


Presented By: Amit Kukreja

Updated: 02-July-2022



MODEL METRICS

Purchase Prediction for

AllState Car Insurance online customers





Table of Contents

Table of Contents	2
1.0 Model Performance Metrics	3
2.0 Model Hyperparameter Settings	5
3.0 Model Features	6
3.1 Feature Nomenclature:	7

1.0 Model Performance Metrics

To predict the customer's final choice for each of the 7 product vectors, models were built using different classifiers based on the customer's quote 1 and 2 information. Each model's hyperparameters were tuned and thresholding was done for binary vectors B and E.

The models were then used to make predictions on the test set and based on the performance the best model was chosen for each vector. Table 5.1 summarizes the performance of the best models across each of the 7 vectors.

	recall_baseline_model	best_model	recall_best_model	improvement_over_baseline	recall_cust_changed_vector	recall_cust_maintained_vector
vector						
A	81.91	xg	84.23	2.32	21.11	98.17
B	83.96	lr	84.14	0.18	0.48	99.96
C	80.18	xg	82.98	2.80	27.02	96.81
D	85.09	rf	86.71	1.62	16.29	99.04
E	83.80	xg	84.72	0.92	13.07	98.58
F	81.04	rf	82.63	1.59	11.36	99.30
G	74.31	rf	75.98	1.67	11.46	98.28

Table 5.1

Key Highlights

- 1) Random Forest and XGBoost came up as the best classifiers, with each of these classifiers giving the best performance for 3 of the 7 vectors.
- 2) The models gave a performance improvement ranging from **0.18% to 2.8%** over the baseline model (final vector = quote2).
 - a) The average improvement was **1.59%**.
 - b) The best performance was for vector C at **2.8%**.
 - c) **Vector B** was the hardest to predict with only 0.18% improvement.

- 3) Another way to analyze the model performance is by looking at predictive prowess of various models for customers who changed their vectors (from quote 2 to the final purchase) and those who maintained their vector choices (classes).
- a) All models did a very good job in predicting **customers who didn't change** their vector choice. The recall of the models ranged from **96.81%** for vector C to **99.61%** for vector B. On average, the models predicted **98.54%** of such customers correctly.
 - b) For customers **that changed vector choice**, the task of prediction was harder. The models had to predict which customers would change vector choice and what the new choice would be. Here the model performance ranged from **0.48%** improvement for vector B over baseline model to **27.02%** for vector C. On average, the models predicted **14.72%%** of such customers correctly.

2.0 Model Hyperparameter Settings

VECTOR	BEST MODEL	HYPERPARAMETER SETTINGS
A	XGBOOST	n_estimators = 300 max_depth = 3 learning_rate = 0.3, colsample_bytree = 0.5
B	LOGISTIC REGRESSION	Best Threshold = 0.412
C	XGBOOST	n_estimators = 200 max_depth = 5 learning_rate = 0.25, colsample_bytree = 0.35
D	RANDOM FOREST	n_estimators=300, min_samples_leaf = 1 min_samples_split = 2
E	XGBOOST	n_estimators = 400 max_depth = 3 learning_rate = 0.25, colsample_bytree = 0.5 Best Threshold = 0.442
F	RANDOM FOREST	n_estimators=300, min_samples_leaf = 1 min_samples_split = 2
G	RANDOM FOREST	n_estimators=550, min_samples_leaf = 1 min_samples_split = 5

3.0 Model Features

Charts 1 through 7 show the top 10 features used for predicting a vector by the respective best model.

Chart 1

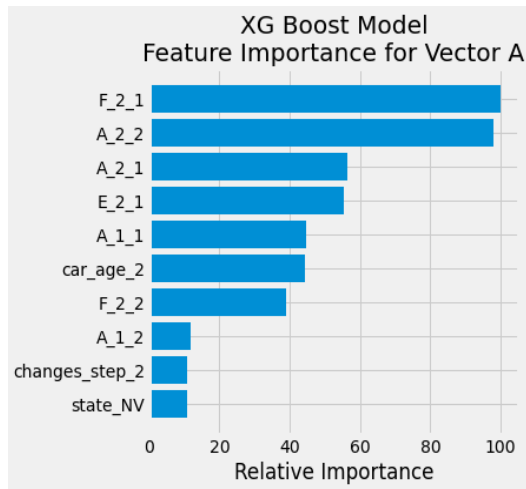


Chart 2

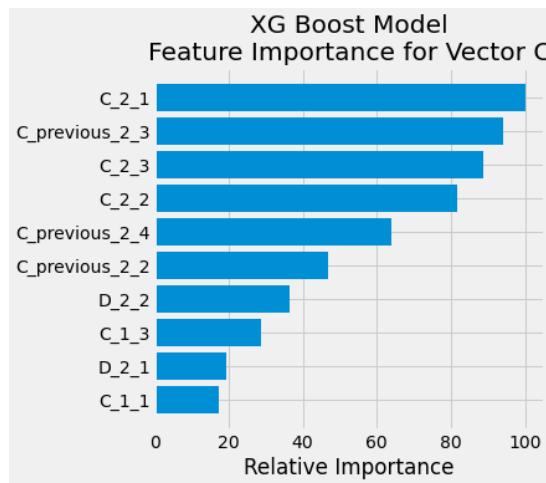


Chart 3

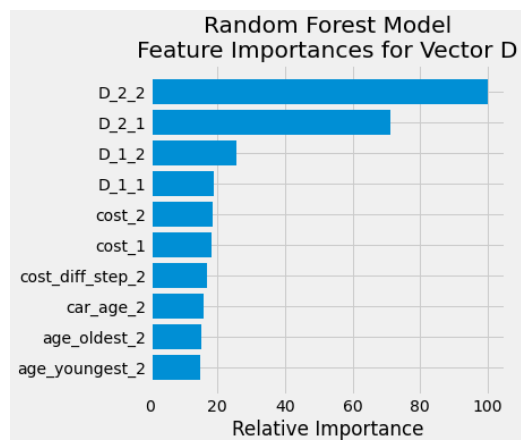


Chart 4

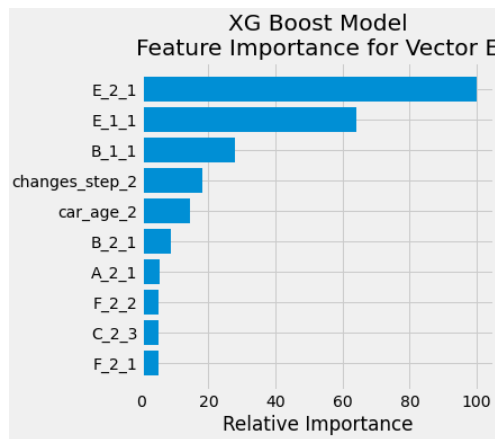


Chart 5

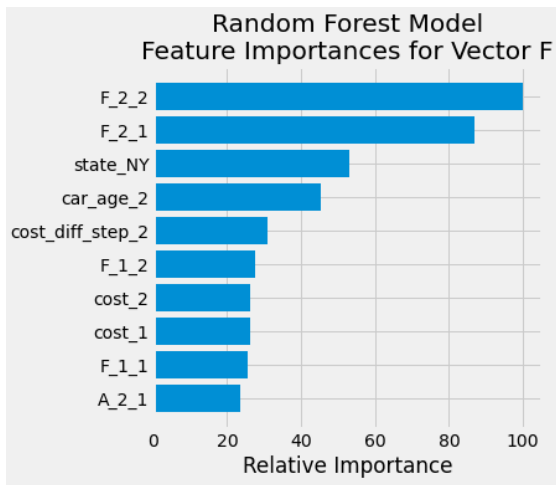


Chart 6

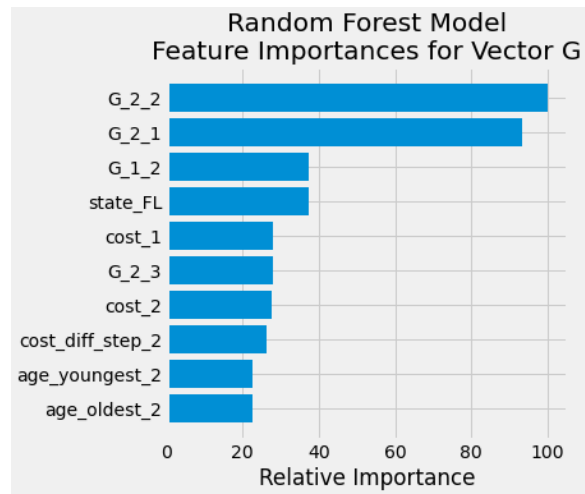


Chart 7

3.1 Feature Nomenclature:

'A': Customer's final choice for vector A i.e. the target we are predicting for vector A.

'cost_2': policy cost of quote 2 taken by the customer.

'A_2_2': One-hot encoded feature for vector A, at shopping pt 2, for class 2.