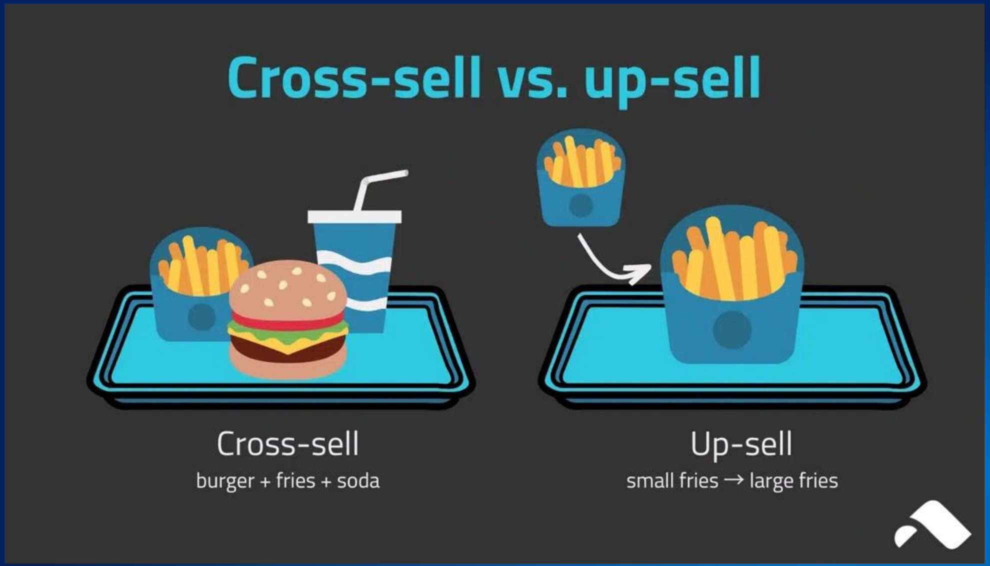
Predicting Vehicle Insurance Cross-Selling Interest Among Existing Health Insurance Customers

Our client is an Insurance company that has provided Health Insurance to their customers, now they need your help in building a model to predict whether the customers from past year will also be interested in Vehicle Insurance.

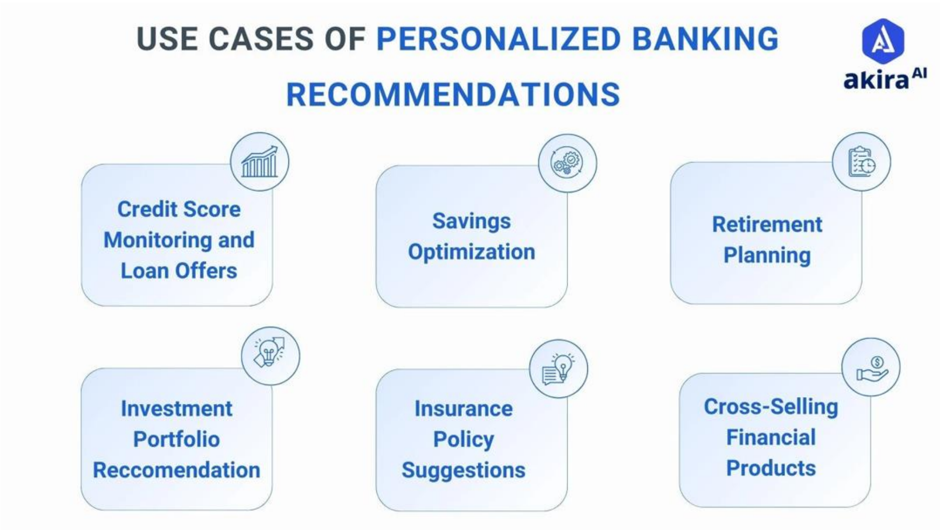


Variable	Definition
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Driving_License	1 : Customer already has DL, 0 : Customer does not have DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1 : Customer already has Vehicle Insurance, 0 : 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
PolicySalesChannel	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

Sorawit Huang (Joe)

Data Source : https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction?select=train.csv

Project Scope and Agenda



Introduction:

Cross-selling prediction is an AI application expected in the banking industry occurs when a **bank attempts to sell an existing customer additional financial products** like insurances, cards, auto loans, or investment services.

> This project showcased how a comprehensive data science framework could inform strategic decision-making and optimize conversion rates.

Sorawit Huang

Today's Agenda:

1) **EDA Insights and Target Customer Analysis**
2) **Data Preprocessing and Model Pipelines**
3) **Feature Selection for Model Training**
4) **Model Optimization & Evaluation Metric**
5) **Conclusion and Implementation**

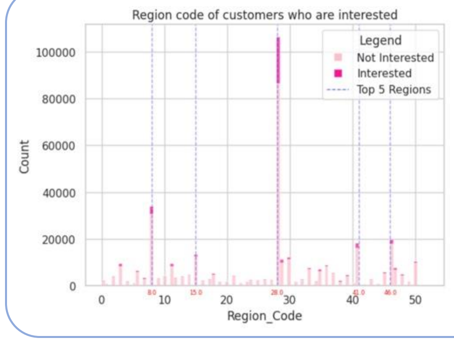
Customer Insights from EDA

381,109 Existing health insurance Customers:

12.23 %
Interested in cross-selling
(imbalance dataset)

99.7%
Interested customers have a driving license

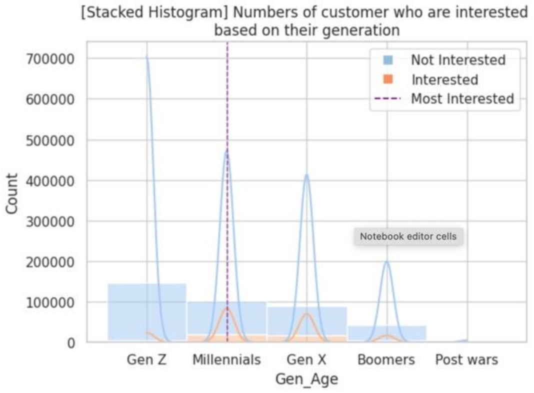
75 % come from 4 Sales Channel : 26,124,152,156



Region Code

The top 5 region codes have the highest number both existing and interested customers.

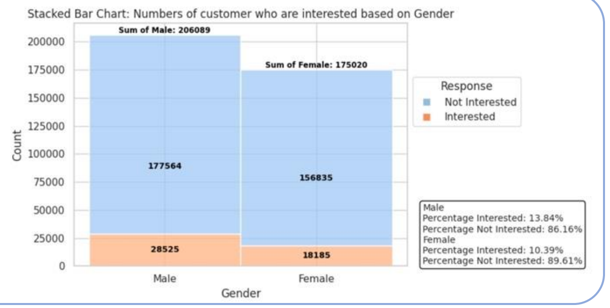
Age Class :

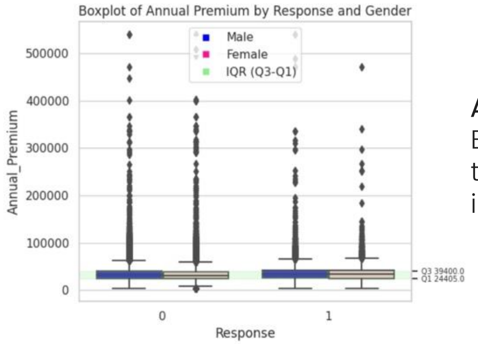


Most existing customers are Gen Z but showing less of interest on cross-selling.

Millennials and Gen X are majority of interested customers .

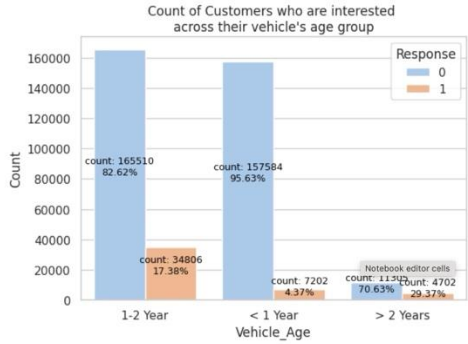
Gender : male customers show slightly higher interest percentage



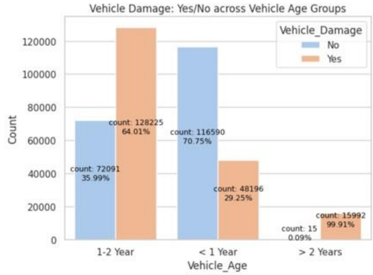


Annual Premium :
Both gender spend nearly the same annual premiums in range of IQR

Trends across Vehicle features

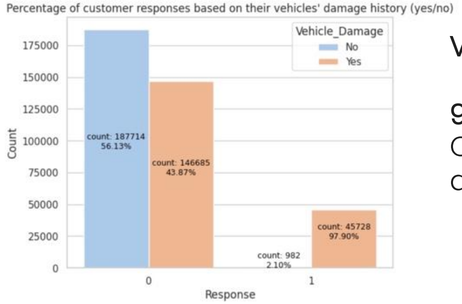


Higher vehicle age
Higher of expressing interest in vehicle insurance.



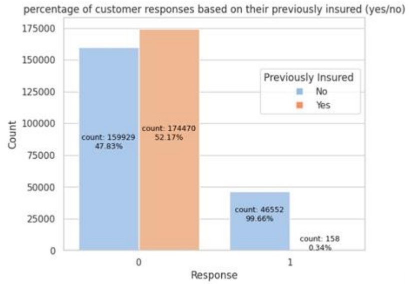
Older vehicle age, more probability of damage

99.9% of vehicle aged more than 2 year has their car got damaged.



Vehicle Damage :

97.90% Interested Customer experienced damage to their cars



Previously Insured
99% of interest customers haven't had previous insured Vehicle Insurance.

Customer Insights from EDA

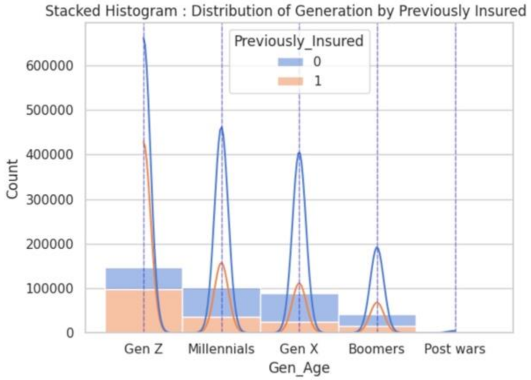
Target Customer Personas

Customer Profile:

- Eligible Driver
- Millennials and Gen X
- Male
- Never has any previous vehicle insurance
- Got their vehicle damaged in the past
- Vehicle aged more than 1 year

Potential Channels and Areas for Engagement:

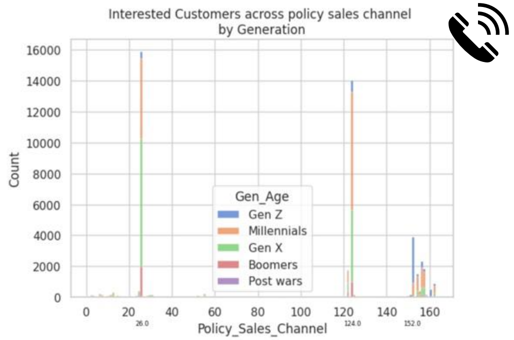
- Policy Sales Channel: 26, 124
- Region Code: 28



While most **Gen Z** have already had previously insured.

Majority of **Millennials and Gen X** haven't had previous insured Vehicle Insurance.

Why? Potential Areas& Channels

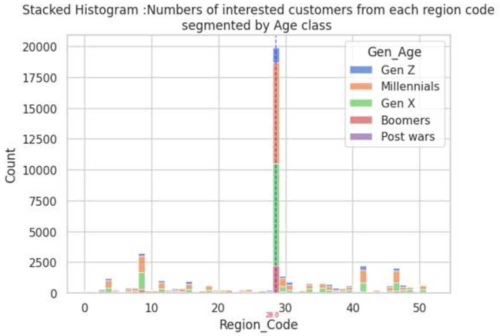


we should prioritize

Sales Channel 26, 124
highest concentration of Millennials and Gen X

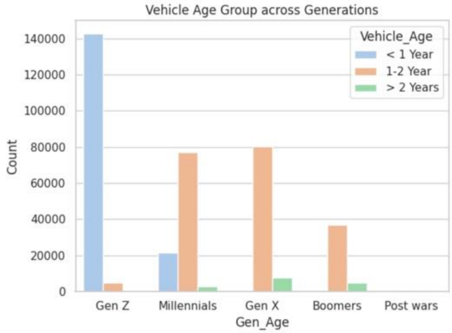
Sales Channel 152 :

- attract almost Gen Z customer
- Also, limitations for Boomers and Gen X

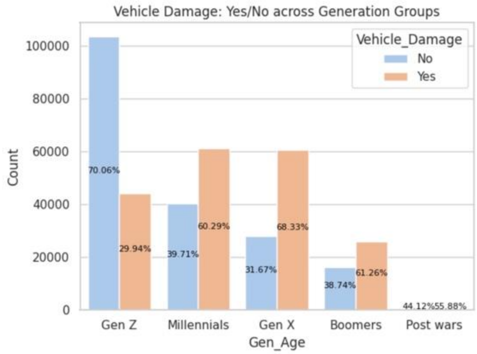


we should prioritize **Region Code 28**

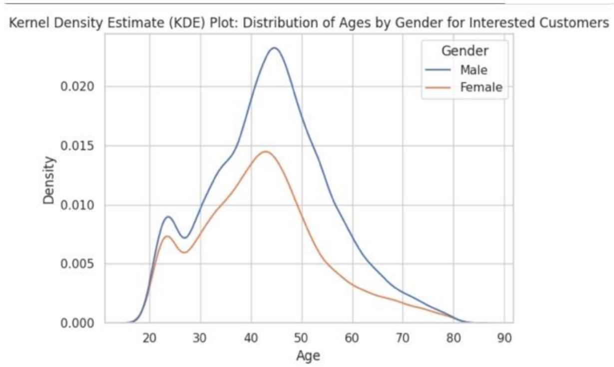
- highest concentrated of Millennials and Gen X
- Also, show highest vehicle's damage



The higher vehicle age seems correlate with higher owner's age



Majority of Gen Z tends to own non-damaged vehicles while almost **of Millennials and Gen X**



Notable of **MALE** interested customers who aged over 30 which are **Millennials, Gen X and boomers**

Data Preprocessing

Cleaning data

no missing values and duplicate rows.

Outliers Handling

Only outliers detected on **Annual_Premium** , we would cut those extreme values from the dataset.

Feature Engineering :

Categorical Features Encoding

Gen_Age, Vehicle_Age : as their trends seem having correlation with other features which increase and decrease ordinally so, we choose **Ordinal Encoder**

One Hot Encoder : 'Region_Code', 'Policy_Sales_Channel'

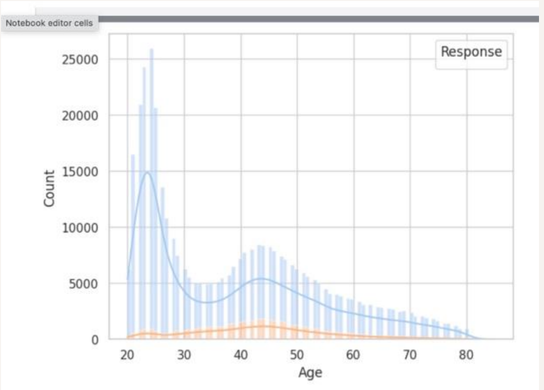
Log and Polynomial Transformation

- Log(Age) , Age^2
- Log(Annual_Premium), Annual_Premium^2

Binning techniques :

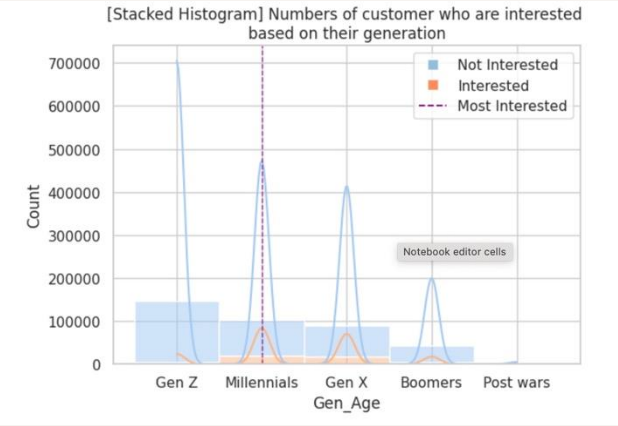
- Age binning by generation
- Age binning by quartile

Key Feature from EDA : "Gen_Age"



Age distribution of interested customers seems equally spread to all age range,

age class segmentation could be beneficial for our analysis



Gen_Age

- 'Gen Z 20-27',
- 'Millennials 28-43',
- 'Gen X 44-59',
- 'Boomers 60-78',
- 'Post wars 79+'

Result

- **Clearer informative insights**
- **Better model target capturing**

Feature Selection :

EDA insights comparing with Pearson Correlation

Interesting Features

- **Gen_Age, Previously_Insured, Vehicle_Damage,**
- Some classes from **Region_Code** and **Sales Channel**

have some clear trends which captured potential customers in their ways :

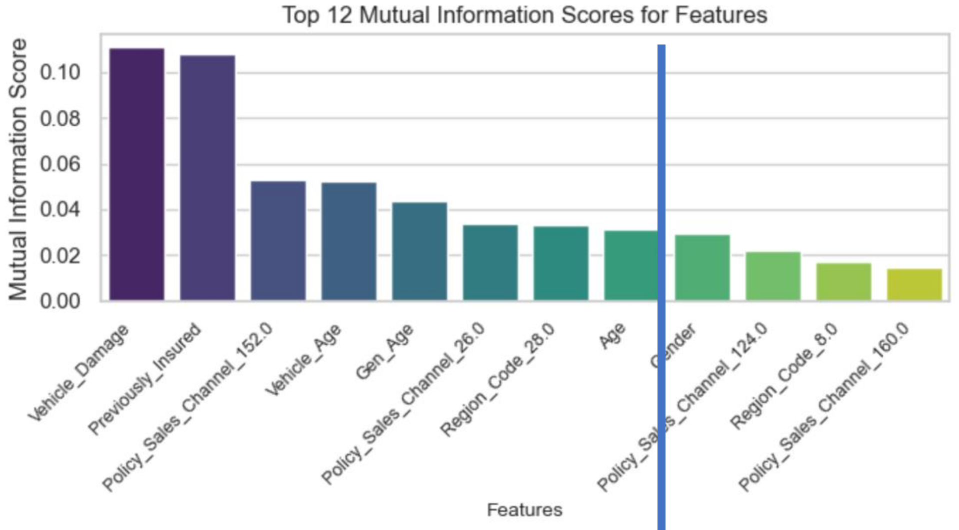
Some features were dropped

- **Vintage** and **Annual_Premium**
- Some classes from **Region_Code** and **Policy_Sales_Channels :**

as them seems very low correlation with other features including the target variable, didn't contribute any notable patterns of data.

Select KBest method - mutual_info_classif

for Classification, **mutual_info_classif** is adept at handling a mixed of categorical and numerical variables.



8 key factors for target prediction.

Vehicle_Damage, Previously_Insured, Age, Region_Code_28 Policy_Sales_Channel_152, Vehicle_Age and Gen_Age

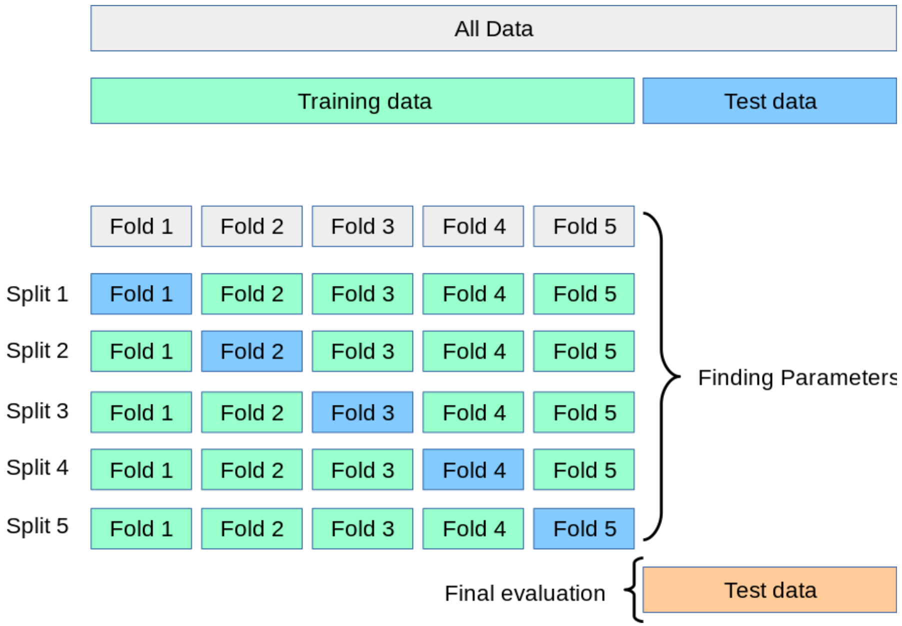
End-to-End Model Pipeline

Data Splitting : Stratified K-fold Cross Validation

Splitting into train (80%) and test (20%) while stratifying by 'Response'

Training set

> - Tuning using Stratified KFold cross-validation to split data while maintaining class balance.
> - build a **pipeline to avoid data leakage** by preprocessing separately within each training fold.



Overall pipeline process in each training split :

1) Adding engineered features
2) Feature Selection with **mutual_info_classif**
3) Oversampling technique (handle imbalance issue)
4) Finding the right hyperparameters with **Randomized Search CV**
5) Classifier Model e.g. Decision Tree, Random Forest and CatBoost
6) Evaluate with **F1 Score**

>> We considered synthetic oversampling techniques comparing

SMOTE :
- generates synthetic instances equally for all minority class
- May not adapt to complex decision boundary potentially leading to overfitting.

ADASYN (adaptive version of SMOTE based on density):
- more synthetic samples in regions where the minority class are sparse and harder to learn

Steps on Model Optimization & Evaluation Metric

Algorithms and Pipelines

- **single model**: Decision Tree
- **ensemble method**: Random Forest
- **boosting technique**: CatBoost for complex patterns.

Comparing pipelines with Stratified KFold & RandomizedSearch Hyperparameters Tuning.

1. **Baseline Features**
   - Uses original features as a benchmark for comparison.
2. **Baseline + Feature Engineering**
   - feature selection using mutual_info_classif.
   - Improves pattern recognition and model generalization.
3. **Baseline + Feature Engineering + ADASYN**
   - Enhances minority class learning with oversampling

Evaluation Metric : F1 score

Why? F1 score

✓ **Better Business Impact**

ensures performance both precision and recall

> A high **Recall** indicates we cover nearly all of the target class customer

> A high **Precision** indicates our identified target customers are highly accurate.

✓ **Reliable for Imbalanced Data**

> The F1 score reflects performance for predicting both classes

> Metric like Accuracy or ROC AUC can be misleading, as the model may predict the majority class well but fail on the minority class.

Final Decision : Best Model

After comparing, **CatBoost model with all pipeline process** is indeed performing better than the other models with F1 score = 0.8191

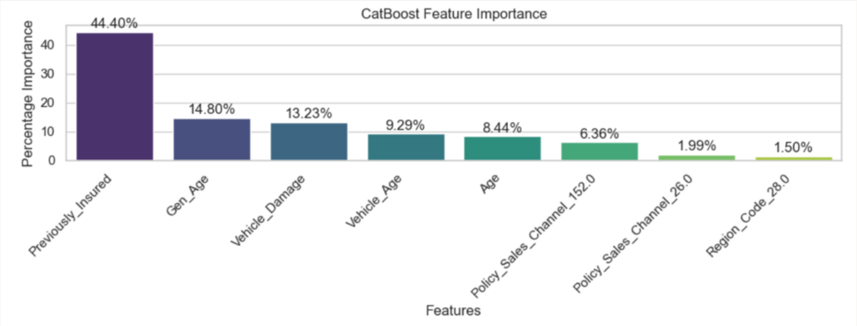
Baseline + Feature Engineering + ADASYN ✅

```
Decision Tree
          ROC AUC: 0.8426 – F1 Score: 0.4119 – PR AUC: 0.3270 – Recall: 0.9501
Random Forest
          ROC AUC: 0.8418 – F1 Score: 0.4126 – PR AUC: 0.3253 – Recall: 0.9485
CatBoost Average Score
          ROC AUC: 0.8500 – F1 Score: 0.8191 – PR AUC: 0.7899 – Recall: 0.8588
```

Implementation and Recommendations for raising customer interest

Final Model Feature Importance

Model explainability on understanding key drivers for making predictions.



1. **Previously_Insured**	**44.40 %**
2. **Gen_Age**	**14.80 %**
3. **Vehicle_Damage**	**13.23 %**
4. **Vehicle_Age**	**9.29 %**
5. **Age**	**8.44 %**
6. **Policy_Sales_Channel 152**	**6.36 %**
7. **Policy_Sales_Channel 26**	**1.99 %**
8. **Region_Code 28**	**1.50 %**

Lead Scoring System

	id	Interested	Response
0	381110	0.002237	0
1	381111	0.630170	1
2	381112	0.668322	1
3	381113	0.077095	0
4	381114	0.002237	0
...	...	...	...
127032	508142	0.000354	0
127033	508143	0.599169	1
127034	508144	0.002225	0
127035	508145	0.002545	0
127036	508146	0.001442	0

127037 rows × 3 columns

- **Predict Probability of Cross-sell Interest** with the final model.
- **Rank leads by confidence** (higher probability = higher priority).
- **Segment into deciles** (Top 10% = Decile 1, highest priority).
- **Focus cross sell efforts on top deciles** for marketing campaign & resource allocated
- **Track and Adjust** continuously refine. lead scoring based on conversion rates.

Implementing recommendations for raising customer interest :

Target Customer Profile

- Plan marketing campaign towards Millennials and Gen X,
- Develop key messages that appeal to individuals who have never had vehicle insurance before.

Potential Channels and Areas for Engagement:

Policy Sales Channels 26 and 124:

- highest numbers of interested customers
- high concentration of Millennials and Gen X

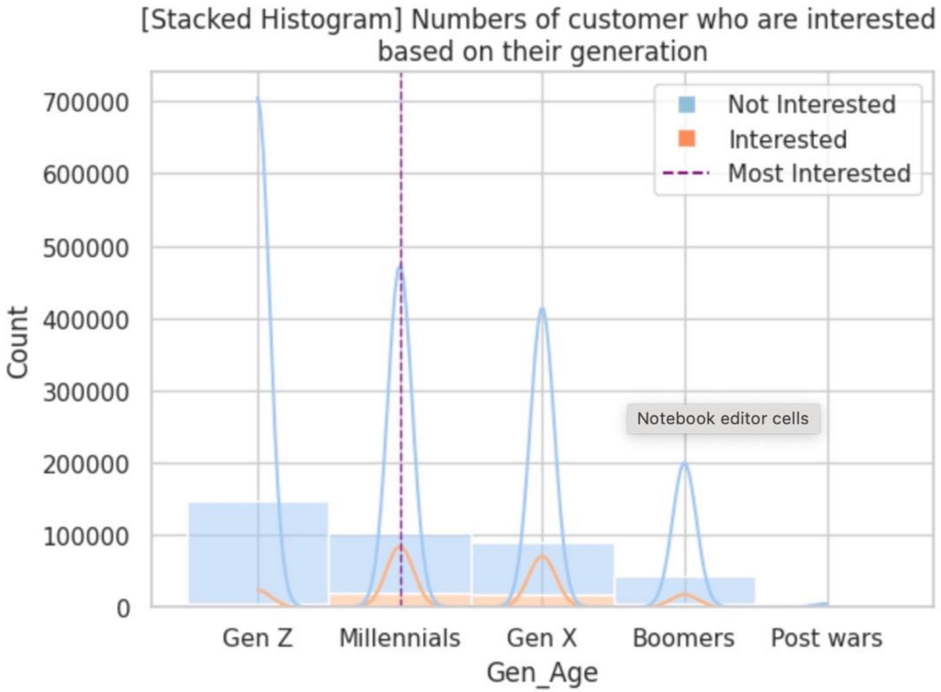
Potential Area : Region Code 28

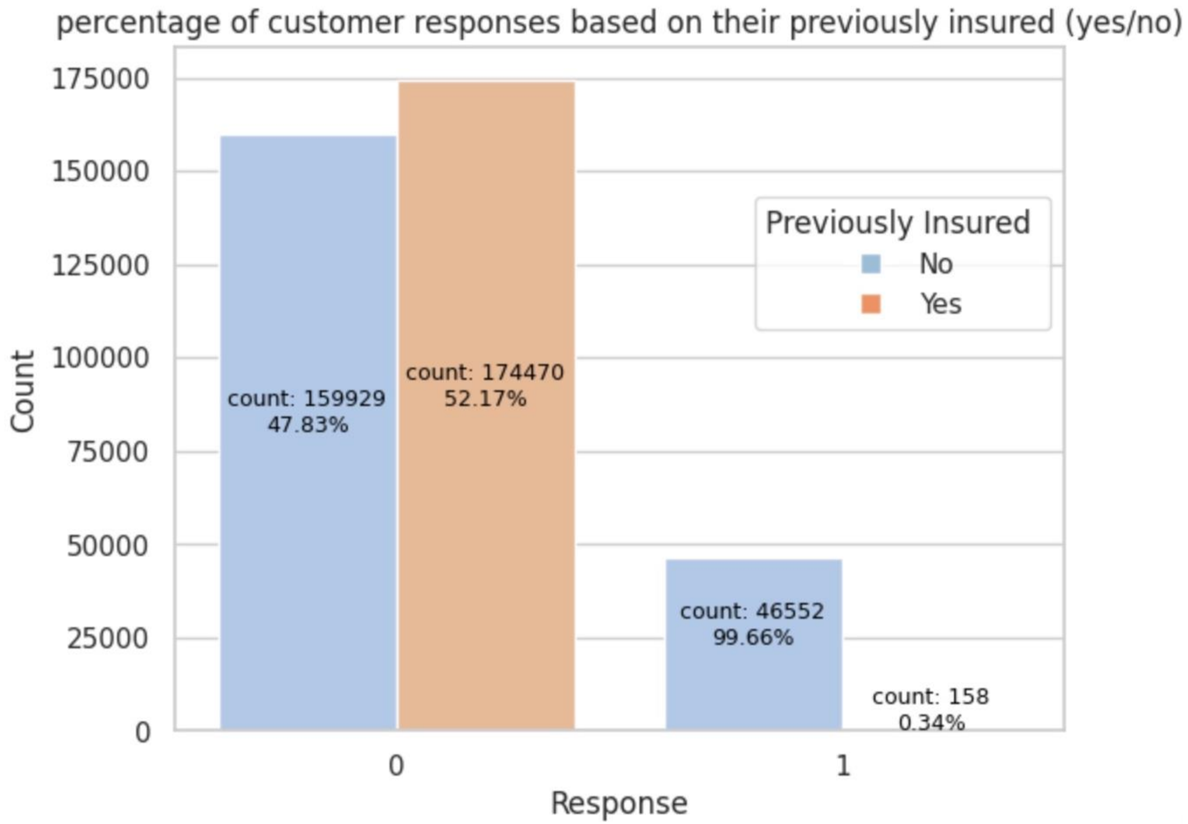
- Focusing efforts on Region Code 28, due to a notable amounts of interested customer.

Q&A

Appendix

EDA graphs & Visualizations



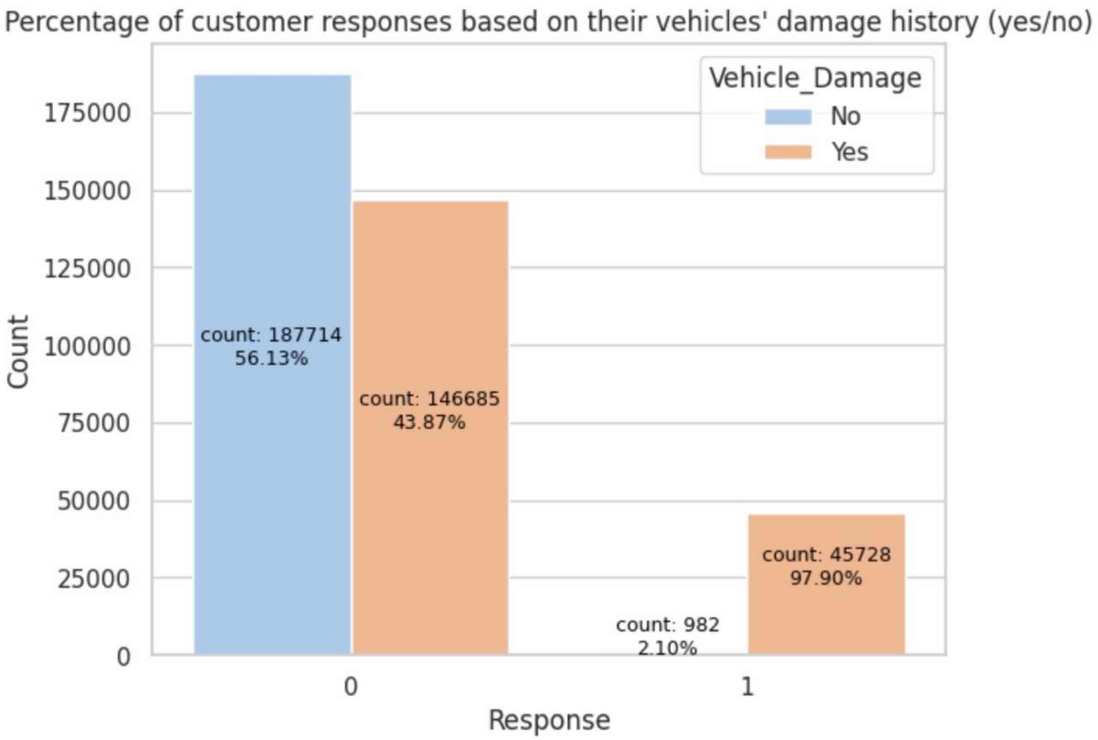


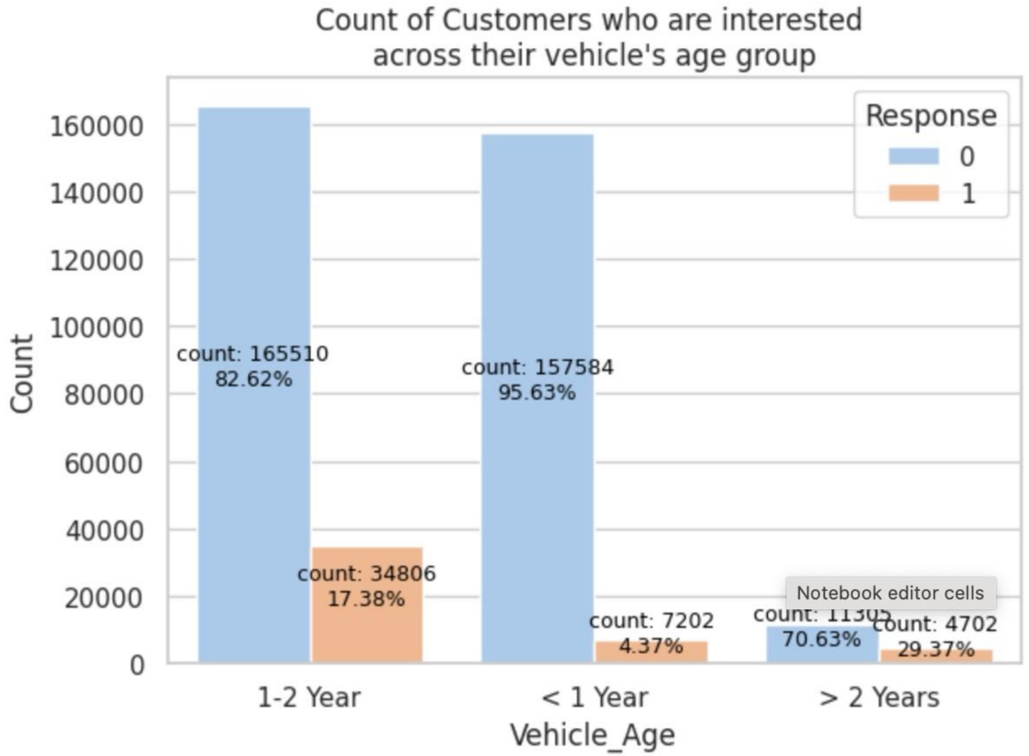
While most of existing customers are Gen Z but they show less of interest on cross-selling.

Millennials and Gen X are majority of interested customers .

Previously Insured
99% of customers interest in cross selling do not have existing coverage for their vehicles.

EDA graphs & Visualizations





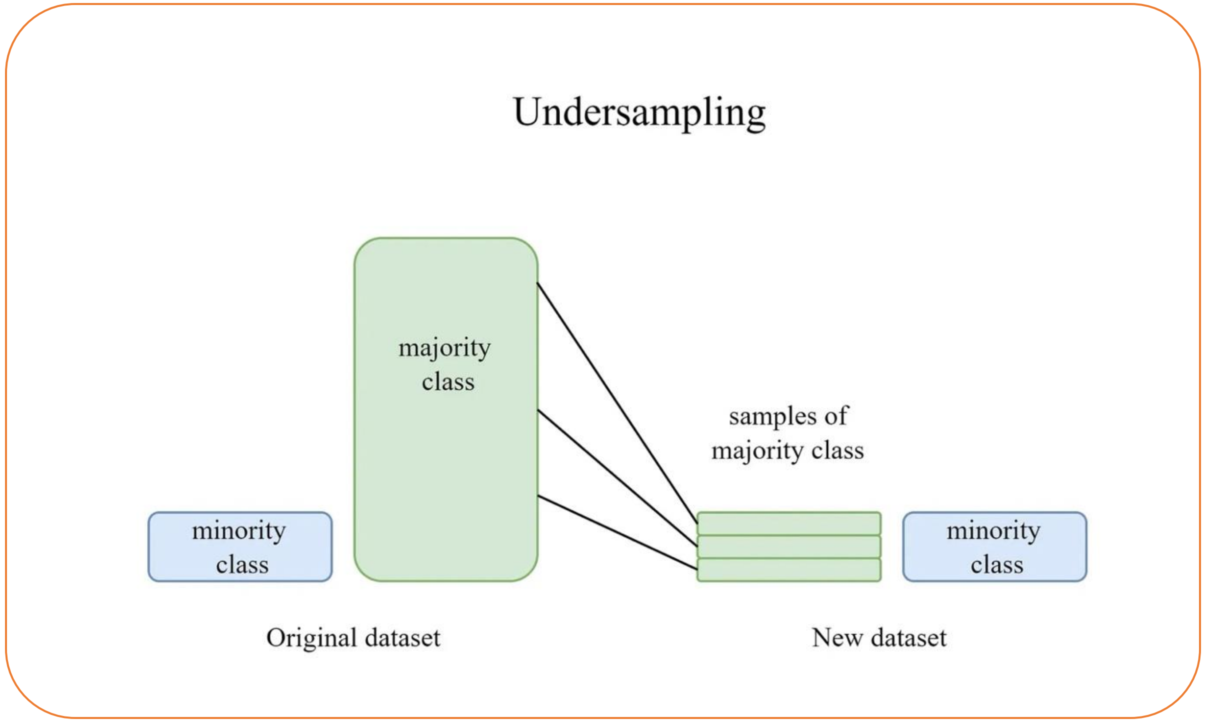
Vehicle Damage :

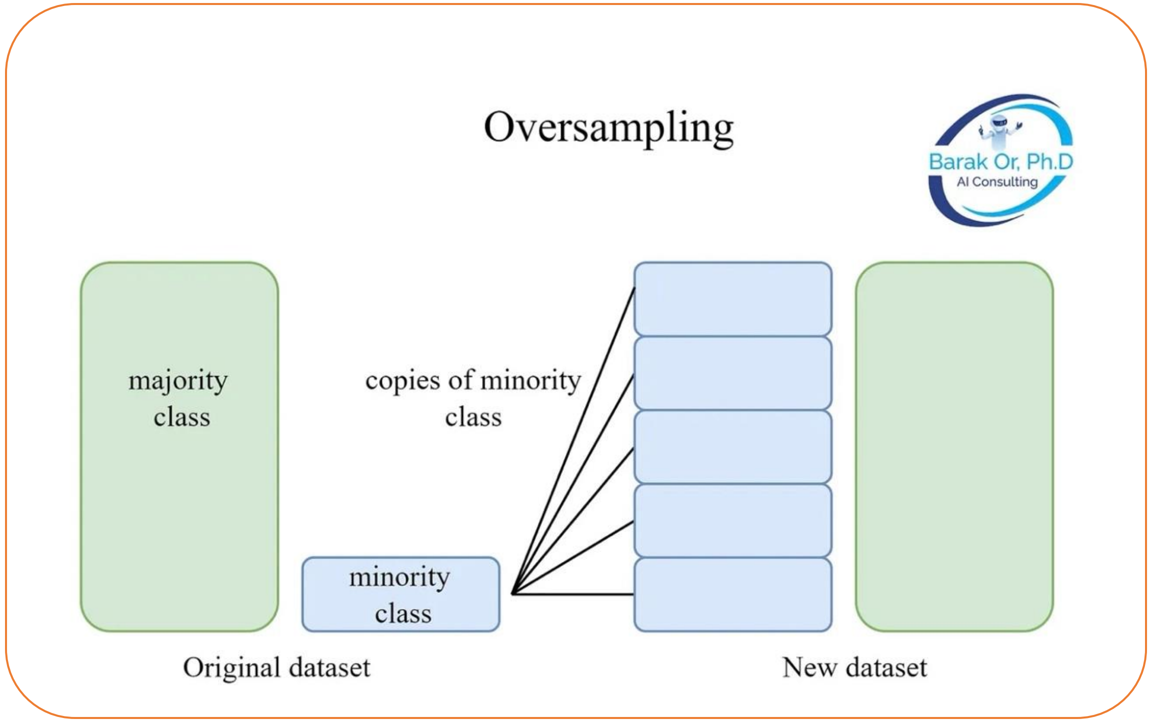
97.90% Customer who are interested in cross selling experienced damage to their cars

Higher vehicle age

greater percentage of customers expressing interest in vehicle insurance.

Undersampling vs Oversampling

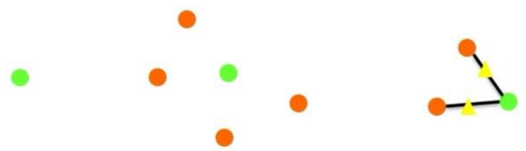




SMOTE vs ADASYN

SMOTE:

The full form of SMOTE, Synthetic **M**inority **O**versampling **T**echnique. Here Synthetic observations are generated from the Minority class



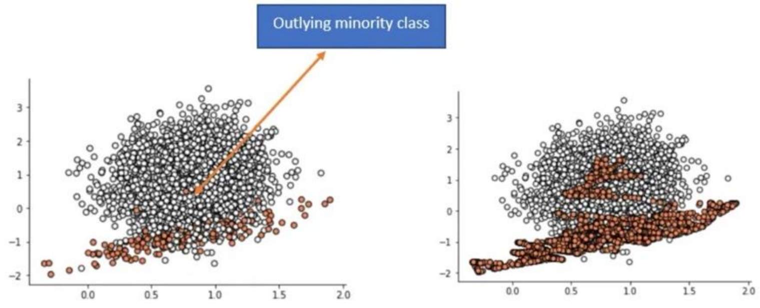
Step 1:
Identify a point from the minority class

Step 2:
Identify it's nearest neighbors. (Here 4 neighbors are taken. All minority neighbors)

Step 3:
Assume we want to add 2 times more synthetic observations
Select 2 nearest neighbor randomly
Draw a line between the point and the selected neighbors
Create a synthetic observation along the line

SMOTE, Synthetic Minority Observation Generation Process (Source: Author)

An issue with SMOTE:



Left Side: Original Data Right Side: Data after SMOTE is Applied (Image Source: Author)

If there are observations in the minority class which are outlying and appears in the majority class, it causes a problem for SMOTE, by creating a line bridge with the majority class.

ADASYN:

ADASYN is a more generic framework, for each of the minority observations it first finds the impurity of the neighborhood, by taking the ratio of majority observations in the neighborhood and k.

Minority Class	Minority Neighbours	Majority Neighbours	Impurity Ratio
Obs 1	3	2	.6
Obs 2	4	1	.4
Obs 3	1	4	.8
Obs 4	5	0	0

ADASYN Impurity Ratio

Now, first of all, this impurity ratio is converted into a probability distribution by making the sum as 1. Then higher the ratio more synthetic points are generated for that particular point. **Hence the number of synthetic observations to be created for Obs 3 is going to be double that of Obs 2.** So it's not so extreme as Borderline SMOTE and the boundary between the noise point, border point, and regular minority points are much softer. (Not a hard boundary). Thus the name adaptive.

To handle imbalance issue and improve F1 score, we considered on synthetic oversampling techniques comparing

While SMOTE :

- generates synthetic instances equally for all minority class
- May not adapt to complex decision boundary potentially leading to overfitting.

we chose ADASYN (adaptive version of SMOTE based on density):

- more synthetic samples in regions where the minority class are sparse and harder to learn

Source : https://towardsdatascience.com/class-imbalance-smote-borderline-smote-adasyn-6e36c78d804

Why F1 Score?

F1 ensures a balance between precision and recall on the positive class

while accuracy looks at correctly classified observations both positive and negative. That makes a big difference especially for the imbalanced problems,

> A high **Recall** indicates we cover nearly all of the target class customer

- A high **Precision** indicates our identified target customers are highly accurate.

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

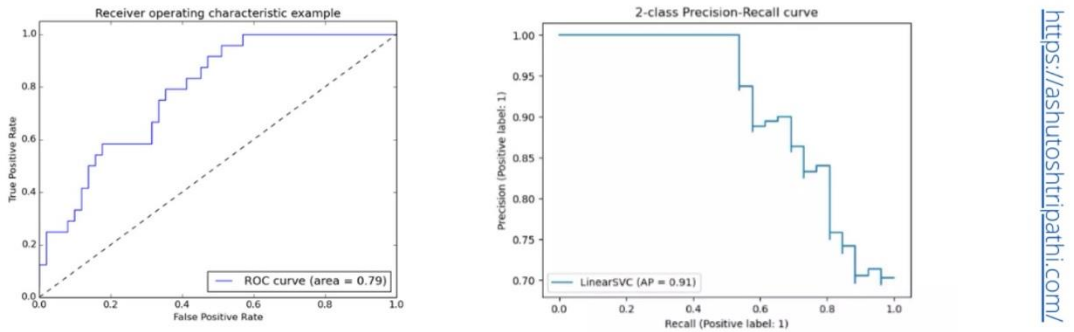
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Image by author

Precision-Recall Curve VS ROC-AUC Curve



Both Precision-Recall Curve and ROC-AUC curve are used:

- To explain model goodness of fit
- To identify the correct threshold to map probabilities value to the actual classes 0/1

When to use which one:

- Precision Recall curve is used when there is imbalance class distribution.
- ROC-AUC curve is used when there is balanced class distribution in data.

Precision Recall Curve

Precision and Recall Aspects in Target Prediction

False Positives (Contacted Uninterested)

- **Customer annoyance:** Unwanted contact might create frustration and damage brand perception.
- **Operational cost:** Resources spent on outreach to uninterested individuals.
- **Data quality concerns:** Can indicate issues with identifying characteristics of interested customers.

vs

False Negatives (Missed Leads)

- **Lost opportunity:** The insurer misses out on potentially valuable customers who might convert and bring in future premiums.
- **Marketing inefficiency:** Resources spent on campaigns that don't reach the right audience.
- **Competitor advantage**: Competitors might capture these missed leads, hurting market share.

Source : **"Modeling Customer Lifetime Value in Insurance" (2015) by David C. Edelman and Peter A. Kropp:**