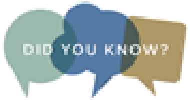
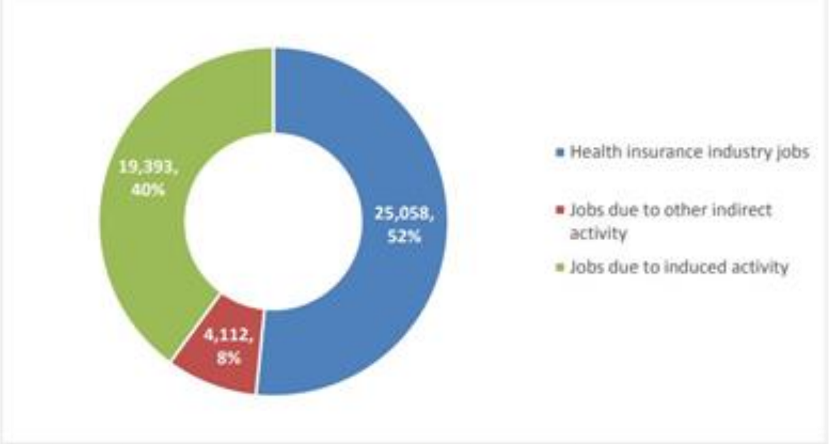
Insurance Cross-Sell Prediction

Impact of Health Insurance Industry

The Health insurance industry generates $15.5 billion in economic activity and pays over $209 million in state taxes.*(CT ERC,2019)*



"One new job in the insurance industry on average adds 4.4 jobs to the Connecticut economy through induced and indirect effects." *(CT IFS, 2021)*



Source: Connecticut Association of Health Plans (CTAHP); Emsi 2019.2; IMPLAN 2017 model for Connecticut; CERC calculations.

Agenda

- Introduction
- Problem Statement
- Data Set Description
- Data Cleaning and Exploration
- Models
- Findings
- Recommendations

Introduction

- ❖ ABC Insurance is an insurance company that currently only sells one product, employer sponsored health insurance.
- ❖ ABC Insurance wants to expand its business by providing Auto insurance to its current customer base.

Problem Statement

- To build a model that predicts whether a Health Insurance customer will purchase an Auto Insurance from ABC Insurance.

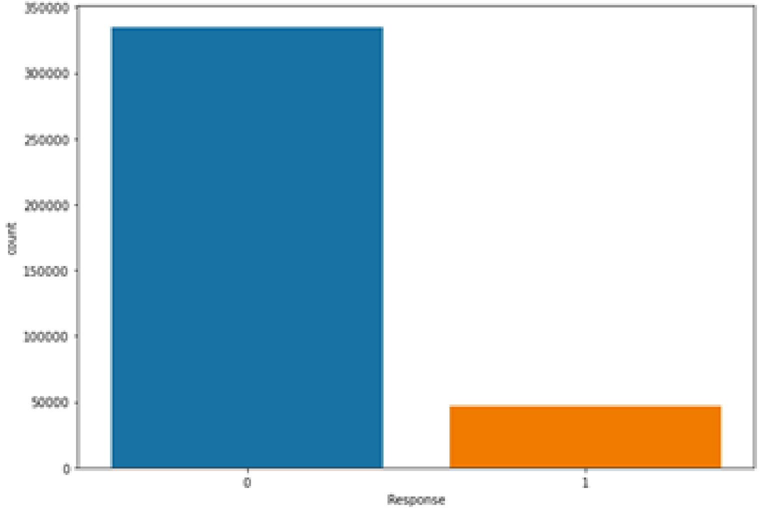
- It would be beneficial for the company to plan its communication strategy appropriately in order to reach out to those customers and optimize its business model and revenue.

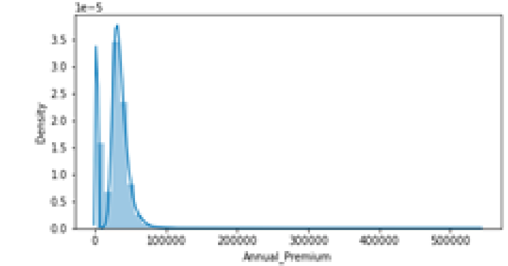
Data Set Description

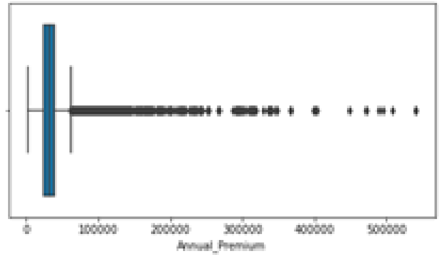
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Driving_License	0 : Customer does not have DL, 1 : Customer already has DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have 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	Anonymised 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

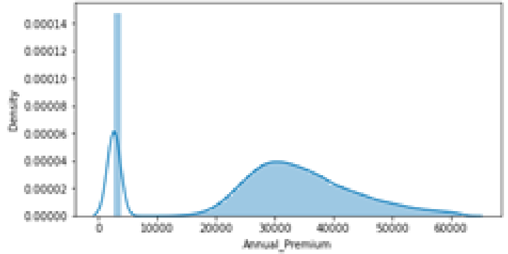
Data Cleaning and Exploration

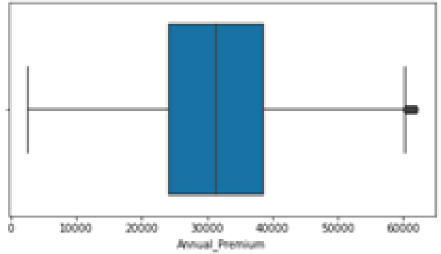
- ❖ No null values
- ❖ Imbalanced Data Set
- ❖ Outlier Treatment

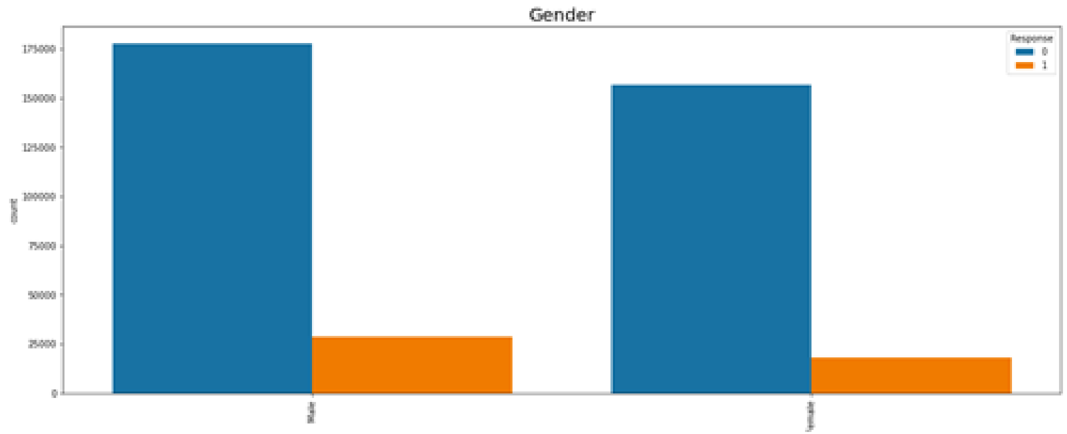


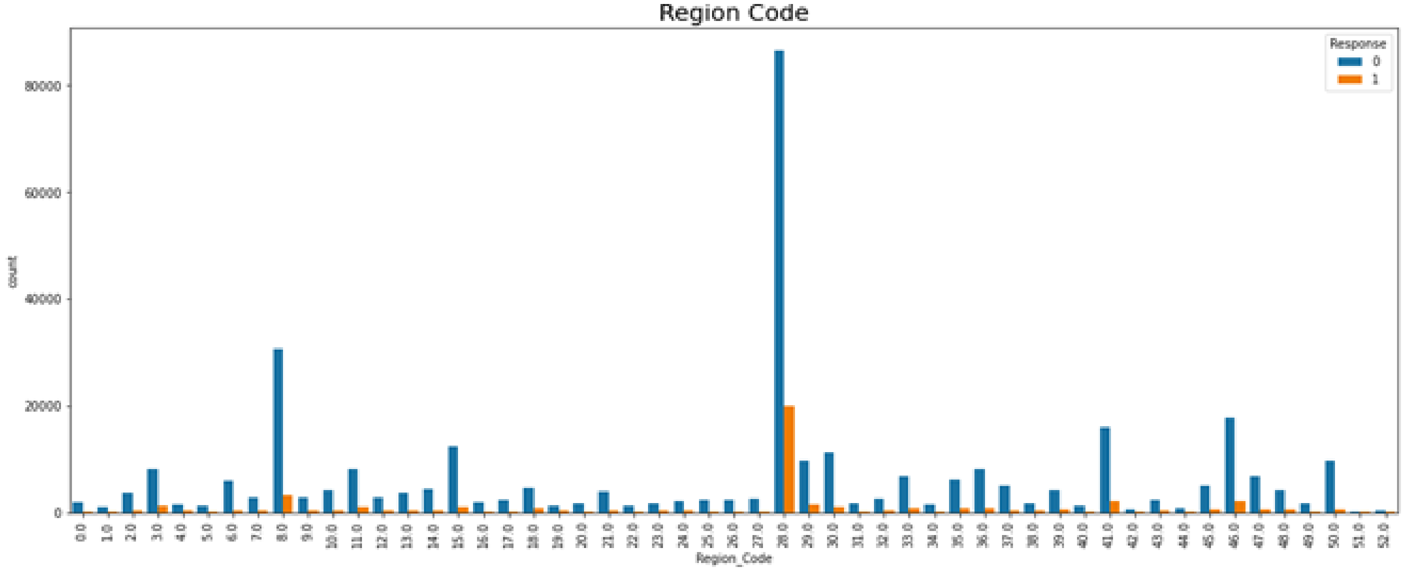


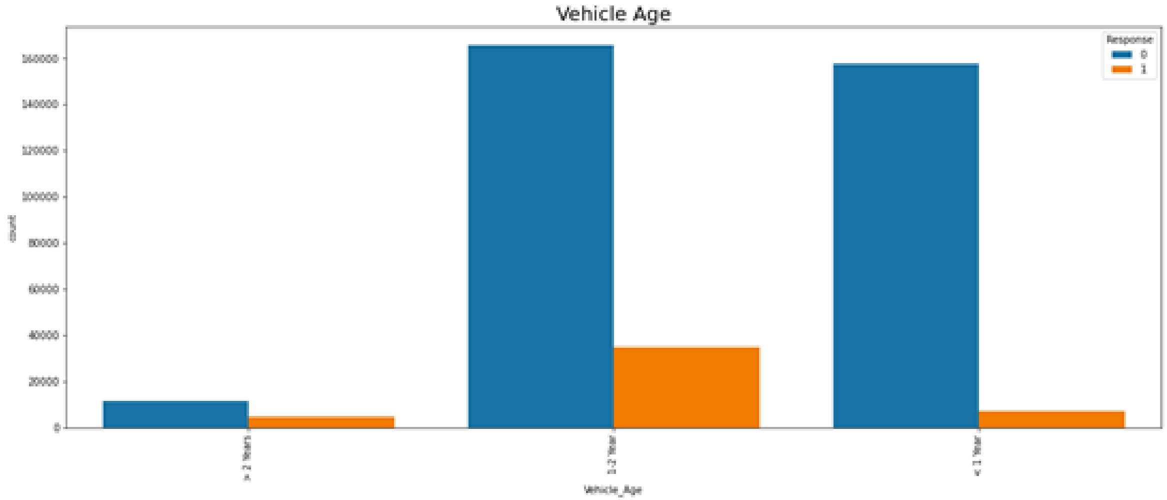


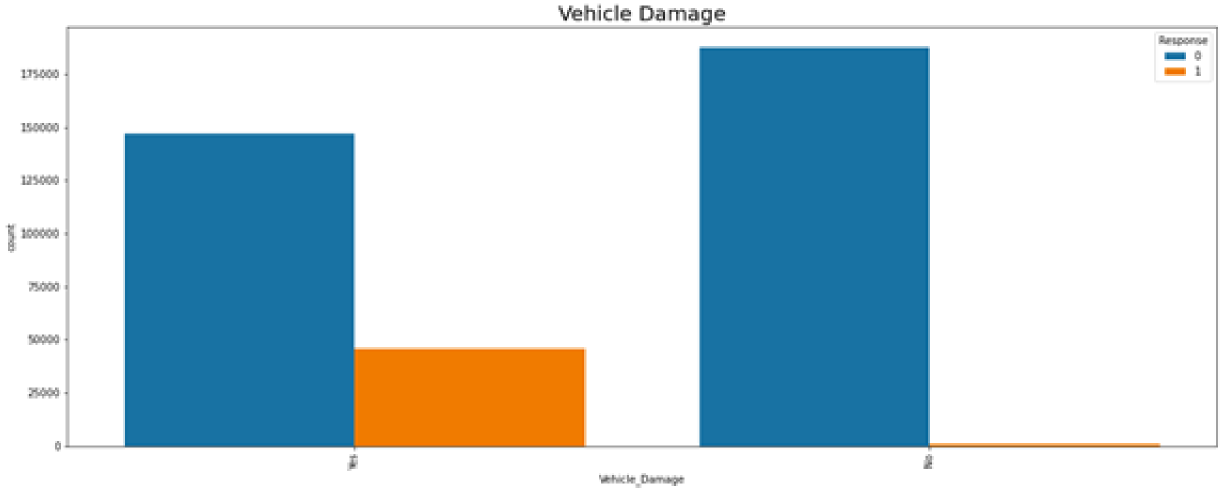


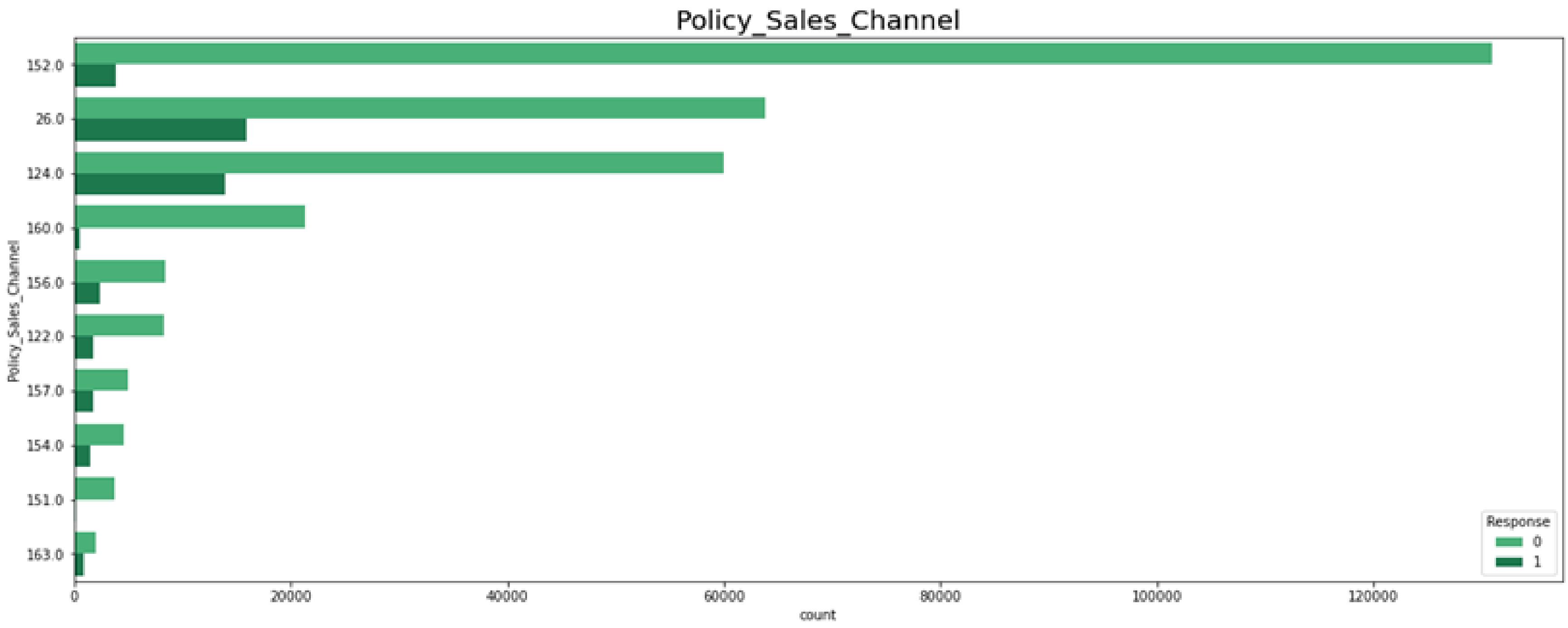



Gender
Response
0
1
count
Male
Female


Region Code
Response
0
1
count
Region_Code
80000
60000
40000
20000
0


Vehicle Age
Response
0
1
count
Vehicle_Age
> 2 Years
1-2 Year
< 1 Year

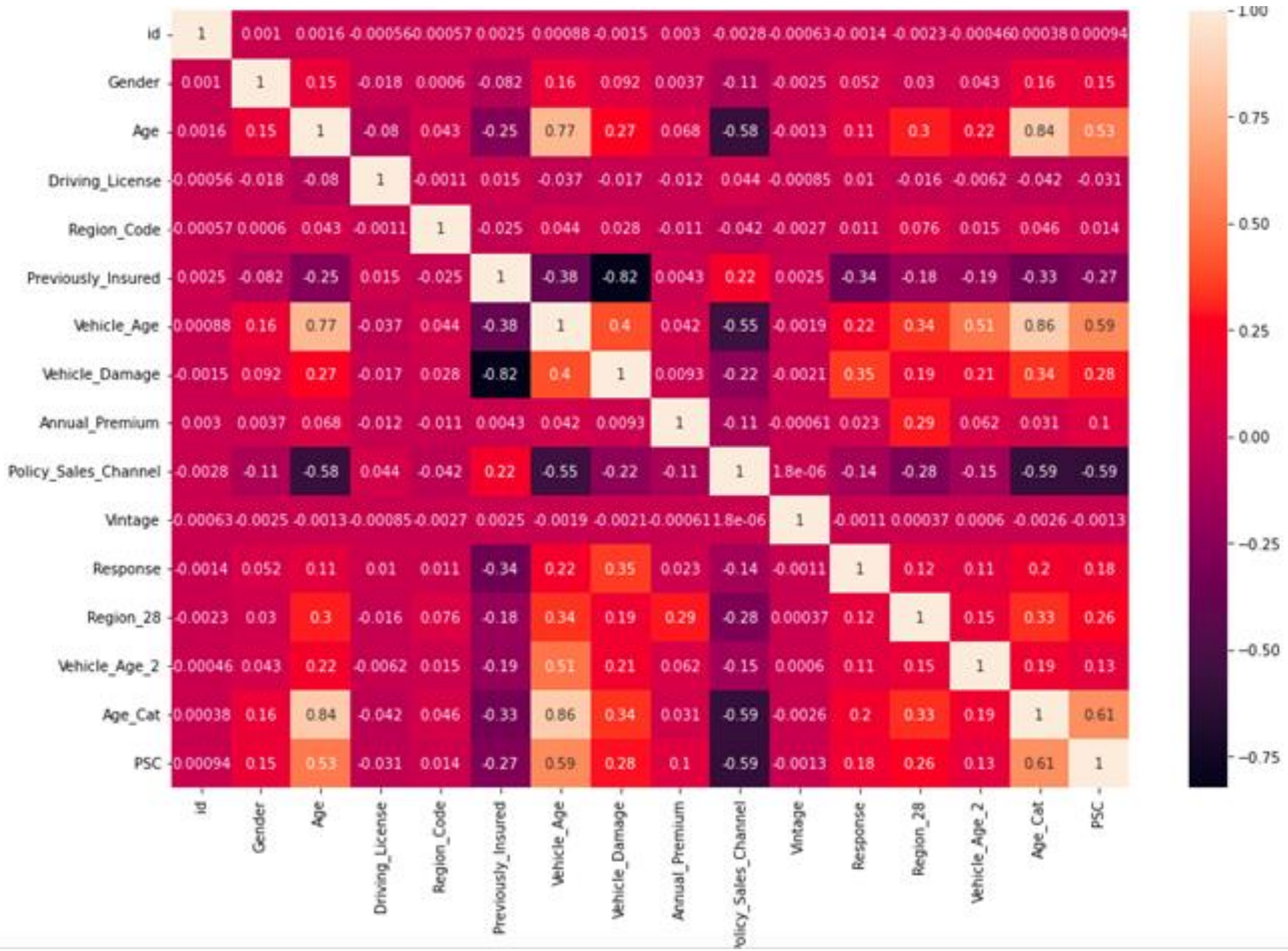

Vehicle Damage
Response
0
1
count
Vehicle_Damage
Yes
No


Policy_Sales_Channel
152.0
26.0
124.0
160.0
156.0
122.0
157.0
154.0
151.0
163.0
Policy_Sales_Channel
0
20000
40000
60000
80000
100000
120000
count
Response
0
1

- Encoding for Gender, Vehicle Age and Vehicle Damage
- Dummy variables for :

  Region 28 from Region

  Greater than 2 from Vehicle Age
- Age column binned to < = 27, >=28 and >=36
- Policy channel binned to 26 - 124 and 152 - 156



- Based on the results of the heat map, 8 columns were used for the model.
- ‘Previously_Insured', 'Vehicle_Age_2', 'Age_Cat', 'Vehicle_Damage', 'Annual_Premium', 'Vintage', 'PSC', 'Region_28' were the final columns used.

SMOTE

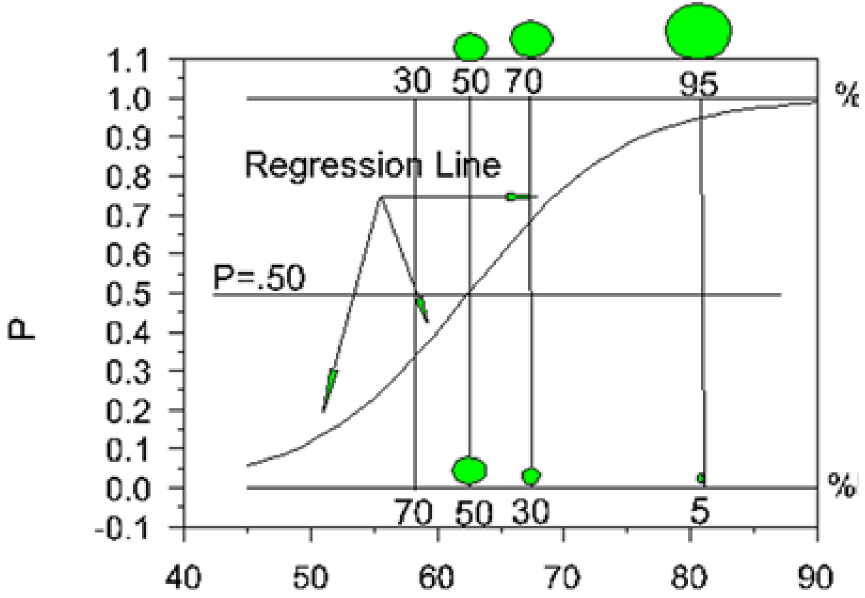
```
samplers = SMOTEENN(0.62,random_state=42)
X_comb, y_comb = samplers.fit_resample(X, y)
X_train, X_test, y_train, y_test = train_test_split(X_comb, y_comb, test_size=0.3, random_state=42)
print(X_train.shape) #Printing shape
print(X_test.shape)
```

```
(186748, 8)
(80036, 8)
```

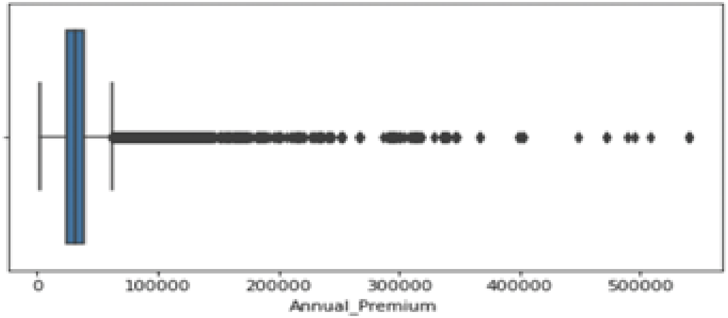
Logistic Regression

- Why Logistic Regression over linear regression?
- Logit Function & Bias Variance Tradeoff
- Maximum Likelihood Estimation
- Threshold Cutoff based on business requirements
- Odds Ratio =

$$\frac{\text{Prob (event)}}{\text{Prob (not event)}}$$



Lasso and Ridge



- Why?
  - Finding the best fit line
  - Overfitting
- Model performance based on L1 - Lasso & L2 Ridge Penalty
- Hyperparameter tuning

- Ridge Regression

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

  - Equivalent to minimize $\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$, subject to $\sum_{j=1}^{p} \beta_j^2 \leq C$

- LASSO Regression

$$\text{Minimize: } \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

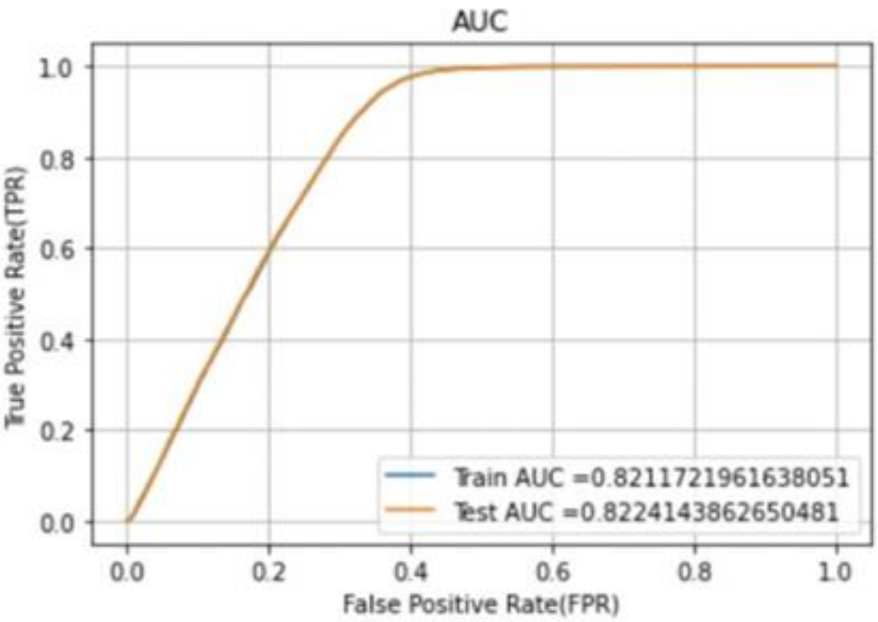
  - Equivalent to minimize $\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$, subject to $\sum_{j=1}^{p} |\beta_j| \leq C$

```python
# Using the Random search method to determine the best parametre of the model
grid_values = {'penalty': ['l1','l2'], 'C': [0,0.1, 0.01,1]}

lr = LogisticRegression()
grid_search = GridSearchCV(lr, grid_values, cv=10,
                           scoring='neg_mean_squared_error',
                           return_train_score=True)
grid_search.fit(X_train, y_train)
```

Model Performance and Findings

- Precision - Positive Predictive Value(PPV)
- Recall - True Positive Rate (TPR)
- F1 Score




```
LogisticRegression(C=0.1)
CF [[42864 12727]
 [ 8107 16338]]
precision 0.562119387579563
recall 0.6683575373286971
f1 0.6106522145393385
```

Classification Model

Step 1

```
# Creating the Randomn forest classifier
rfc=RandomForestClassifier(random_state=42)
param_grid = {
    'n_estimators': [200, 500],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_depth' : [4,5,6,7,8],
    'criterion' :['gini', 'entropy']
}
CV_rfc = GridSearchCV(estimator=rfc, param_grid=param_grid, cv= 3)
```

Step 2

```
# Training the Classifier
CV_rfc.fit(X_train, y_train)#
```

```
GridSearchCV(cv=3, estimator=RandomForestClassifier(random_state=42),
             param_grid={'criterion': ['gini', 'entropy'],
                         'max_depth': [4, 5, 6, 7, 8],
                         'max_features': ['auto', 'sqrt', 'log2'],
                         'n_estimators': [200, 500]})
```

Step 3

```
# Diaplaying the best Parametres
CV_rfc.best_params_
```

```
{'criterion': 'gini',
 'max_depth': 8,
 'max_features': 'log2',
 'n_estimators': 500}
```

Classification Model

Step 4

```
# Applying the traing model to test
rfc1=RandomForestClassifier(random_state=42, max_features='auto', n_estimators= 200, max_depth=8, criterion='gini')
clf=rfc1.fit(X_train, y_train)
final_model = clf. # Model 2
final_predictions = final_model.predict(X_test)

y_pred_new = final_model.predict_proba(X_test)[:,1]
```

Step 5

```
# Creating and displaying  the confusion matrix , Precesion , Recall , F1-Score , AUC score

print('CF', confusion_matrix(y_test, final_predictions))
print('precision', precision_score(y_test, final_predictions))
print('recall', recall_score(y_test, final_predictions))
print('f1', f1_score(y_test, final_predictions))
print('auc-roc-score', roc_auc_score(y_test, y_pred_new))
```

```
CF [[48802  6789]
 [ 4451 19994]]
precision 0.746518313855804
recall 0.8179177745960319
f1 0.7805887405325213
auc-roc-score 0.9380049581138761
```

Findings

```
LogisticRegression(C=0.1)
CF [[42864 12727]
 [ 8107 16338]]
precision 0.562119387579563
recall 0.6683575373286971
f1 0.610652214539385
```

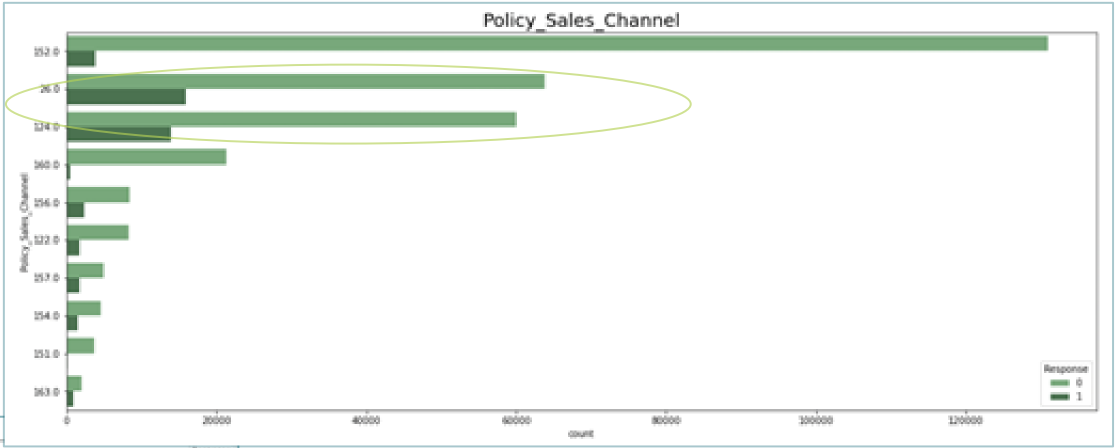
To

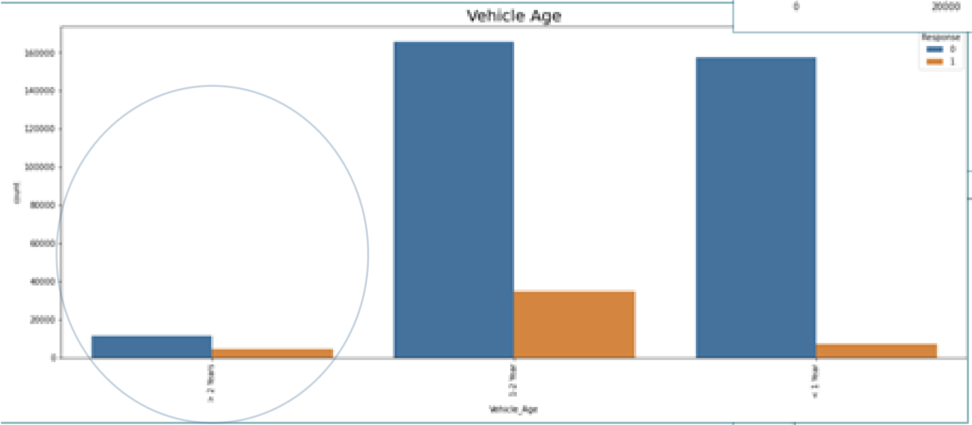
```
CF [[48802  6789]
 [ 4451 19994]]
precision 0.746518313855804
recall 0.8179177745960319
f1 0.7805887405325213
auc-roc-score 0.9380049581138761
```

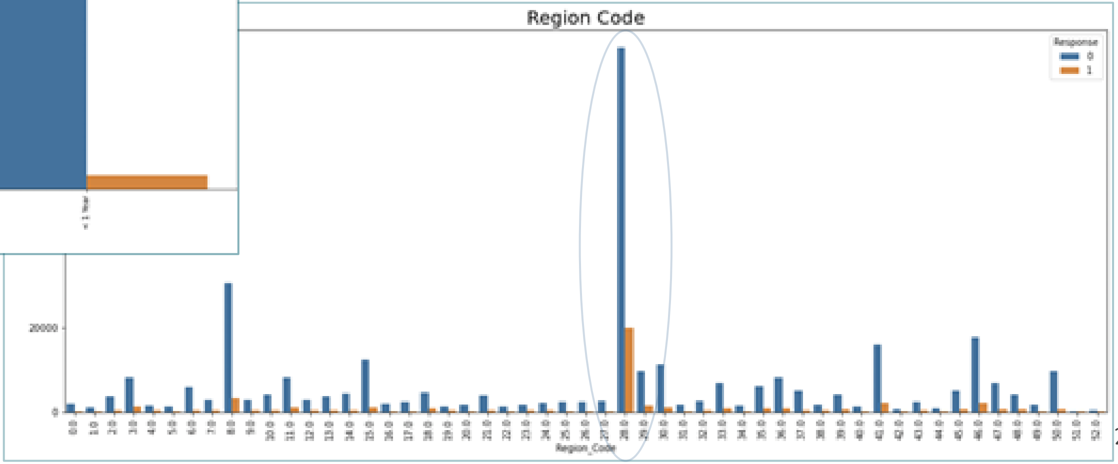
1st Model

2nd Model

Findings


Policy_Sales_Channel
Response
0
1
count
Policy_Sales_Channel

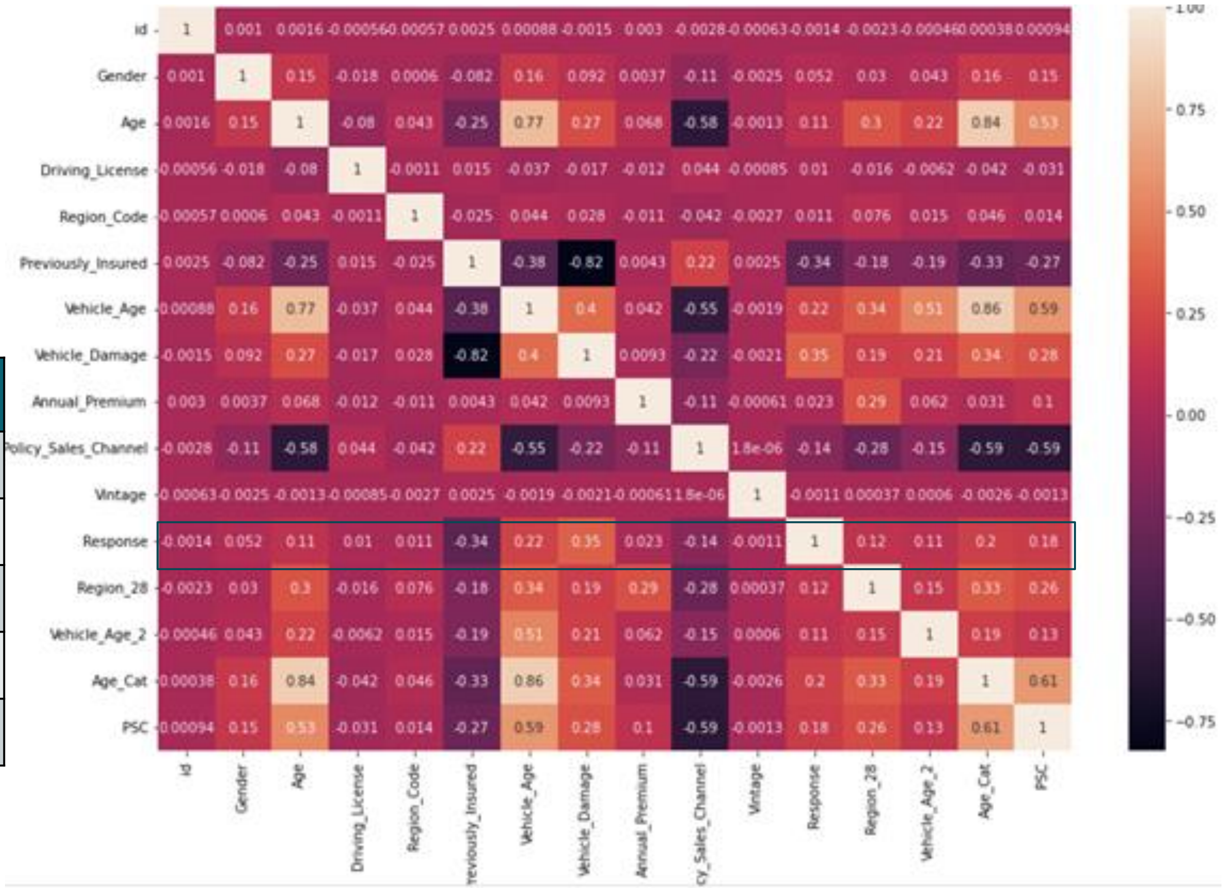

Vehicle Age
Response
0
1
Vehicle_Age
count


Region Code
Response
0
1
Region_Code

Findings

Correlations

Response & Vehicle_Damage	0.35
Response &Previous_Insured	−0.34
Response &Vehicel_Age	0.22
Response &Age_Cat	0.20
Response &PSC	0.18
Response &Region_28	0.12



Recommendations For Business

- ❖ Marketing team can use our cross sell predictions to grow their business and change business strategies
- ❖ Customers got his/her vehicle damaged in the past.
- ❖ Customer doesn't have a vehicle insured experience.
- ❖ Vehicles age greater 1-2 years
- ❖ Policy Sales Channel (*outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.)* :26, 124
- ❖ Customer Region_Code : Region 28 - Marketing Strategies


Thank You