# Predicting Health Insurance Policyholders Interest in Vehicle Insurance

By Brandon Chuck, Yiming Huang, Yujia Zhang

# **Project Problem & Goals:**

#### Problem:

★ Company wants to determine if current policyholders are interested in purchasing vehicle insurance.

#### Goals:

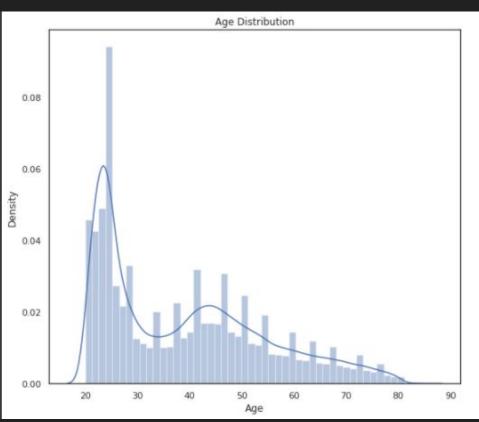
- ★ Test various classification models and find the best for predicting if current health insurance policyholders will buy vehicle insurance.
- $\star$  Predict the vehicle insurance Response based on the most impactful features.
- ★ Identify best predictors for a positive Response of 1.

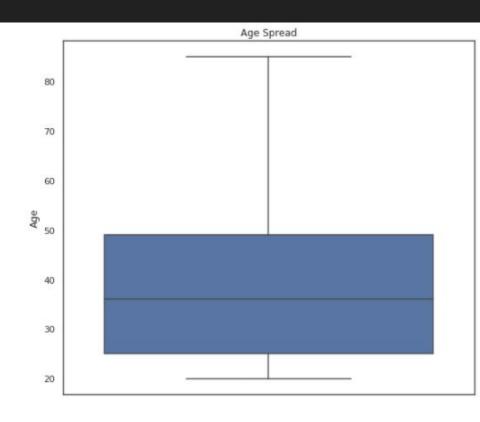
# **Browse Dataset: Existing Policyholder Features**

3	Cont	tinuous Features:
		Annual_Premium
		Vintage
		Age
3	Nom	inal Features:
		Vehicle_Age (<1 yr, 1-2 yrs, >2 yrs)
		Gender (Male or Female)
		Previously_Insured (0 or 1)
		Vehicle_Damage (Y or N)
		Driving_License (Y or N)
		Policy_Sales_Channel
		Region_Code
<b>1</b>	Pred	licting For:
		Response (0 or 1)
		0 - Not Interested
		□ 1 - Interested

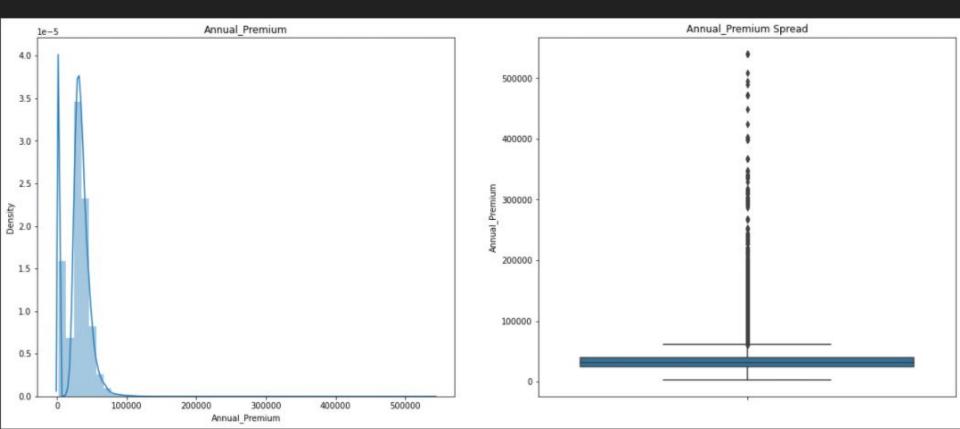
#### **Feature Clarifications: Region\_Code:** Unique code for the region of the customer Policy\_Sales\_Channel: Anonymised code for the channel of outreaching to the customer ie. Different Agents, By-Mail, By-Phone, In-Person, etc. Vintage: Number of days customer has been associated with the company

# **EDA Pt.1 - Age distribution**





# **EDA Pt2 - Annual Premium**



# **Correlation Matrix**

#### **Highly Correlated Features:**

- **□** Response:
  - **□** Vehicle\_Damage\_Yes
  - ☐ Vehicle\_Damage\_No
  - ☐ Previously\_Insured
- Previously\_Insured:
  - □ Vehicle\_Damage\_Yes
  - □ Vehicle\_Damage\_No

Previously_Insured	1	0.083	-0.083	-0.28	0.36	-0.19	0.82	-0.82	-0.051	-0.08	-0.25	0.34	-0.26	0.014	-0.024	0.0048	0.22	0.0017	-0.29
Gender_Female	0.083	1		-0.15	0.17	-0.043	0.092	-0.092	-0.017	-0.008	-0.14	0.16	-0.15	0.018	-0.00068	3-0.0035	0.11	0.0021	-0.045
Gender_Male	-0.083	-1		0.15	-0.17	0.043	-0.092	0.092	0.017	0.008	0.14	-0.16	0.15	-0.018	0.00068	0.0035	-0.11	-0.0021	0.045
Vehicle_Age_1-2 Year	-0.28	-0.15	0.15	1	-0.92	-0.22	-0.29	0.29	0.14	0.18				-0.037	0.038	-0.0024	-0.51	-0.0018	0.14
Vehicle_Age_< 1 Year	0.36	0.17	-0.17	-0.92		-0.18	0.37	-0.37	-0.17	-0.17				0.039	-0.044	-0.023		0.0015	-0.18
Vehicle_Age_> 2 Years	-0.19	-0.043	0.043	-0.22	-0.18	1	-0.21	0.21	0.07	-0.043	0.15	-0.17	0.22	-0.0054	0.015	0.062	-0.15	0.001	0.093
Vehicle_Damage_No	0.82	0.092	-0.092	-0.29	0.37	-0.21		-1	-0.061	-0.074	-0.26	0.35	-0.27	0.016	-0.027	-0.0095	0.23	0.0011	-0.3
Vehicle_Damage_Yes	-0.82	-0.092	0.092	0.29	-0.37	0.21	-1	1	0.061	0.074	0.26	-0.35	0.27	-0.016	0.027	0.0095	-0.23	-0.0011	0.3
age_range_Elderly	-0.051	-0.017	0.017	0.14	-0.17	0.07	-0.061	0.061	1	-0.15	-0.29	-0.31	0.41	-0.067	0.018	0.025	-0.14	0.0013	-0.02
age_range_Middle	-0.08	-0.008	0.008	0.18	-0.17	-0.043	-0.074	0.074	-0.15	1	-0.28	-0.3	-0.09	0.016	0.0069	-0.064	0.02	-0.0014	0.086
age_range_Older	-0.25	-0.14	0.14			0.15	-0.26	0.26	-0.29	-0.28			0.54	-0.00099	0.027	0.039	-0.45	-0.00097	0.13
age_range_Young	0.34	0.16	-0.16	-0.78		-0.17	0.35	-0.35	-0.31	-0.3				0.036	-0.044	-0.011	0.53	0.001	-0.17
Age	-0.26	-0.15	0.15	0.69	-0.79	0.22	-0.27	0.27	0.41	-0.09	0.54	-0.75	1	-0.08	0.042	0.067	-0.58	-1e-05	0.095
Driving_License	0.014	0.018	-0.018	-0.037	0.039	-0.0054	0.016	-0.016	-0.067	0.016	-0.00099	0.036	-0.08	1	-0.0002	-0.012	0.043	-0.00085	0.0083
Region_Code	-0.024	-0.00068	20.00068	0.038	-0.044	0.015	-0.027	0.027	0.018	0.0069	0.027	-0.044	0.042	-0.0002		-0.012	-0.042	-0.0016	0.0086
Annual_Premium	0.0048	-0.0035	0.0035	-0.0024	-0.023	0.062	-0.0095	0.0095	0.025	-0.064	0.039	-0.011	0.067	-0.012	-0.012	1	-0.11	-0.00085	0.019
Policy_Sales_Channel	0.22	0.11	-0.11	-0.51		-0.15	0.23	-0.23	-0.14	0.02	-0.45	0.53		0.043	-0.042	-0.11	1	-0.00081	-0.12
Vintage	0.0017	0.0021	-0.0021	-0.0018	0.0015	0.001	0.0011	-0.0011	0.0013	-0.0014	-0.00097	0.001	-1e-05	-0.00085	5-0.0016	-0.00085	0.0008	1	0.00087
Response	-0.29	-0.045	0.045	0.14	-0.18	0.093	-0.3	0.3	-0.02	0.086	0.13	-0.17	0.095	0.0083	0.0086	0.019	-0.12	-0.00087	1
	eviously_Insured	Gender_Female	Gender_Male	cle_Age_1-2 Year	le Age < 1 Year	e_Age_> 2 Years	icle_Damage_No	cle_Damage_Yes	ge_range_Elderly	ge_range_Middle	age_range_Older	ge_range_Young	Age	Driving_License	Region_Code	Annual Premium	/_Sales_Channel	Vintage	Response

- 0.25

- 0.00

- -0.50

# Feature Engineering: Create New Features based on existing features

#### **New Features:**

1. 'insured\_with\_no\_damage'

df\_numeric['insured\_with\_no\_damage'] = df\_numeric['Previously\_Insured'] \*df\_numeric['Vehicle\_Damage\_No']

'not\_insured\_with\_damage'

 $df_numeric['not_insured_with_damage'] = df_numeric['Previously_Insured'].apply(lambda x: 1 if x == 0 else 0) * df_numeric['Vehicle_Damage_Yes']$ 

# **Preprocessing Steps:**

- 1. Combine each dataset (train.csv, test.csv, sample\_submission.csv) into single dataframe
- Create a new feature 'age\_range' w/ 4 categories: <u>Young (20-30)</u>, <u>Middle (30-40)</u>, <u>Older (40-65)</u>, and <u>Elderly (65-85</u>).
- 3. Drop feature 'id' that have unique values, 'Driving\_License', and 'Vintage' which seems to not affect the model.
- 4. One-hot encode categorical features: 'Gender', 'Previously\_Insured', 'Vehicle\_Age', 'Vehicle\_Damage', 'age\_range'.
- 5. Apply RobustScaler() from Scikit-Learn in the pipeline models to Annual\_Premium to handle outliers.
  - \* Only apply StandardScaler() to 'Age' when training non-tree models.

#### Models:

- lue Baseline Used ightarrow Mode
  - Since we are dealing with a binary classification problem, we decided to compare our predictions with the mode.
- Pipeline Models
  - Logistic Regression + GridSearchCV
  - Decision Tree Classifier + GridSearchCV
  - Random Forest Classifier + GridSearchCV
  - KNN Classifier + GridSearchCV
  - XGBoost Classifier + GridSearchCV

### **Fitting Models with Raw Dataset**

Train.csv

test.csv

Baseline Accuracy= 0.91

Raw dataset

Create age\_range to young/middle/older/elder

Drop id, Driving\_IRaw dataset icense, vintage

Sample\_submission.csv

One hot encoding

'Gender', 'Previously\_Insured', 'Vehicle\_Age', 'Vehicle\_Damage', 'age\_range'.

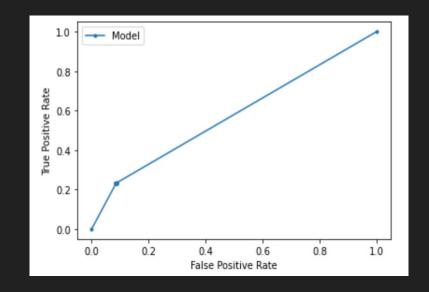
Create two new features:

Insured\_with\_no\_damage = Previously\_Insured \* Vehicle\_Damage\_No Not\_insured\_with\_damage = Previously\_Insured \* Vehicle\_Damage\_Yes

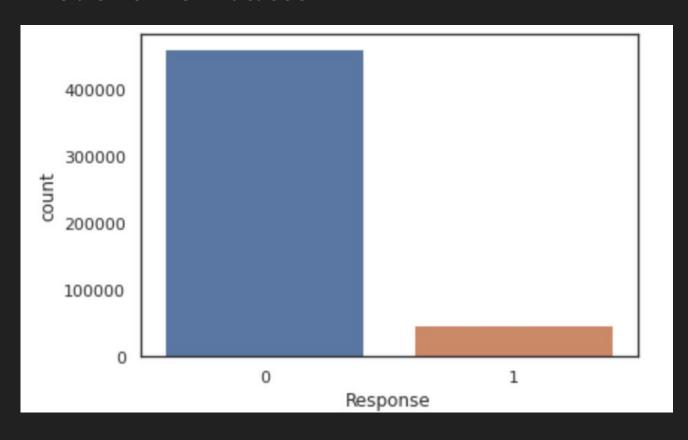
For non\_tree model: StandardScaler() -> Age RobustScaler() -> Annual Premium For tree model:

RobustScaler() -> Annual Premium

Models	F1_Train	F1_Val	ROC_train	ROC_val		
Decision Tree_CV	1	0.85	1	0.573		
Random Forest_CV	1	0.9	1	0.829		
∠ogistic Regression_CV	0.91	0.91	0.827	0.824		
KNN_CV	1	0.88	1	0.574		



# **Problems with Dataset:**



■ Number of
Responses were
highly unbalanced
which affected our
model's predicting
capabilities.

■ Baseline mode accuracy rate is 0.9

# **Balancing the Data**

Response

dtype: int64

467827 461436

To remedy this, we applied the ADASYN model which was designed to balance datasets being used for classification.

```
[] # Use ADASYN to balance the dataset
    from imblearn.over_sampling import ADASYN

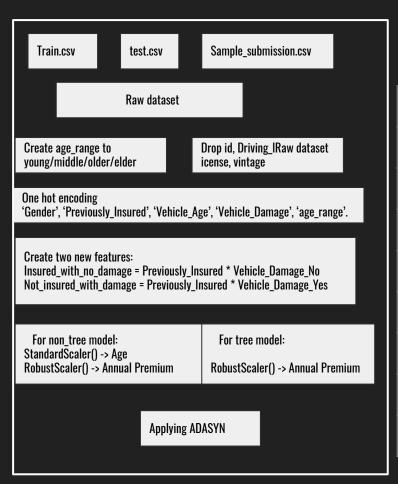
# create ADASYN model
adasyn = ADASYN()

# balance data
    df_X_adasyn, df_y_adasyn = adasyn.fit_resample(df_X, df_y)

[] new_df_y = pd.DataFrame(df_y_adasyn, columns=['Response'])

[] # Confirming that we have balanced Responses after ADASYN
    new_df_y_adasyn.value_counts()
```

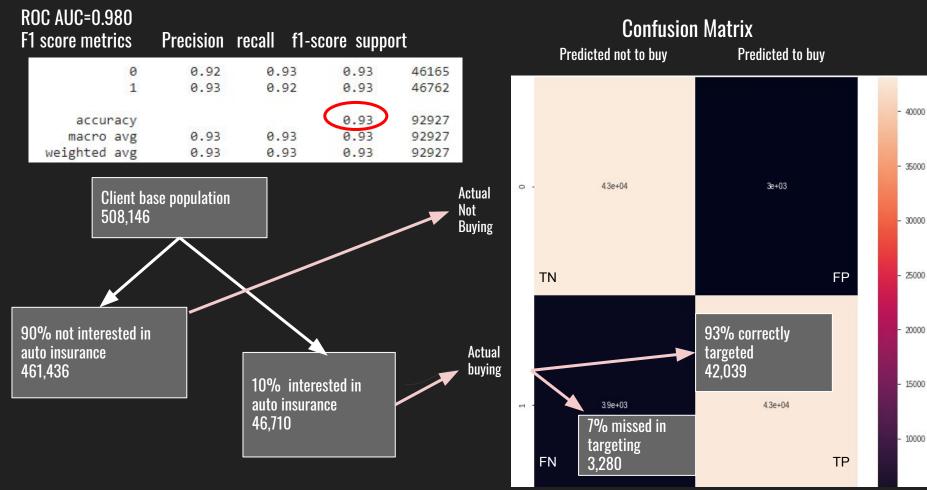
## Improved Results: After Applying ADASYN





Models	F1_Train	F1_Val	F1_Test	ROC_train	ROC_val	ROC_test
Decision Tree	0.93	0.91	0.91	0.988	0.961	0.959
Decision Tree_CV	0.98	0.92	0.92	0.999	0.938	0.935
Random Forest	0.99	0.93	0.93	0.999	0.98	0.98
Random Forest_CV	0.99	0.93	0.93	0.999	0.981	0.98
Logistic Regression_CV	0.78	0.78	0.78	0.829	0.83	0.828
KNN_CV	0.99	0.88	0.88	0.998	0.938	0.937
XGBoost	0.91	0.91	0.91	0.976	0.976	0.976
XGBoost_CV	0.92	0.92	0.92	0.98	0.979	0.979

# **Conclusion: Best Model -- Random Forest On Test Set**



# **Most Impactful Features:**

Feature	Importance Score	Feature In-Detail
not_insured_with_damage	0.12845653	Previously insured but car damage
Region_Code	0.12077327	The region of the customer
Previously_Insured	0.10421228	Previously insured with auto insurance
age_range_Elderly	0.08769957	Age 65-85
Vehicle_Damage_Yes',	0.08241084	Car previously damaged
Annual_Premium	0.07346153	The amount customer needs to pay as premium in the year
Vehicle_Age_1-2 Year	0.05321889	Car 1-2 year
Policy_Sales_Chanel	0.04980615	Different Agents, Over Mail, Over Phone, In Person,

# Thank you!