Insurance Cross Sale Prediction

Brown University

Dongjun Shin

(https://github.com/josh7197/midterm)

11th Dec, 2021

What matters?

Q : How can an insurance company sell cost effectively an additional insurance plans(=car insurance) to current customers?

→ *If the company have accurately distinguish whether the current insurance holders are interested in the care insurance????*

Is it important?

- Target marketing
- Cost saving
- Efficient customer service

How to predict?

- Classification problem

- Target variable(=Respond) is a Dummy variable
- If someone is interested, the dummy is 1, if not, 0

Where I get the data

- Kaggle dataset

EDA - General Information

Variable	Classification	Explanation
Gender	Categorical	Customer Gender(M:1, F:0)
Age	Numerical	Customer Age
Driving_License	Categorical	Having a DL has 1 or 0
Region_Code	Categorical	Customer region code
Previously_Insured	Categorical	Already having a car insurance has 1 or 0
Vehicle_Age	Categorical	Vehicle Age
Vehicle_Damage	Categorical	Damaged car has 1 or 0
Annual_Premium	Numerical	The annual insurance premium
Policy_Sales_Channel	Categorical	Contact channel Code
Vintage	Numerical	Days when customers has been with the company
Response(Target)	Categorical	Being interested has 1 or 0

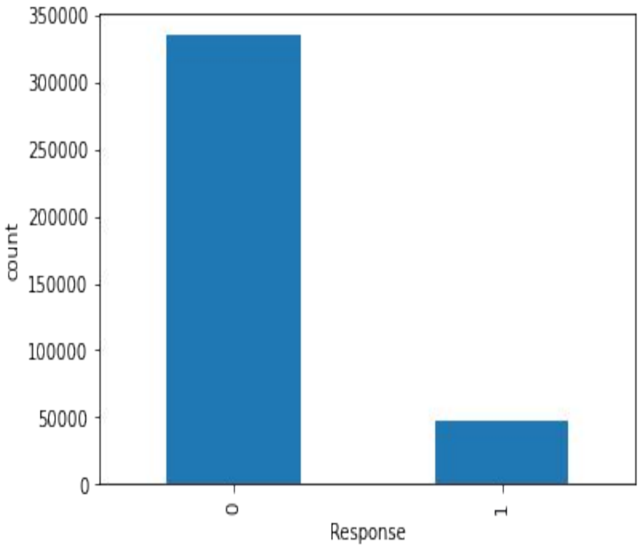
- **A Shape of Dataset**

- 12 columns
  -> **1 Target, 10 features**
  (ID was excluded)

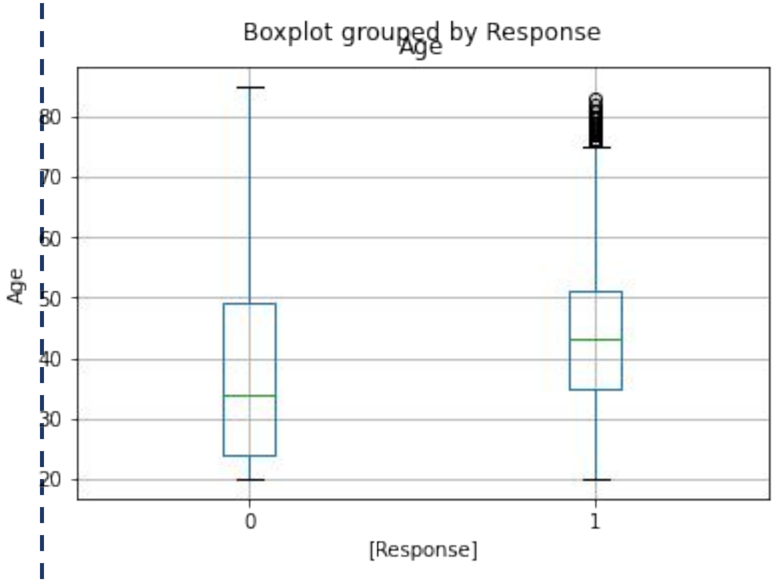
- 381,109 rows
  -> **No Missing Data**

EDA - Remarkable Findings

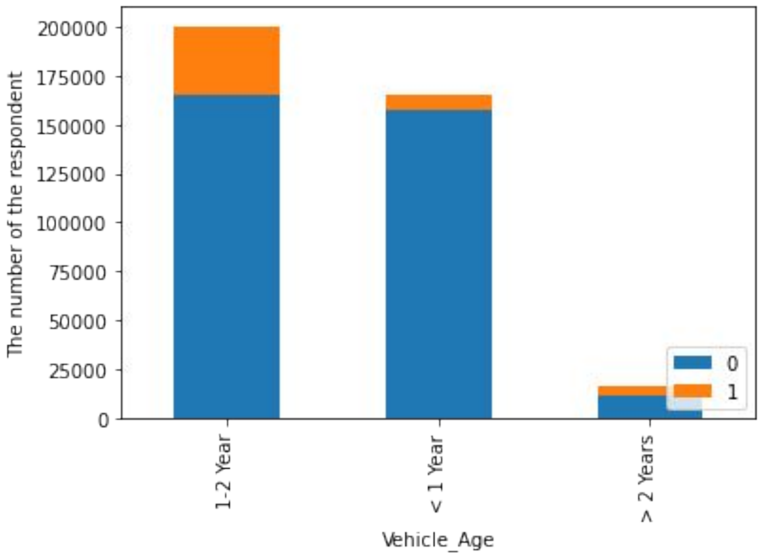
Target variable



Customer Age & Target



Vehicle age and Target



- Imbalance data(87.74%)
- Cost saving
- Efficient customer service

- Why 1~2 vehicle owner are interested in?

- 40~50 more interest

Cross Validation

1. Reducing Data Points

- Too large data points : **380K**
  - My laptop does not work with full data
- **Random Stratification Sampling**
  - Same Target variable ratio : 88%
  - 5 Sampling with different random states

2. Splitting

- **General Splitting**
  - Why ? IID (Unique ID)
- **Test data ratio** : 20%

3. Kfold Setting

- **3 folds** for robust validation

4. Preprocessing

- **OneHotEncoder**
  : Seven **categorical** features
- **StandardEncoder**
  : Three **continuous** features

5. Parameter setting

- **Diverse hyper parameters per 4 different ML models**

6. Grid Search

- **Evaluation metric**
  : Accuracy Score
- **Best parameter combination**

7. Models

- **KNN Classification**
- **XGboost Classification**
- **Logistic Regression**
- **Random Forest**

8. Function

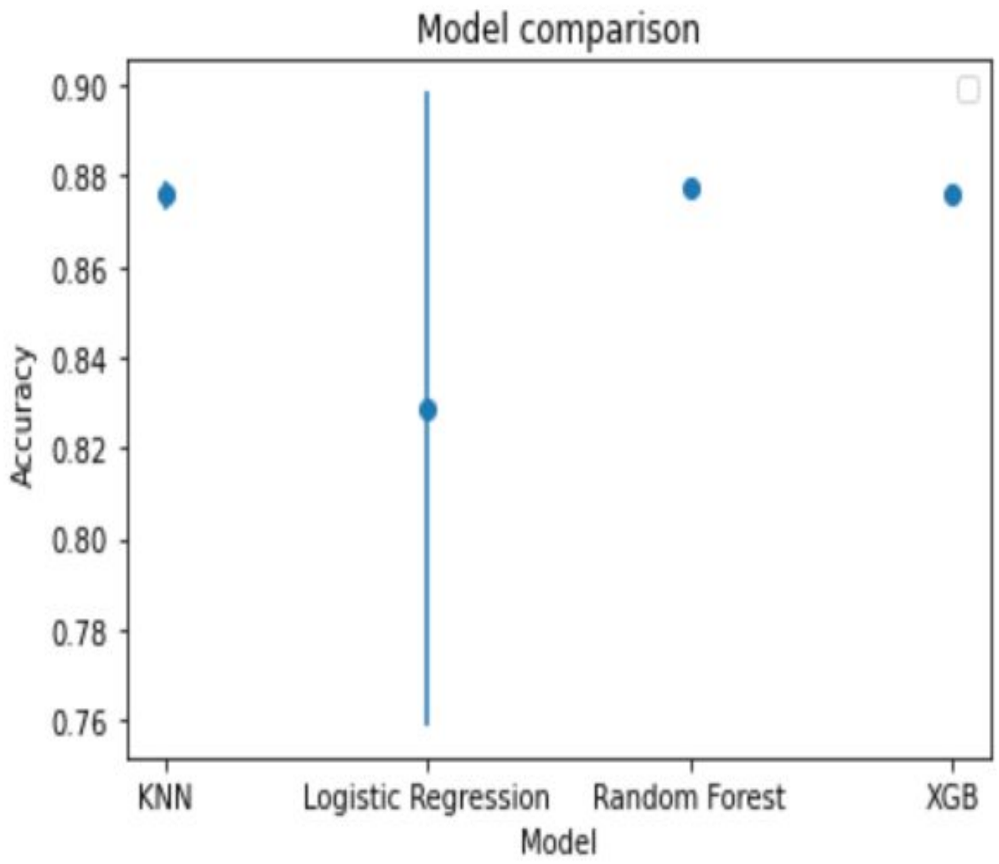
- **5 different Random states**
- Returning **Best model & score** per each random state
- **Yielding means and std** of Scores per each ML model

Cross Validation

ML Model	Parameters
KNN Classification	• **N_neighbors** : 3,5,7,20,30 • **Weights** : uniform, distance
XGboost Classification	• **Learning_rate** : 0.03, 0.05 • **model__max_depth** : 5,10,50,30,50
Logistic Regression	• **C** : 0.01, 0.1, 10, 50, 100 • fit_intercept : 0, penalty: ['l2'] , solver : lbfgs , max_iter : 10000000
Random Forest	• **Min_samples_split** : 16, 32, 64, 128 • **Max_depth** : 10,30,100,300

Result

Error bar



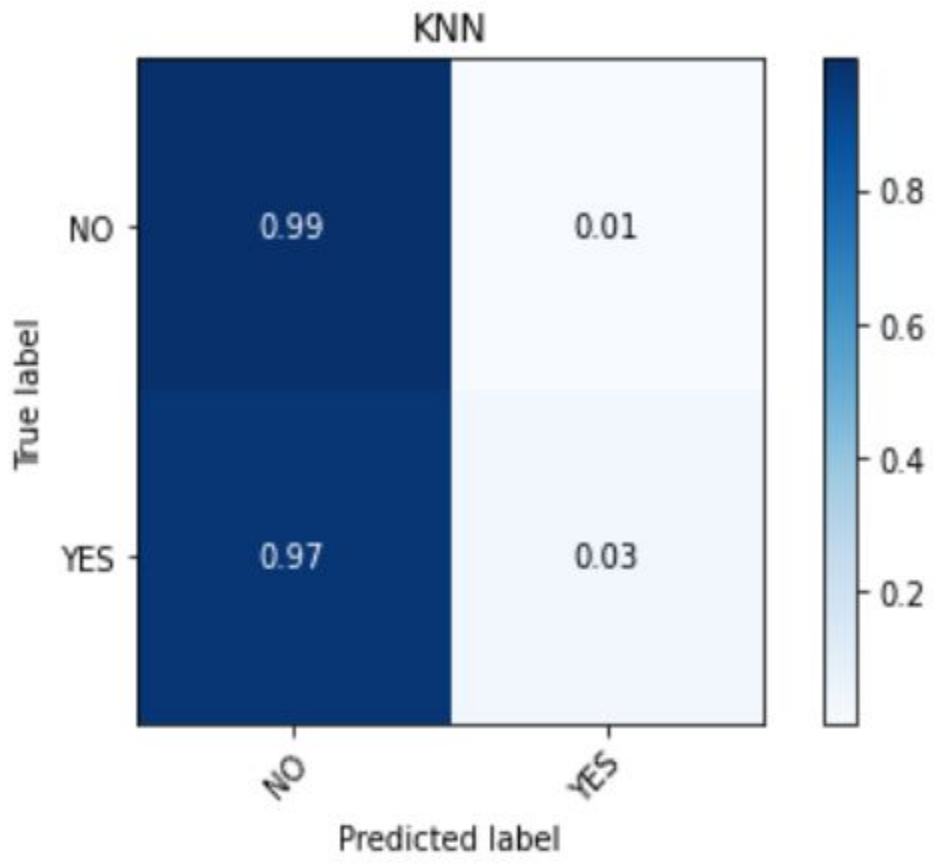
Model comparison

Model	Means	std	std from the baseline
KNN	87.56%	0.0034	-0.5246
Logistic Regression	82.87%	0.0698	-0.6987
Random Forest	87.76%	0.0010	0.1448
XGboost	87.63%	0.0017	-0.6518

Base line accuracy : 87.74%

Result

Confusion matrix



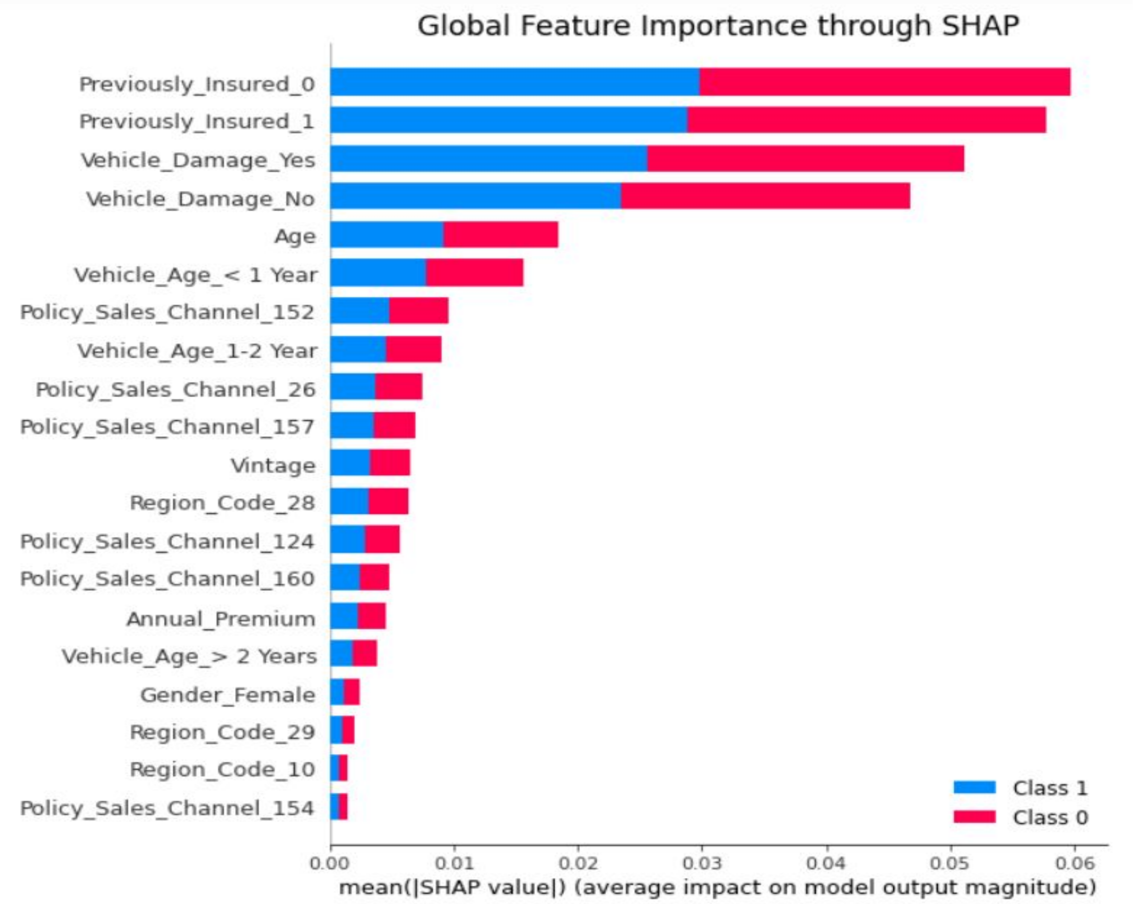
Implication

- **All four matrices are similar like the left**
  - Most are predicted to “Negative Response(=0)
- **Alternative Trials**
  - Starification splitting
  - Hyper-parameter re-setting

No Impact

Result

Global feature importance



Random forest result

Model comparison

- **Expected results**
  - **Age, vehicle damage,** and **previous insured** are commonly most important to the period
  - **Driver licence is less important**
- **Unexpected results**
  - **Gender** is **less critical** than I expected

Result

Local feature importance

- **Main Features that lower the possibility from baseline**
  - Previously_insurance_0
  - Vehicle_Damage_Yes
  - Previous-insurance_1

- **A feature that higher the possibility from baseline**
  - Vintage

Outlook

- **Weakness**
  - All four models' performances are almost the same as the base line
  - Almost 99% of responses are predicted to 0

- **Potential trials for performance improvement**
  - Interaction variables
    - For instance, an interaction term between age and previous insured.

  - New variables for largely dispersed categorical variables
    - Region codes and Sales channels have more than 50 categories.
    - But, most data points are densely populated in a small number of categories.