INSURANCE CROSS SELL PREDICTION

Roma Coffin
Brown University
Data 1030 Project
12/3/20
https://github.com/romacoffin/Health-Insurance-Cross-Sell-Prediction

RECAP-INTRODUCTION

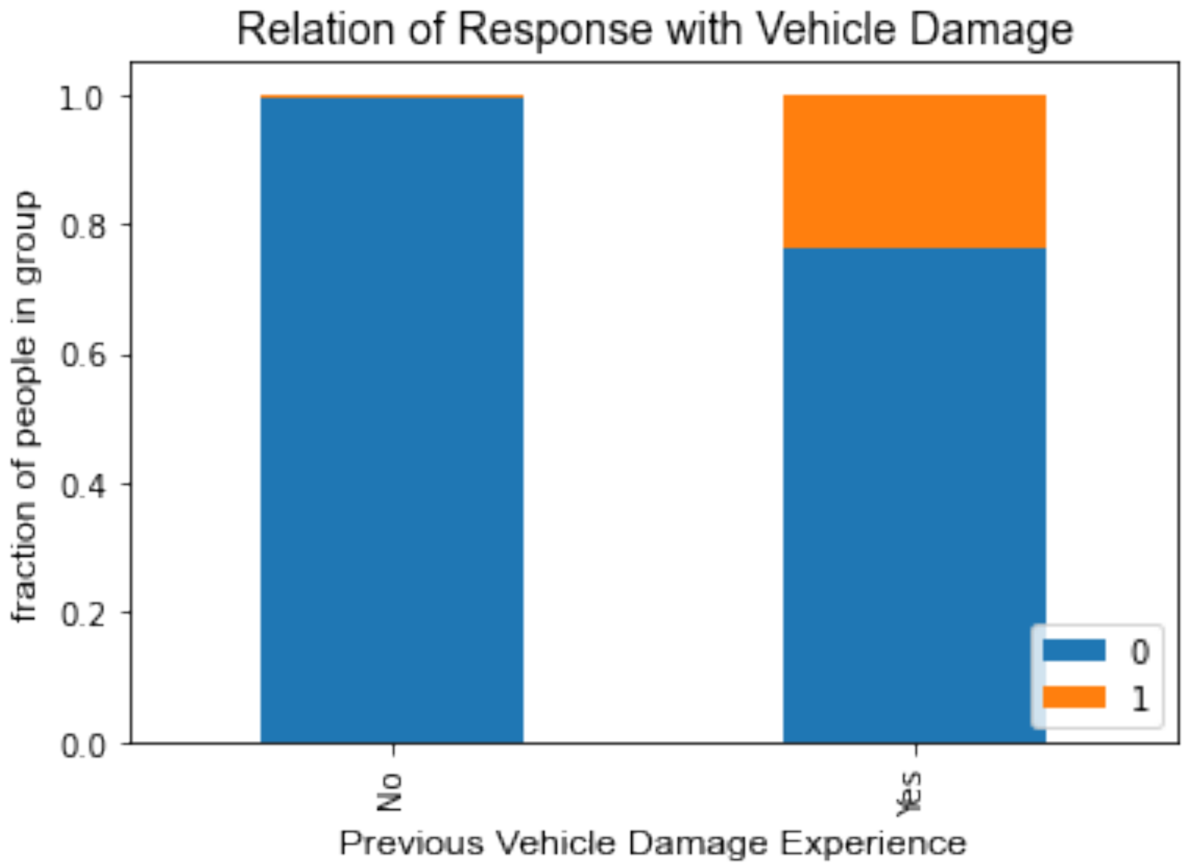
Leverage machine learning techniques to ***predict*** if someone will be interested in purchasing vehicle insurance (classification)

Dataset is from a health insurance provider leveraging their own customer's data

The ***target variable*** is the ***customer's response***, if they are interested in purchasing insurance

RECAP-EXPLORATORY DATA ANALYSIS

Those with prior vehicle damage experiences are more likely to be interested in purchasing vehicle insurance when compared to customers with no past vehicle damage



CROSS VALIDATION

- Data split basic train test and then kfold
- Preprocessed data-numerical(standard scaler), categorical (onehot and ordinal)
- Kfold Cross Validation/GridSearchCV
- Three machine learning algorithms were selected: Logistic Regression, Random Forest Classifier, Support Vector Machine
- Evaluation Metric - Accuracy Score
- See details for hyperparameters which were tuned in table below

ML Algorithm	1-Parameter	1-Values	2-Parameter	2-Values
LR	C	logspace(-2, 2, num=8)	NA	NA
RF	Max Depth	1, 3, 10	Max Features	.5-1
SVM	C	logspace(-3, 4, num=8)	Gamma	logspace(-3, 4, num=8)

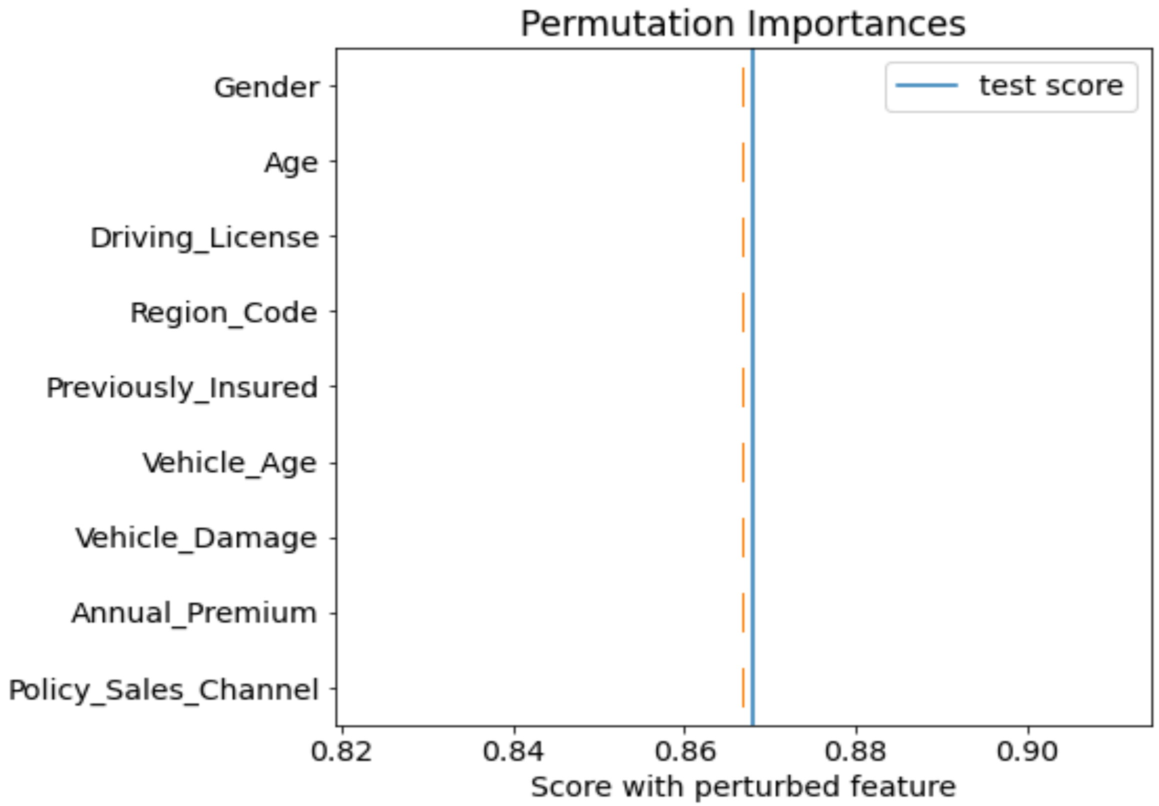
RESULTS-MODEL OUTCOMES

The machine learning algorithms do not reflect a significantly improved accuracy score when compared to the baseline model

Baseline = .877

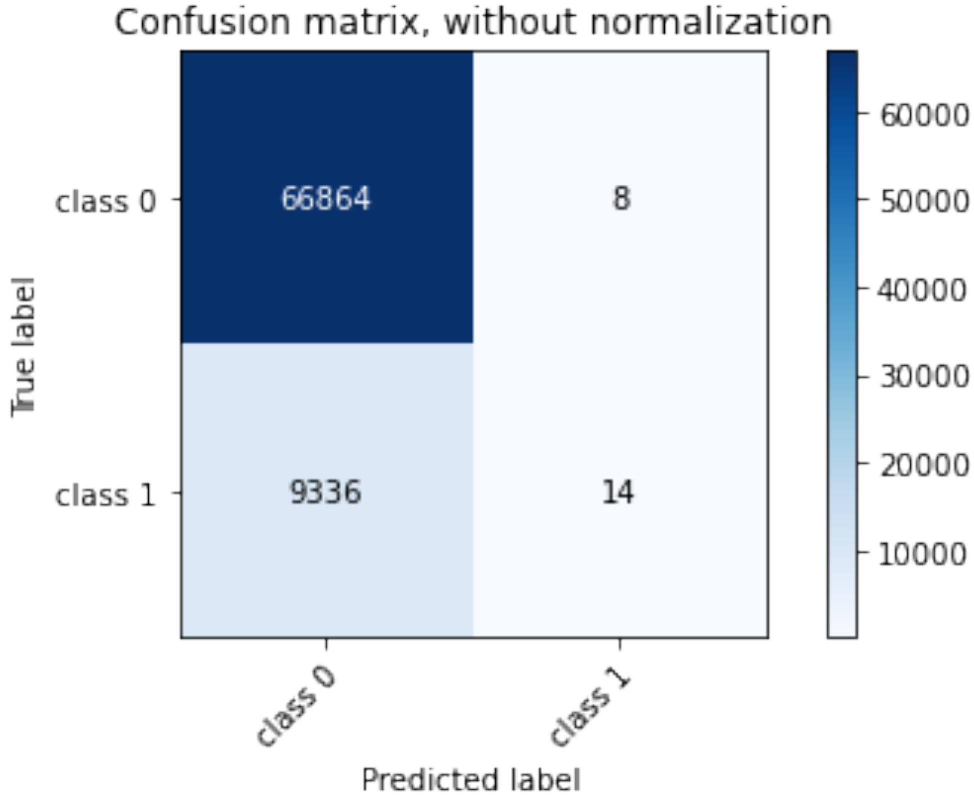
Machine Learning Algorithm	**Mean**	**SD**	**Best Parameters**
Logistic Regression	.867	.001	C= .01
Random Forest Classifier	.892	.011	Max Depth = 1, Max Features = .5
Support Vector Classifier	.892	.012	C= .001, Gamma= .001

RESULTS-FEATURE IMPORTANCE & CONFUSION MATRICES



*Random Forest

Feature importance shows us that the features were really not that important



*Logistic Regression

Confusion Matrices were developed for each algorithm to help measure effectiveness of the model

OUTLOOK

Look into more than ***three machine learning algorithms*** used in the model approach

Tune ***additional hyperparameters*** which can provide additional confidence in the predictions

Dive deeper into feature selection by performing ***a more comprehensive EDA process*** or use various mixtures of features to produce additional features

QUESTIONS?

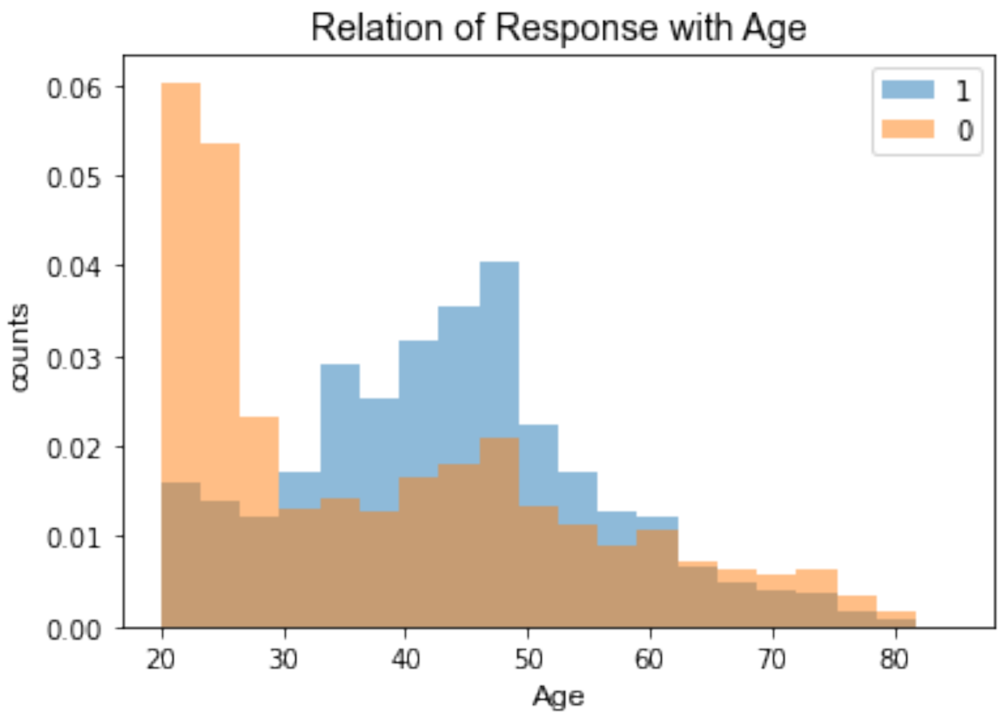
APPENDIX

PREPROCESSING DATA

The data was preprocessed inside the machine learning pipeline. See below on why a specific preprocessor was chosen for each feature.

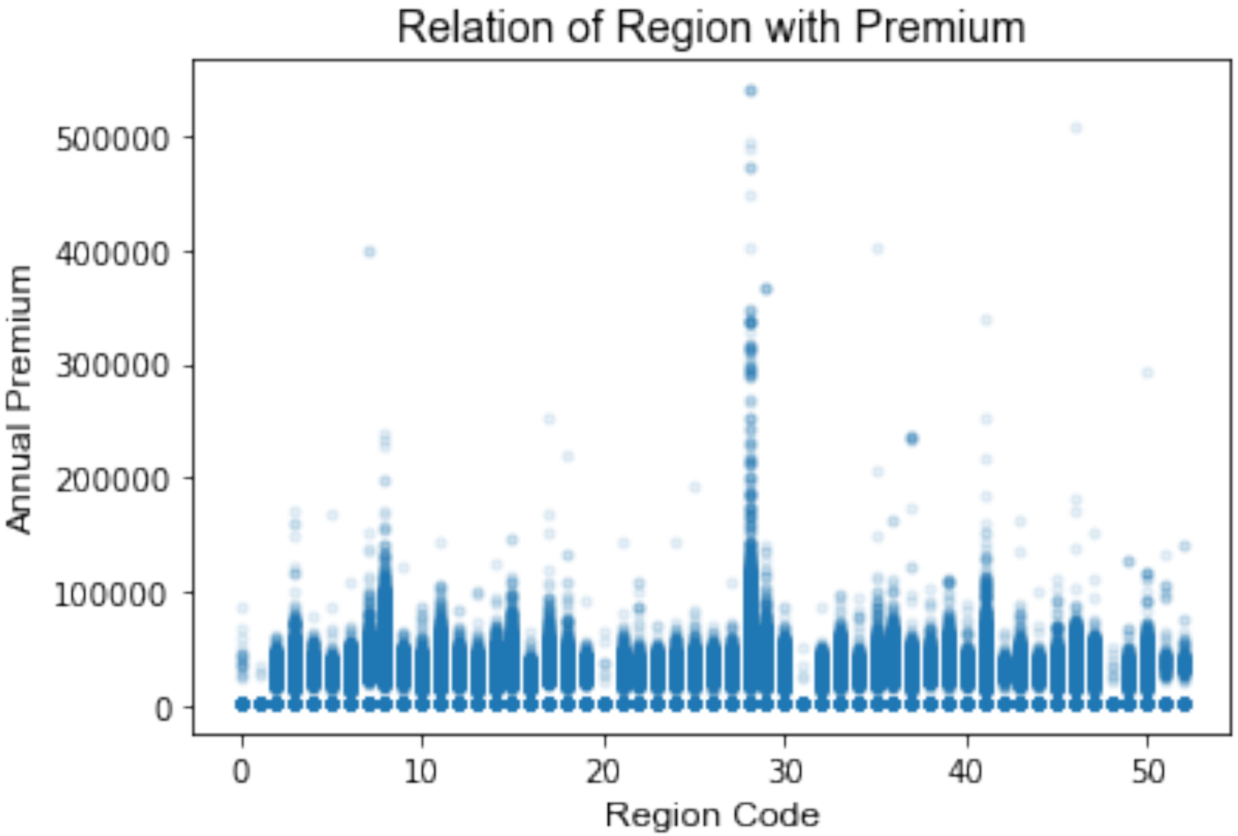
Feature	Numerical/Categorical	Preprocessor	Reason
Gender	Categorical	OneHotEncoder	Categorical can't be ordered
Age	Numerical	StandardScaler	Numerical reasonable bounds
Driving License	Categorical-0,1	OneHotEncoder	Categorical can't be ordered
Region Code	Numerical	StandardScaler	Numerical
Previously Insured	Categorical-0,1	OneHotEncoder	Categorical can't be ordered
Vehicle Age	Categorical	OrdinalEncoder	Categorical can be ordered
Vehicle Damage	Categorical-0,1	OneHotEncoder	Categorical can't be ordered
Annual Premium	Numerical	StandardScaler	Numerical, tailed distribution
Policy Sales Channel	Numerical	StandardScaler	Default
Response/Target	Categorical-0,1	OneHotEncoder	Default

EXPLORATORY DATA ANALYSIS

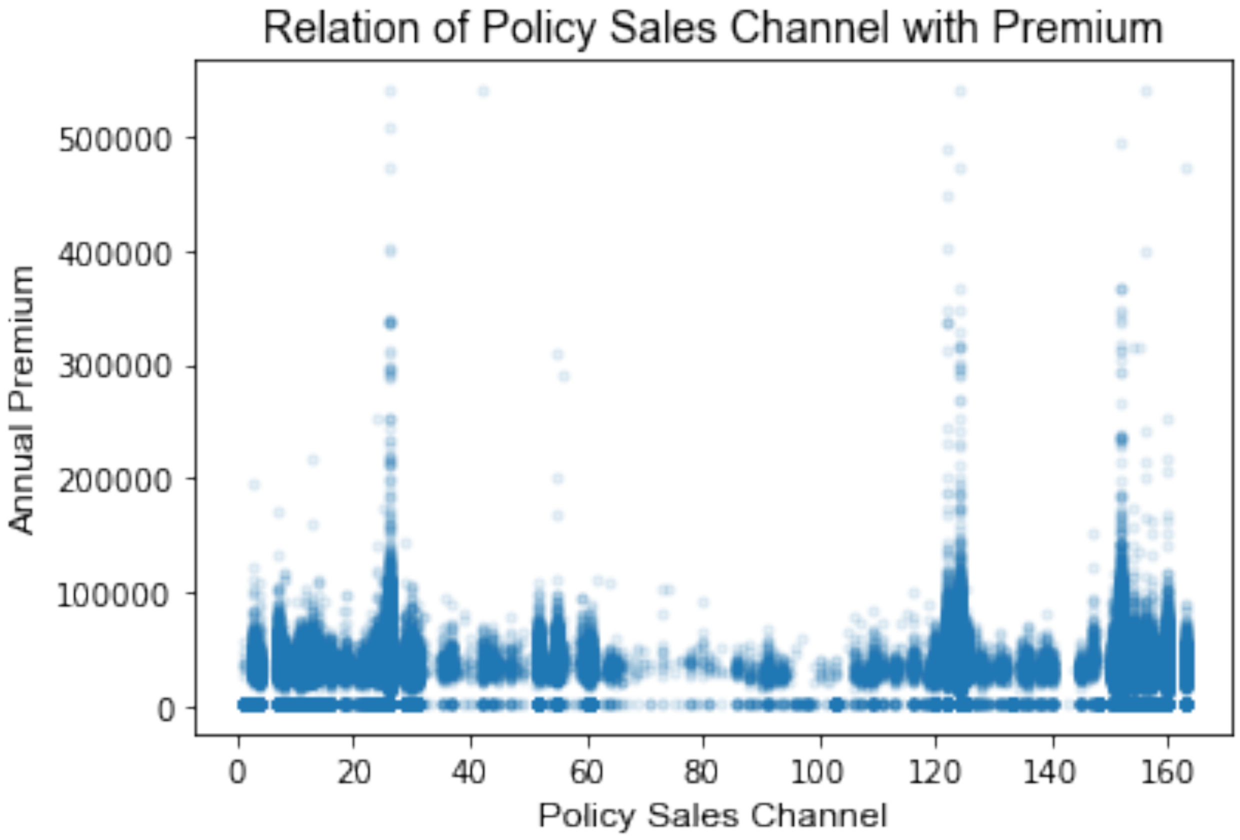


Customers between the ages of 35-50 are more likely to be interested in purchasing vehicle insurance when compared to customers between the ages of 20 to 30

EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS

Region Code and Policy Sales Channel show us a relationship that can better understand sales distribution options