Travel Insurance Claim Predictor

Problem Statement

To predict if the claim request of a travel insurance is genuine or fake

- **Potential Business Problem:** With so many claims on daily basis it is difficult to study the authenticity of claims manually . Thus a automated system is needed to predict If a claim should be approved or not

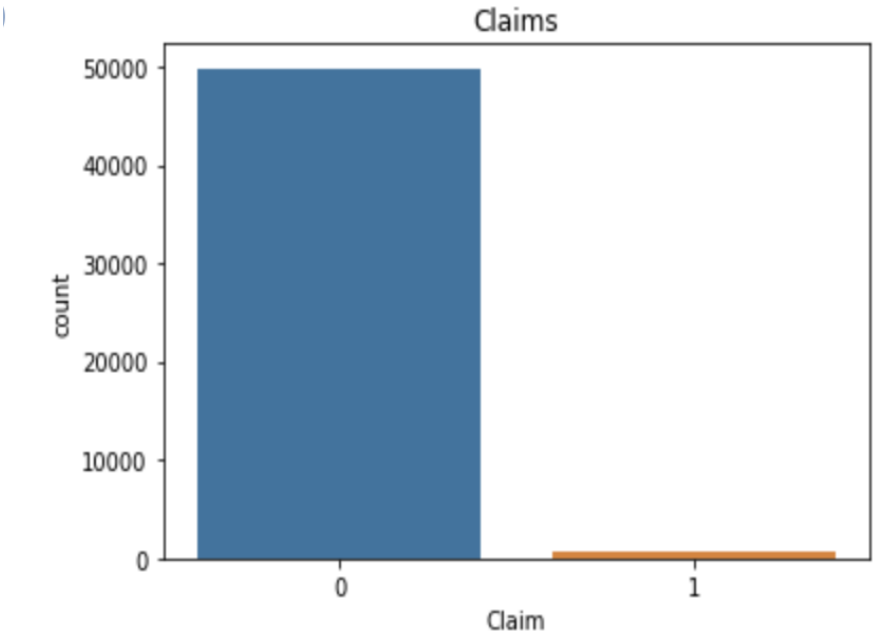
We have information of roughly 50k consumer claims. There are around 11 features and 1 target variable.

Feature Name	Type	
Agency Type	Object	It is eithe
Distribution Channel	Object	It is eithe
Destination	Object	What was
Net Sales	float64	What was insurance

Evaluation metric

- Since this is highly imbalanced data, accuracy is not the right measure for the performance of our model.
- In this case we are going to consider Precision and Recall as the best measure for the performance of the model.

Exploratory Data Analysis- Target variable



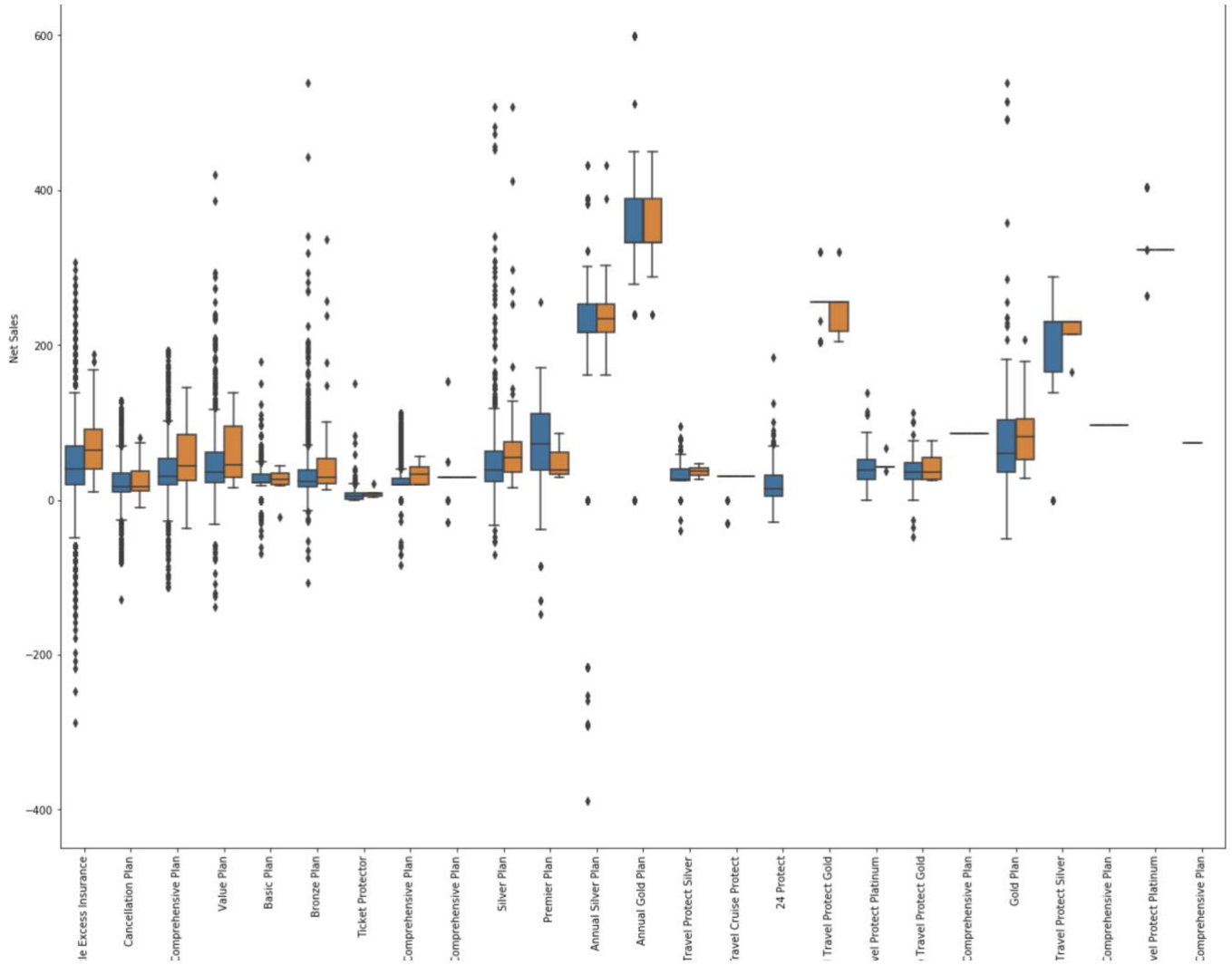
Claim
Boolean

Distinct count	2
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0
Mean	0.014658

Toggle details

Value	Count	Frequency (%)
0	49812	98.5%
1	741	1.5%

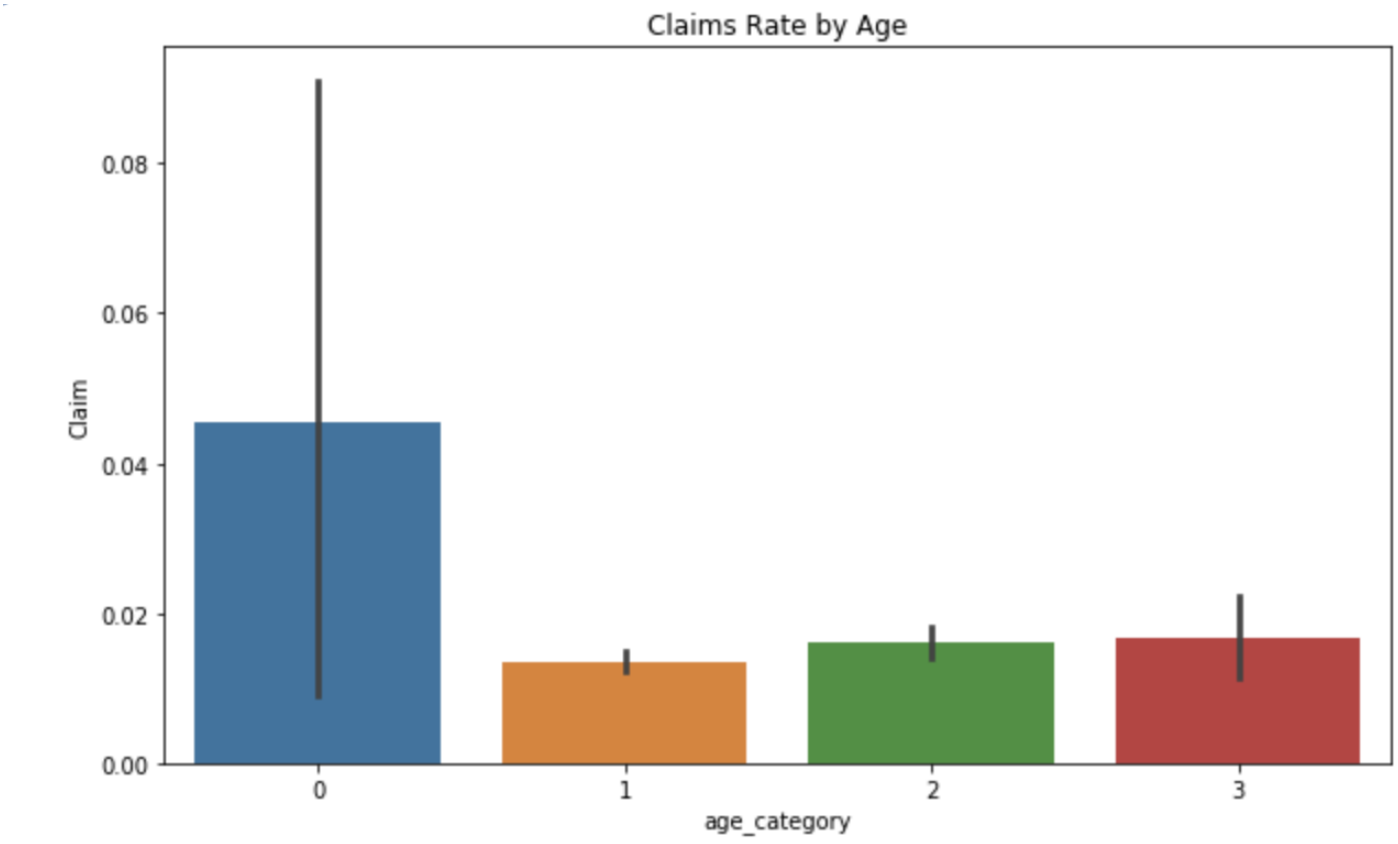
This is a boxplot **for net sales vs various products**



Since in this null check analysis, we can see that Gender has more than 50% null values.
Thus it holds no relevance and can be dropped

```
ID                          0
Agency                      0
Agency Type                 0
Distribution Channel        0
Product Name                0
Claim                       0
Duration                    0
Destination                 0
Net Sales                   0
Commision                   0
Gender                  35953
Age                         0
age_category                0
dtype: int64
```

- Since Age feature has so many values almost all unique do for better understanding we have divided age into categories.



Model Tuning

One Hot Encoding :

Since most of the features are categorial, we used one hot encoding to transform the data

PCA:

Since destination variable had high cardinality, the features size increased to 156. We used PCA for dimensionality reduction and chose 130 features.

Undersampling:

Since the data set is highly imbalanced, we did undersampling and made data with 50% positive claims data and 50% negative claims data.

Model	Precision	Recall
Logistic regression	0.82	0.75
Random Forest	0.74	0.69
SVM	0.79	0.74

Next Steps:

1. Using SMOTE on the data for data balanacing.
2. Hyper Parameter tuning.