	-0.001412	-0.003431	0.002610	-0.002039	-0.002402	-0.002044	-0.000127	1.000000	-0.001025
Response	-0.052387	0.111053	0.010280	0.011651	-0.340417	0.221079	0.353340	0.002727	-0.137533	-0.001025	1.000000

After checking the correlation matrix, we obtain that Previously Insured and Vehicle Damage has correlation with absolute value greater than 0.8. Thus, we can drop one of both. Because the absolute value of correlation of Vehicle Damage and Response is higher than Previously Insured and Response, we decided to drop Previously Insured.



SCALING

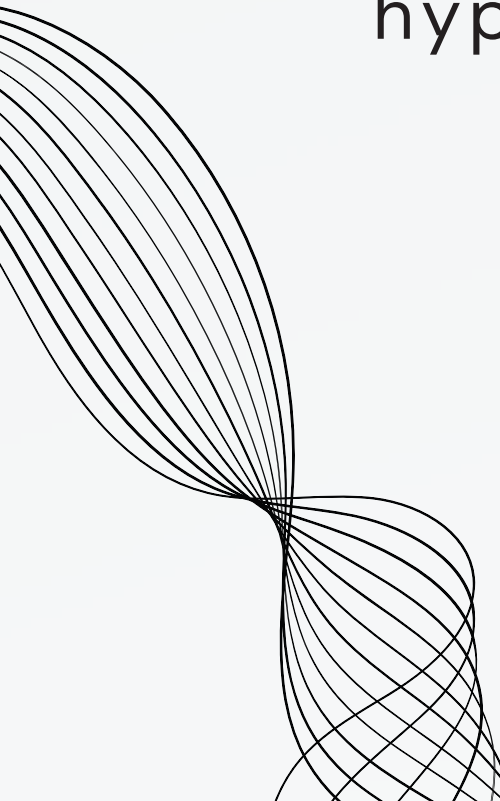
Since we want to get a better interpretation during the partial dependence plot and mainly apply tree-based models, we decided not to use any scaling for this study case. However, we will still use a standard scaler for the logistic regression to optimize its result using a pipeline.


$$z = \frac{x - \mu}{\sigma}$$



SAMPLING

As we know, the target data is imbalance, that is the number of interested customer is far less than the number of uninterested customer. Therefore, we will experiment with SMOTE, No Sampling, and Undersampling during the hyper parameter tuning by using pipeline to avoid data leakage.



The background features a light gray gradient with abstract, wavy, golden-yellow lines in the upper corners and a solid black horizontal band across the middle. The text is centered within this black band.

MODEL TRAINING

WITH SMOTE

- We train four models: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and XGB Classifier.
- We use a pipeline with smote to avoid data leakage.
- The hyperparameters that we set for the grid search cv are:
 1. Logistic regression: 'smote__sampling_strategy': [0.5, 0.75, 1.0], 'model__C': [0.1, 1, 10], 'model__solver': ['liblinear', 'lbfgs']
 2. Random forest classifier: 'smote__sampling_strategy': [0.5, 0.75, 1.0], 'model__n_estimators': [100, 200], 'model__max_depth': [None, 10, 20]
 3. Decision tree classifier: 'smote__sampling_strategy': [0.5, 0.75, 1.0], 'model__max_depth': [None, 10, 20], 'model__criterion': ['gini', 'entropy']
 4. XGB classifier: 'smote__sampling_strategy': [0.5, 0.75, 1.0], 'model__n_estimators': [100, 200], 'model__max_depth': [3, 6], 'model__learning_rate': [0.01, 0.1, 0.3]

WITHOUT SAMPLING

- We train four models: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and XGB Classifier.
- The hyperparameters that we set for the grid search cv are:
 1. Logistic regression: 'C': [0.1, 1, 10], 'solver': ['liblinear', 'lbfgs']
 2. Random forest classifier: 'n_estimators': [100, 200], 'max_depth': [None, 10, 20]
 3. Decision tree classifier: 'max_depth': [None, 10, 20], 'criterion': ['gini', 'entropy']
 4. XGB classifier: 'n_estimators': [100, 200], 'max_depth': [3, 6], 'learning_rate': [0.01, 0.1, 0.3]

WITH UNDERSAMPLING

- We train four models: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and XGB Classifier.
- We use a pipeline with a random under-sampler to avoid data leakage.
- The hyperparameters that we set for the grid search cv are:
 1. Logistic regression: 'undersampling__sampling_strategy': [0.5, 0.75, 1.0], 'model__C': [0.1, 1, 10], 'model__solver': ['liblinear', 'lbfgs']
 2. Random forest classifier: 'smote__sampling_strategy': [0.5, 0.75, 1.0], 'model__n_estimators': [100, 200], 'undersampling__max_depth': [None, 10, 20]
 3. Decision tree classifier: 'undersampling__sampling_strategy': [0.5, 0.75, 1.0], 'model__max_depth': [None, 10, 20], 'model__criterion': ['gini', 'entropy']
 4. XGB classifier: 'undersampling__sampling_strategy': [0.5, 0.75, 1.0], 'model__n_estimators': [100, 200], 'model__max_depth': [3, 6], 'model__learning_rate': [0.01, 0.1, 0.3]

The background features a light gray field with abstract, wavy, golden-yellow lines in the upper corners and a solid black horizontal band across the middle. The text 'MODEL EVALUATION' is centered within this black band in a bold, white, sans-serif font.

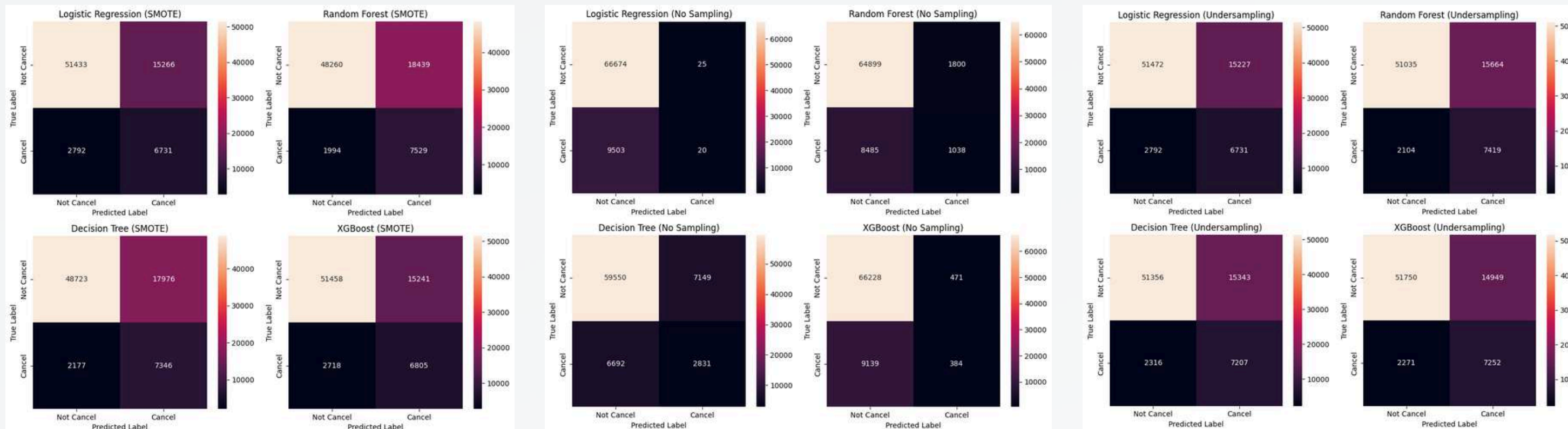
MODEL EVALUATION

METRIC SCORES

	Model	F1 Train	F1 Test	Recall Train	Recall Test	Precision Train	Precision Test	AUC Train	AUC Test
0	Logistic Regression (SMOTE)	0.415119	0.427094	0.699061	0.706815	0.295211	0.305996	0.819237	0.823458
1	Random Forest (SMOTE)	0.420754	0.424277	0.799231	0.790612	0.285537	0.289934	0.828794	0.827781
2	Decision Tree (SMOTE)	0.415432	0.421639	0.773550	0.771396	0.283968	0.290103	0.823553	0.820085
3	XGBoost (SMOTE)	0.421797	0.431119	0.711673	0.714586	0.299717	0.308673	0.828685	0.832139
4	Logistic Regression (No Sampling)	0.003958	0.004181	0.001990	0.002100	0.360976	0.444444	0.819076	0.823336
5	Random Forest (No Sampling)	0.999260	0.167948	0.998736	0.108999	0.999785	0.365751	1.000000	0.825073
6	Decision Tree (No Sampling)	0.999448	0.290314	0.998897	0.297280	1.000000	0.283667	1.000000	0.595059
7	XGBoost (No Sampling)	0.147485	0.074003	0.081346	0.040323	0.788993	0.449123	0.889418	0.845743
8	Logistic Regression (Undersampling)	0.414993	0.427623	0.698040	0.706815	0.295266	0.306540	0.819282	0.823437
9	Random Forest (Undersampling)	0.452453	0.455070	0.791863	0.779061	0.316706	0.321405	0.855852	0.848905
10	Decision Tree (Undersampling)	0.444641	0.449412	0.763842	0.756799	0.313594	0.319601	0.847856	0.839711
11	XGBoost (Undersampling)	0.449248	0.457193	0.763977	0.761525	0.318173	0.326652	0.850617	0.850206

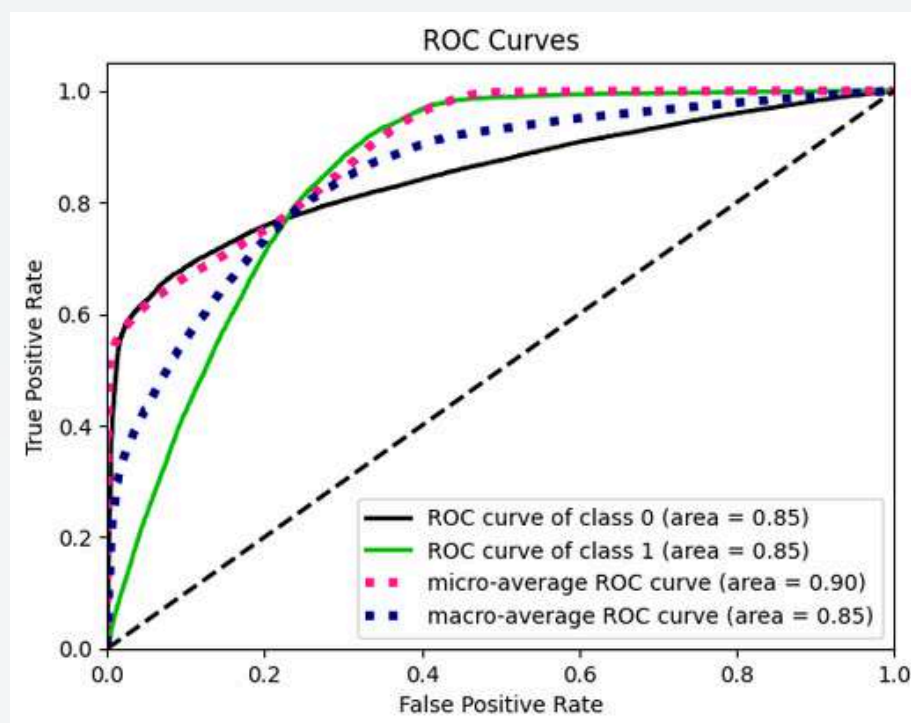
We attained that all of the models with no sampling are either overfit nor underfit while the models with SMOTE or Undersampling are best fit. Among all of this models, the best model for this study case is the XGB Classifier with undersampling because f1 scores in the test dataset is the highest.

CONFUSION MATRIX



- All of the models varied in the prediction result, with the models without any sampling having the smallest number of true positive predictions. Both SMOTE and Undersampling can increase the number of true positives. However, as a consequence, the number of false positives also increases.
- Compared to the models with SMOTE, the models with Undersampling have fewer False Positives but also fewer True Positives, indicating that they have higher Precision scores and lower Recall scores. This is why the F1 scores of the models with undersampling in the test dataset are the highest. Because we aim to optimize the F1 Score, we decided that the XGB with undersampling is the best model for this study case.

ROC CURVE



	FPR	TPR	Threshold
12382	0.299990	0.881865	0.373303
12381	0.299885	0.881865	0.373436
12380	0.299885	0.881760	0.373456
12379	0.299855	0.881760	0.373593
12378	0.299765	0.881550	0.373600
...
9223	0.200093	0.707550	0.529834
9222	0.200078	0.707550	0.529851
9221	0.200078	0.707445	0.529861
9220	0.200033	0.707445	0.529933
9219	0.200033	0.707340	0.529953

- The figure illustrates the ROC Curve of the selected model: XGBClassifier with random under sampler.
- The curve shows that the model still needs improvement since it does not fit the upper left corner.
- Based on the ROC curve, we try to find the best threshold to optimize the selected model's F1 score. Thus, we will filter the FPR, TPR, and Threshold data frame to the intersection of the lines, that is, to FPR between 0.2 and 0.3.
- If we calculate the median and mean of the data frame, we find that the threshold is around 0.47.
- Then, we ensure the best threshold by checking for each value from 0 to 0.01, ... to 1, and we obtain the best F1 score when the threshold is 0.48.
- Hence, we decided to choose XGBClassifier with undersampling and threshold 0.48.

FUTURE DEVELOPMENT

Based on the previous result, there are some recommendations for future developments of the prediction model.

Experiment with more hyperparameters during the hyperparameter tuning stage using grid search cross validation technique.

RECOMMENDATION 1

Create more new features from the existing features.

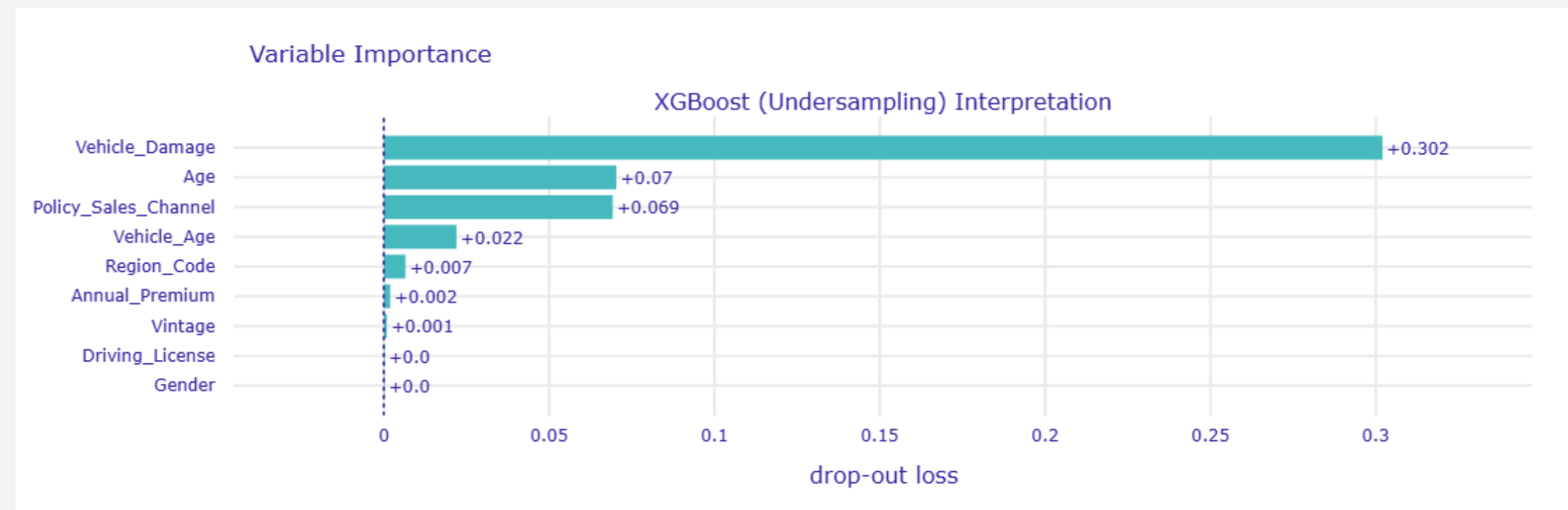
RECOMMENDATION 2

We can use other encoding techniques, such as one hot encoding. Furthermore, we can also experiment with other scaling techniques such as robust scaler.

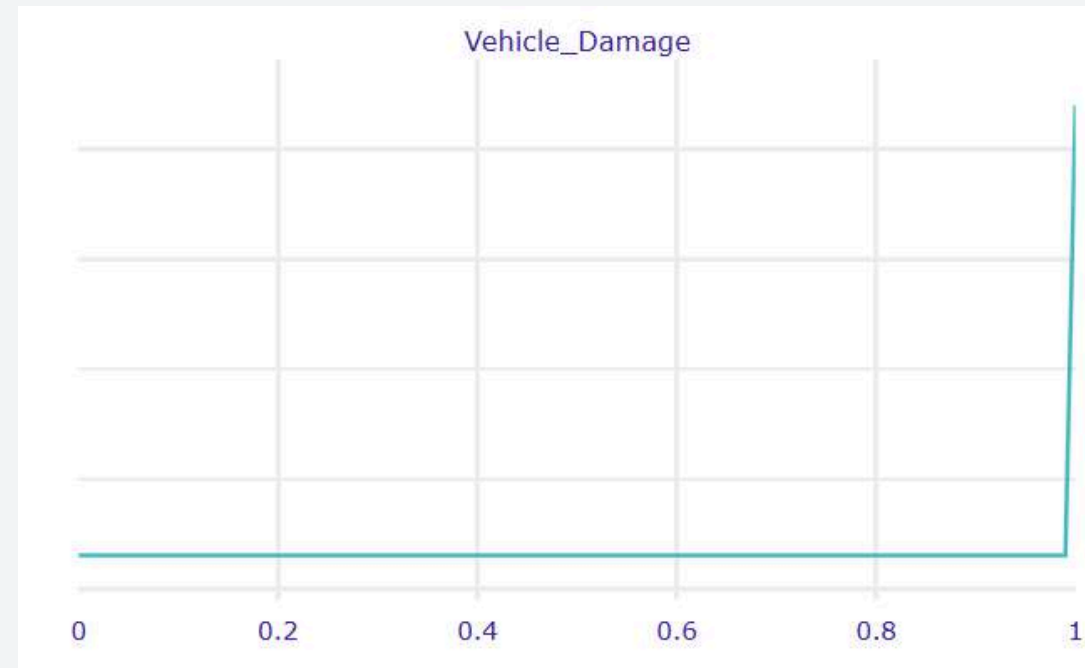
RECOMMENDATION 3

FEATURE IMPORTANCE ANALYSIS

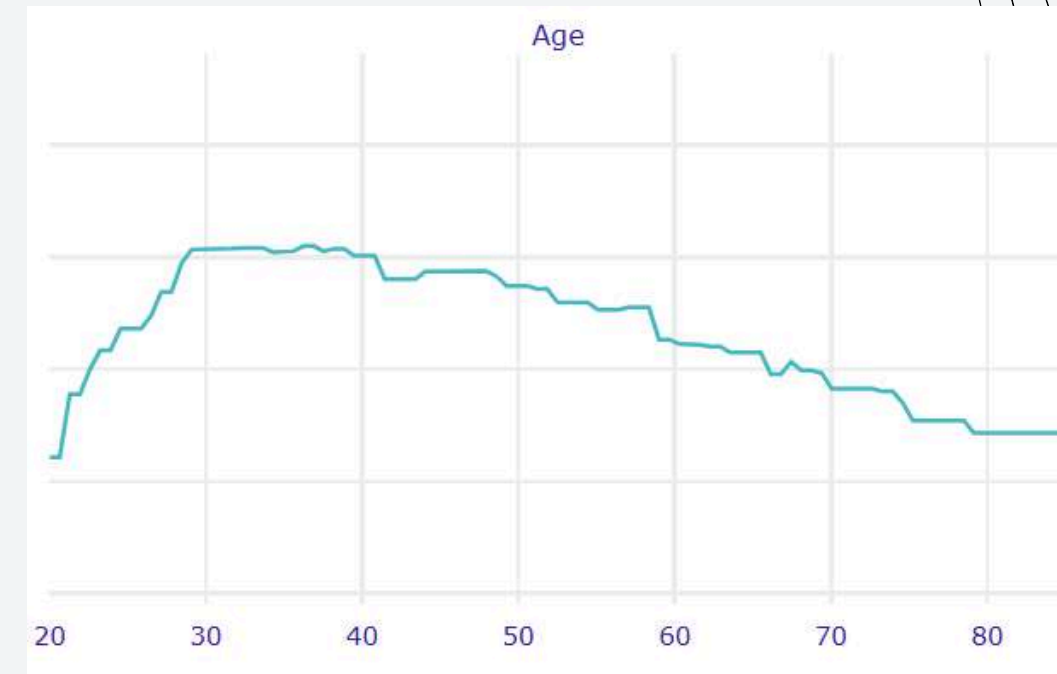
Based on the visualization above, we obtain that the top four most influential features for the model's prediction are vehicle_damage, age, policy sales channel, and vehicle_age. To know more about the direction of each factor, we can visualize it using a partial dependence plot.



PARTIAL DEPENDENCE PLOT



The partial dependence plot has a positive influence, indicating that the more vehicle damage, the higher the chance of the customer being interested.

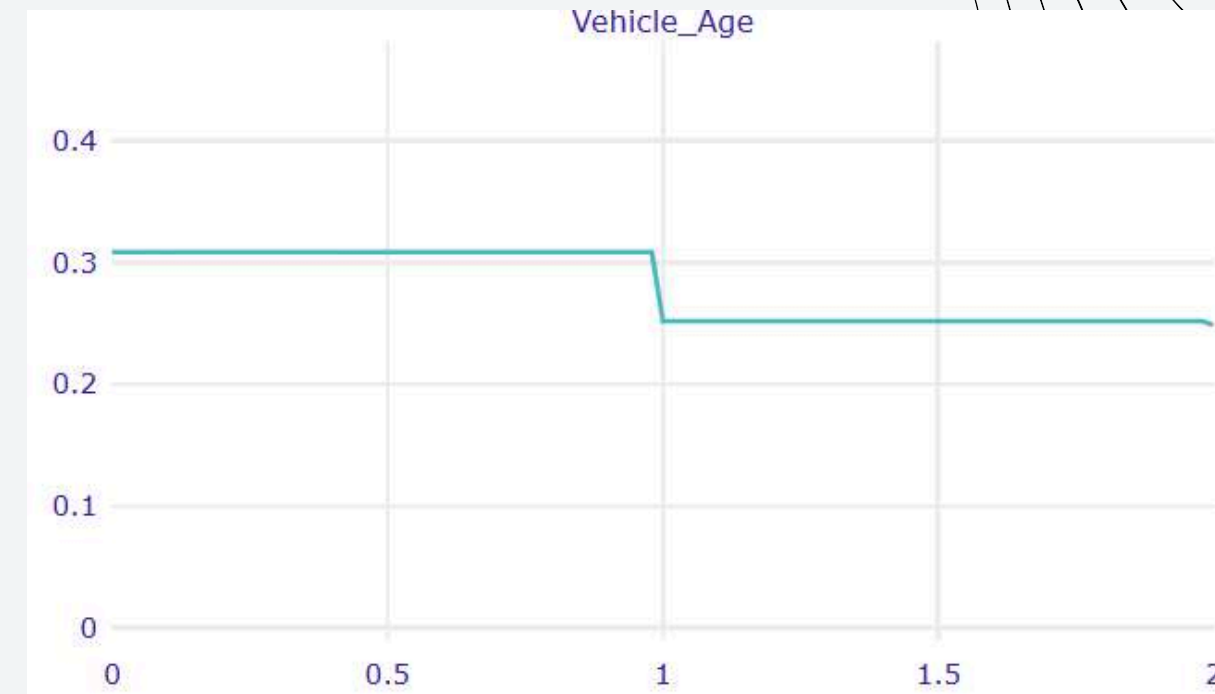


The graphic shows a decreasing trend, which means that the older the customer, the less likely the customer is to be interested. On the contrary, there is a positive influence on the younger generation, especially at age 20 – 30.

PARTIAL DEPENDENCE PLOT



The chart oscillates in each policy sales channel with some increasing trend at some points such as channel number 2, 25, 122, 145, 158, and 163.



After reaching vehicle age category 1 and above (1-2 years and >2 years), the probability of the customer being interested in vehicle insurance decreases.

RECOMMENDATION

Derived from the feature importance analysis, we can give some suggestions for the company to improve the business.



Attract more customers with previously damaged vehicles by giving free monthly vehicle inspection to get to know the damage earlier.

RECOMMENDATION 1



Adjust the marketing technique so that it is suitable for the young generation's latest trends.

RECOMMENDATION 2

RECOMMENDATION

Derived from the feature importance analysis, we can give some suggestions for the company to improve the business.



Do more research about these policy sales channels, especially channels 2, 25, 122, 145, 158, and 163 to know more about the reasons.

RECOMMENDATION 3



Give special offers/promotions for customer with brand new vehicles by giving three months subscription discount.

RECOMMENDATION 4

THANK YOU!

