Travel Insurance

HACKATHON 1

16TH -23RD NOV'19

VAIBHAV SINGH

BRUNO FERNANDES

SWAPNIL MASTEKAR

Problem Statement

Predict whether to sanction the claim

Description of problem:

- Many risk factors for customers during travel like loss of baggage , airline cancellations , health issues etc.
- Company offers insurance against these risks across multiple products like 1-way travel insurance, 2-way insurance, insurance against cancellations and so on
- Company receives thousand of claims which need to be automatically and accurately predicted
- Wrongly denying a genuine claim could lead to lawsuits against the company and approving the wrong claim would lead to a loss.
- Vast amount of information about policyholders is available which need to be analyzed to develop profiles of high and low insurance risks

Business Problem

Potential Business Problems :

- Predicting the claims automatically will lead to faster processing of claims with minimal manual intervention
- This will eventually lead to customer satisfaction which in turn will potentially bring more customers and purchase of policies

Why Solve this problem?:

- Identify common features of policies where claim is processed
- Predicting the claims accurately will avoid any losses or any legal investment in case of wrong prediction

Dataset

Dataset Information :

- The data consists of records of roughly 62288 clients and 12 features
- There are 11 predictors and 1 target that describes whether the claim will be processed or not

Below are some of the features and the target variable

Feature	Feature Type	Feature Description
Claim	Binary	Target: Claim Status
Agency	categorical	Name of agency
Agency.Type	categorical	Type of travel insurance agencies
Distribution.Channel	categorical	Distribution channel of travel insurance agencies
Product.Name	categorical	Name of the travel insurance products
Duration	numeric	Duration of travel
Destination	categorical	Destination of travel
Net.Sales	numeric	Amount of sales of travel insurance policies
Commission	numeric	The commission received for travel insurance agency
Gender	categorical	Gender of insured
Age	ordinal	Age of insured
ID	nominal	Identification of policy

Evaluation Metrics

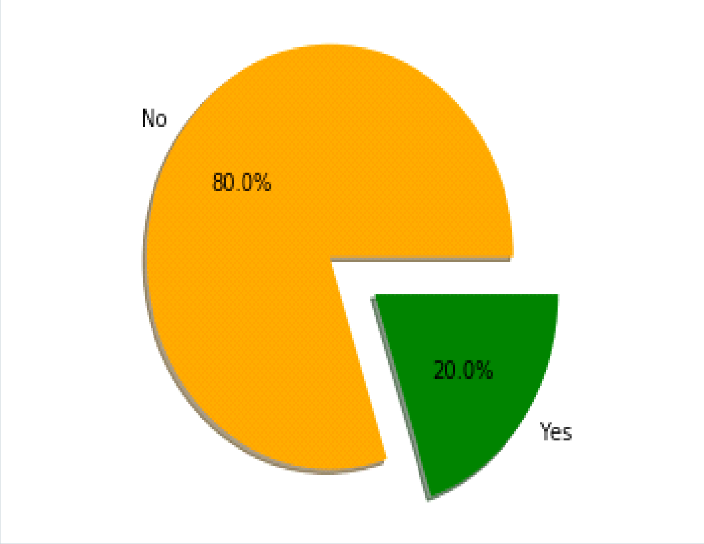
The evaluation metric for this project is precision score.

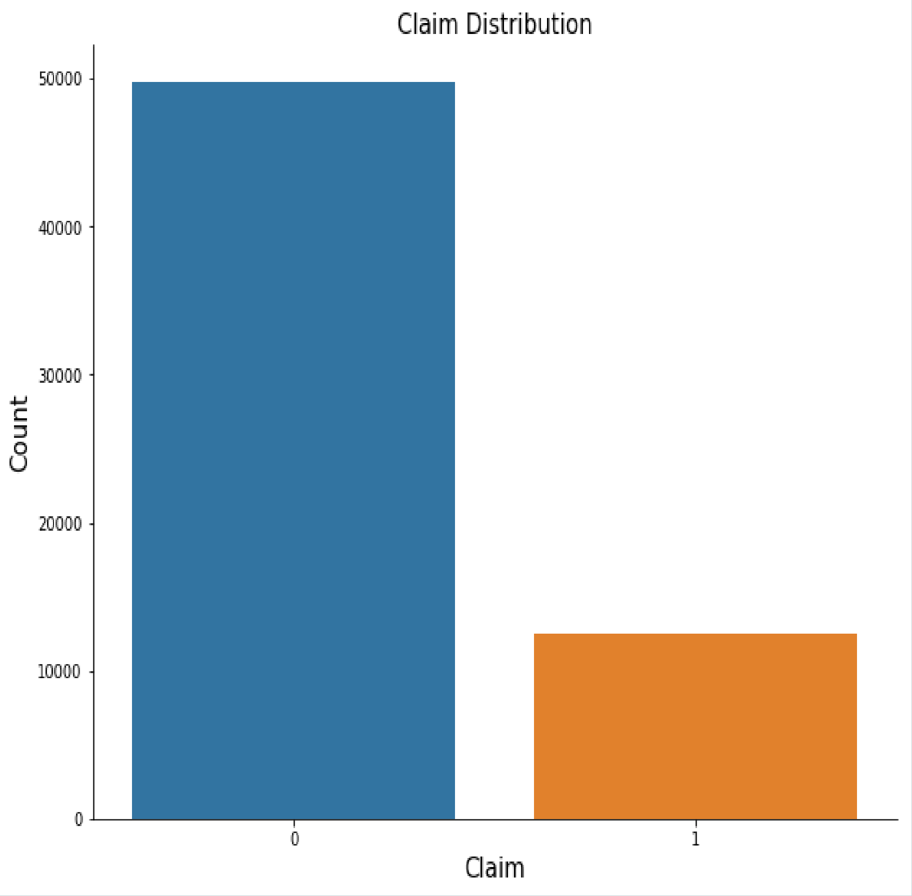
- False Positive – predicted claim is processed, but there is no claim.
- True Positive- predicted there was a claim and there was a claim
- For the use case, False positives must be reduced. So precision to be given more importance.
- Precision=True positive/True Positive + False Positive

Target Variable distribution

Class Imbalance

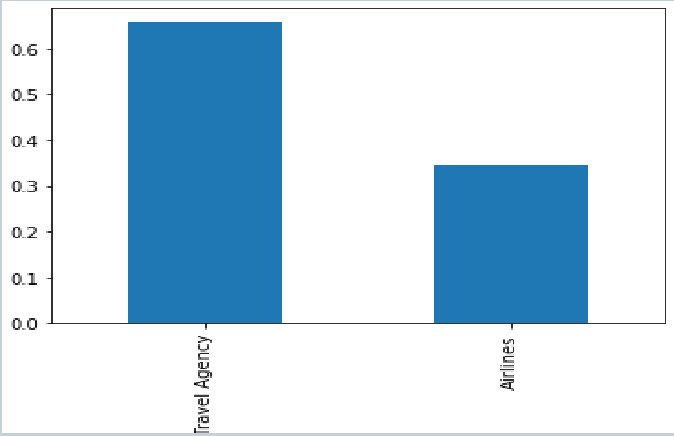
Class Imbalance is identified with distribution of 80% where claim is not processed and 20% where claim is processed



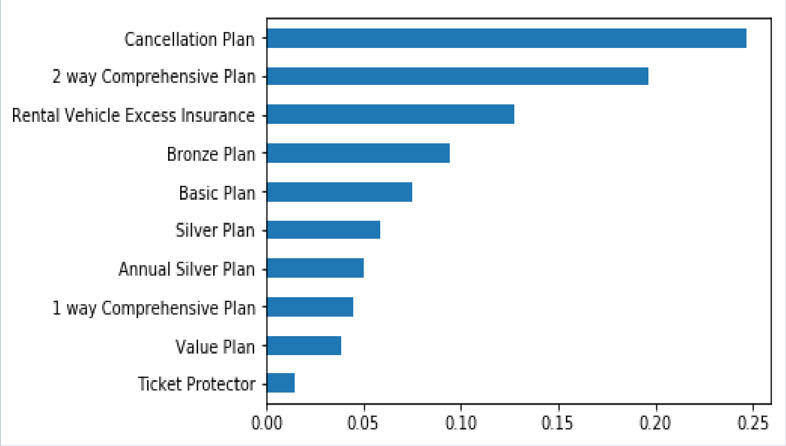


Exploratory Data Analysis

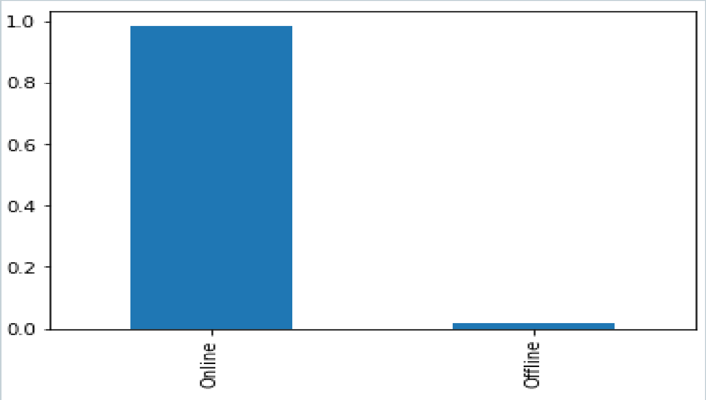
Agency Type



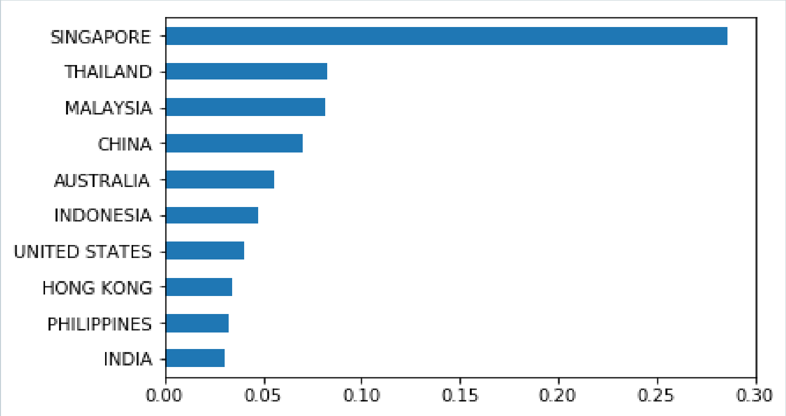
Top 10 products



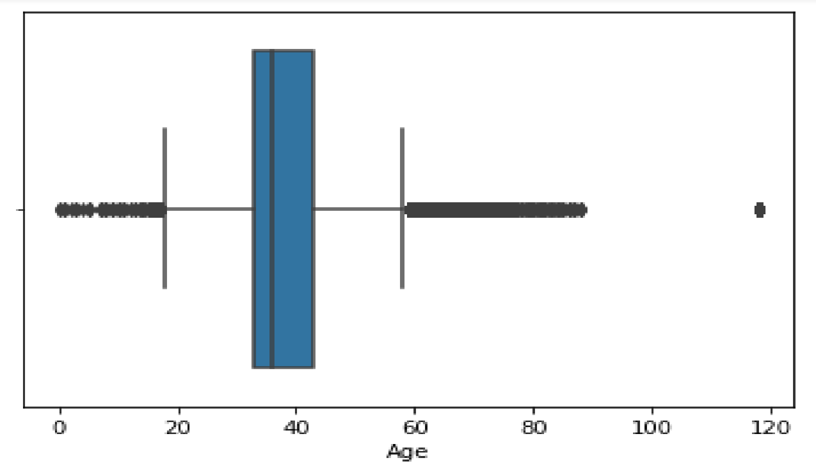
Distribution Channel

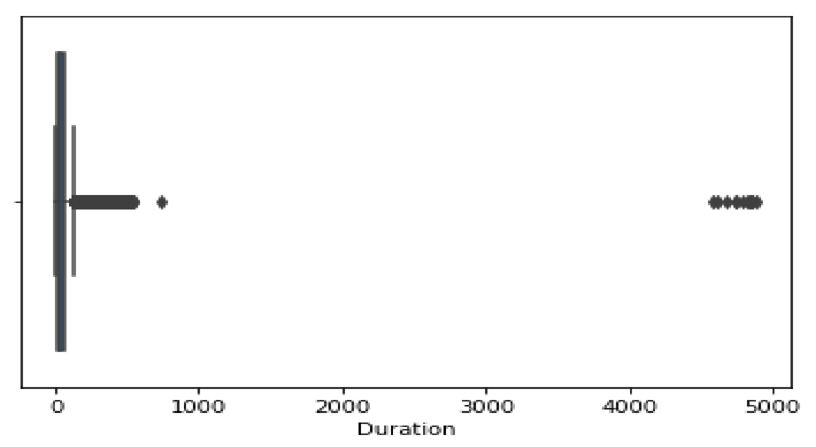


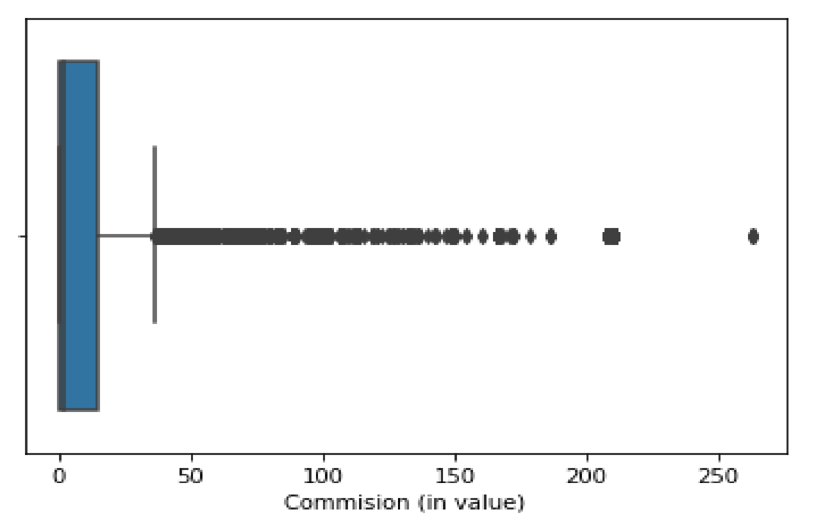
Top 10 Destinations

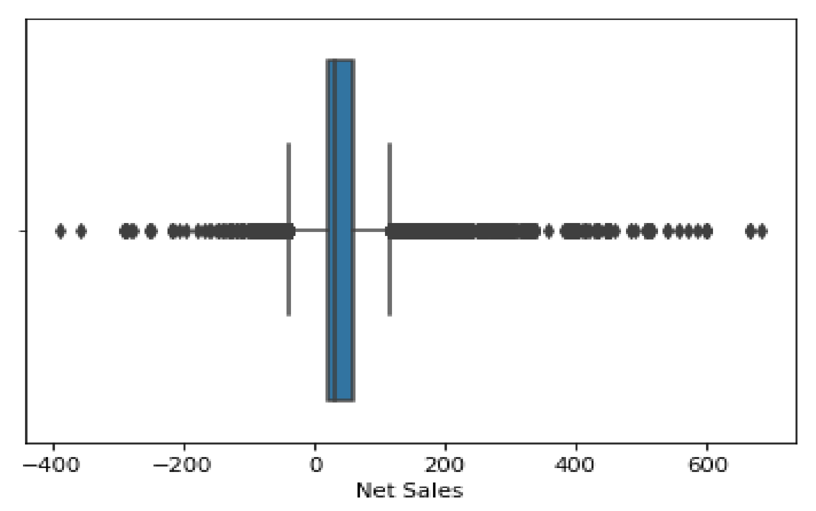


Box plot for numeric variables



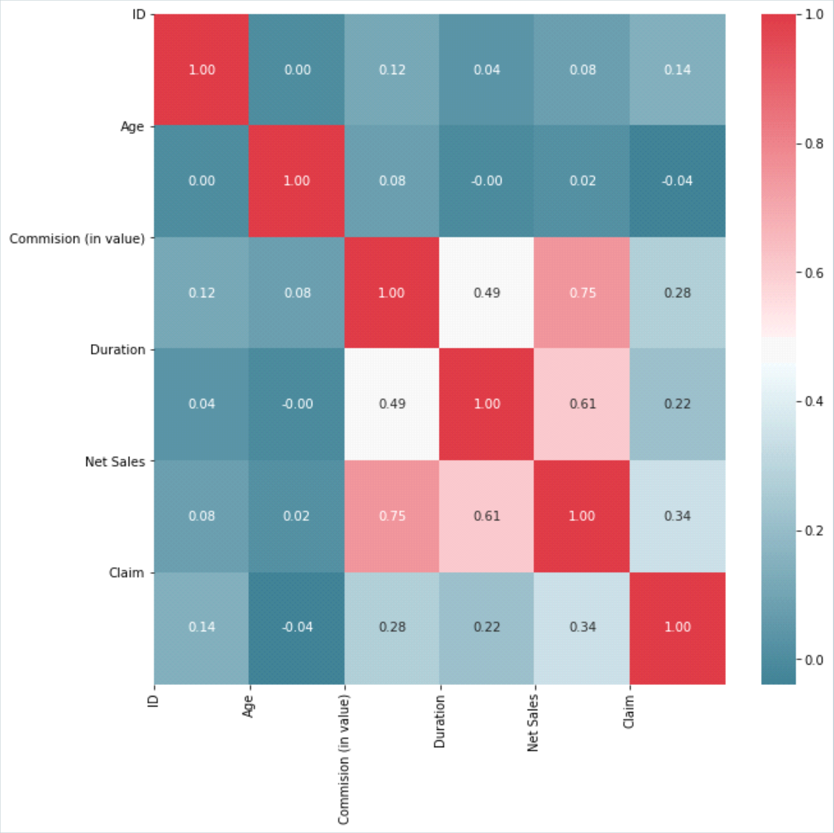




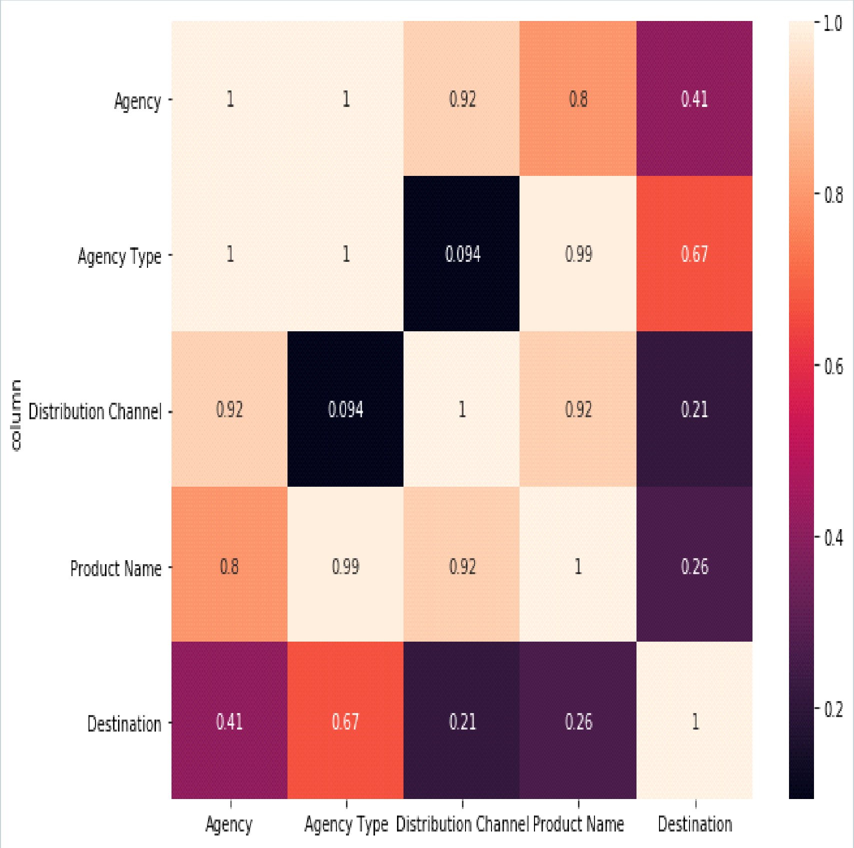


Heat Map to determine relationship of variables

Numeric Variables



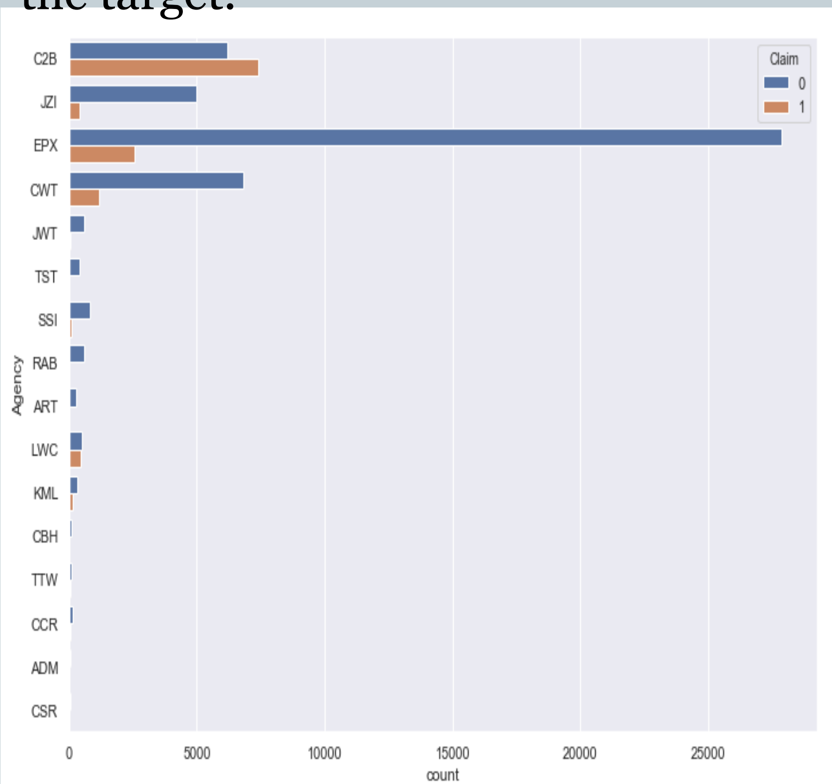
Categorical Variables

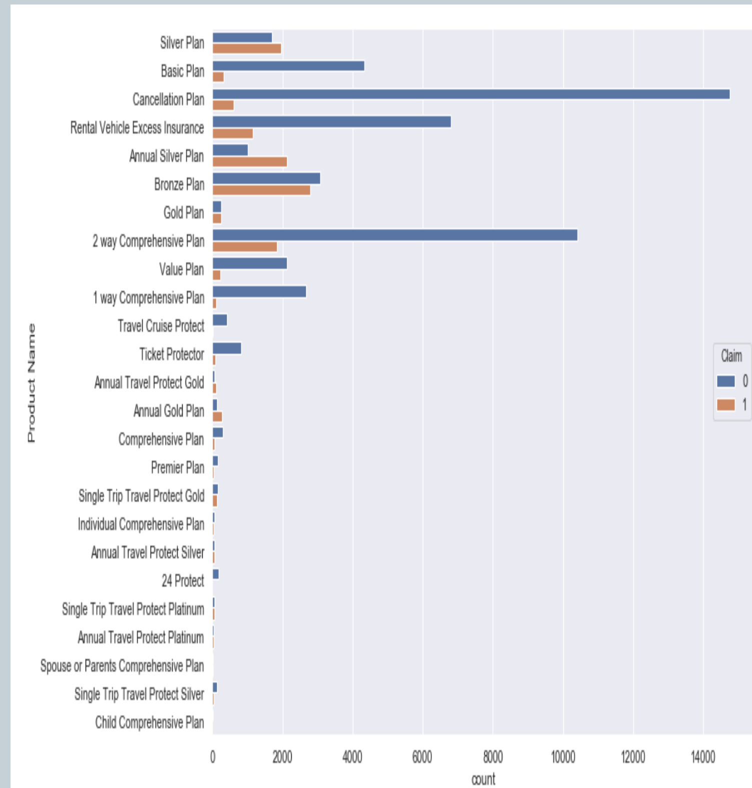


Label Encoding was performed on categorical variables

Bivariate Analysis

Below are the bivariate analysis of features Agency Type and Destination w.r.t the target.





Most of the claims processed are policies sourced through Agency - C2B

Bronze plan products have more claims processed

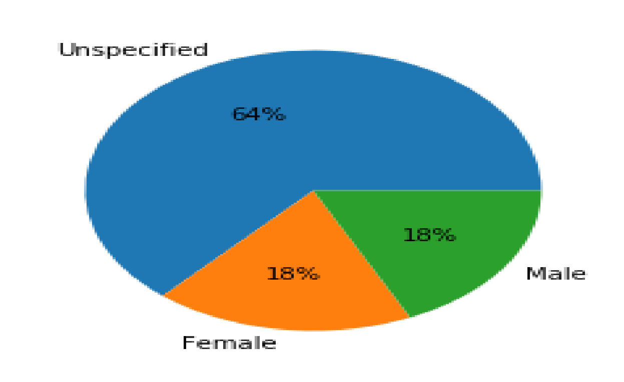
Pipeline

Outlier treatment :

Variables	Outlier before winsorization	Outlier after winsorization
Age	4687	0
Commision (in value)	7360	7360
Duration	6758	6758
Net Sales	5563	0

•The Outliers in the continuous features were detected and treated using Winsorization

Missing Values



64% missing values in Gender Column and hence dropped

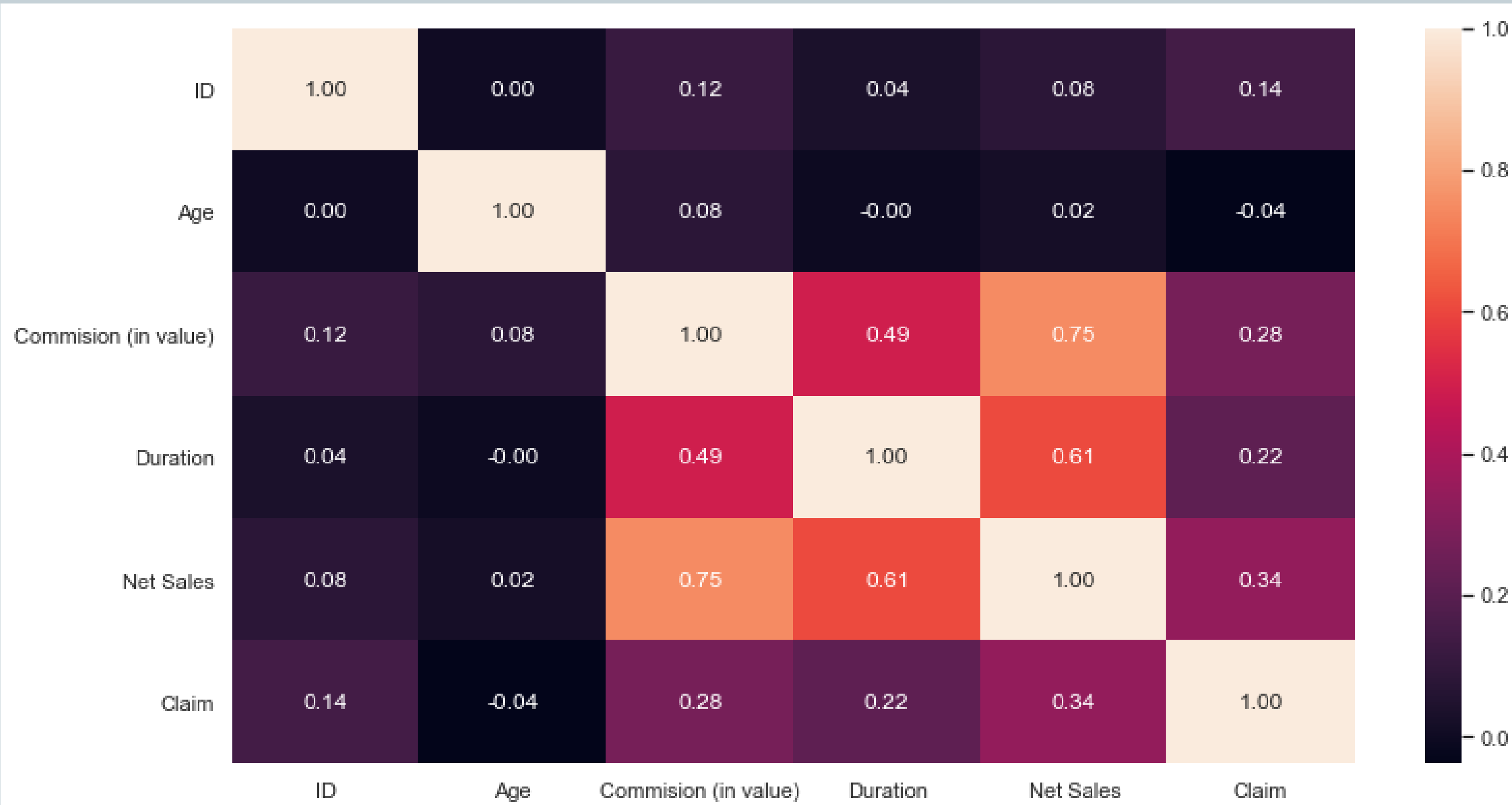
Test Statistics on Numerical values

	ID	Age	Commision (in value)	Duration	Net Sales
count	62288.000000	62288.000000	62288.000000	62288.000000	62288.000000
mean	32844.953458	39.666324	12.829703	60.958804	50.717064
std	18065.417216	14.014652	23.498745	114.325330	63.166715
min	0.000000	0.000000	0.000000	-2.000000	-389.000000
25%	17579.000000	33.000000	0.000000	10.000000	20.000000
50%	33446.500000	36.000000	1.880000	25.000000	29.700000
75%	48532.250000	43.000000	14.440000	59.000000	58.000000
max	63323.000000	118.000000	262.760000	4881.000000	682.000000

Negative values in duration which is not possible

Feature Selection

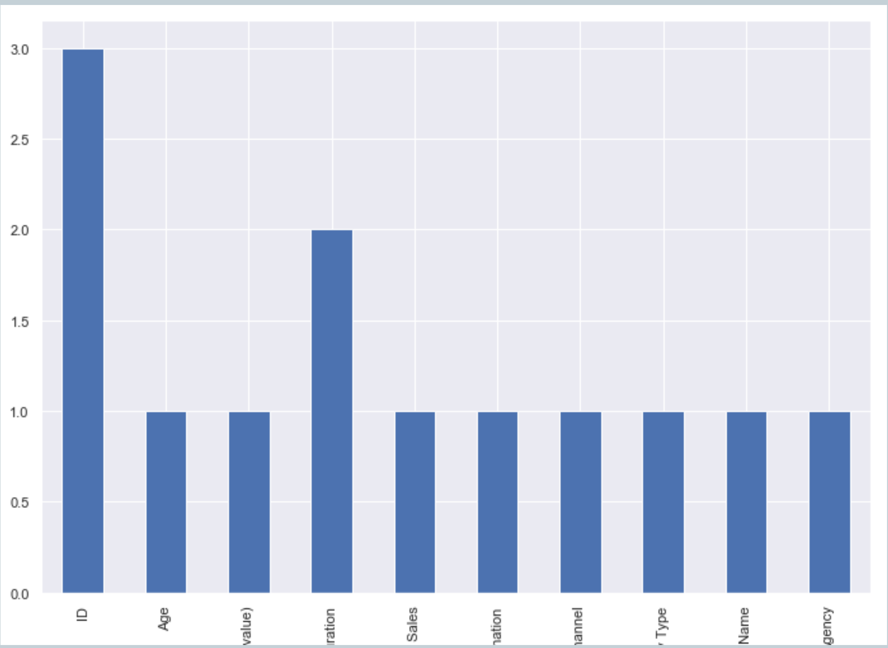
- Correlation



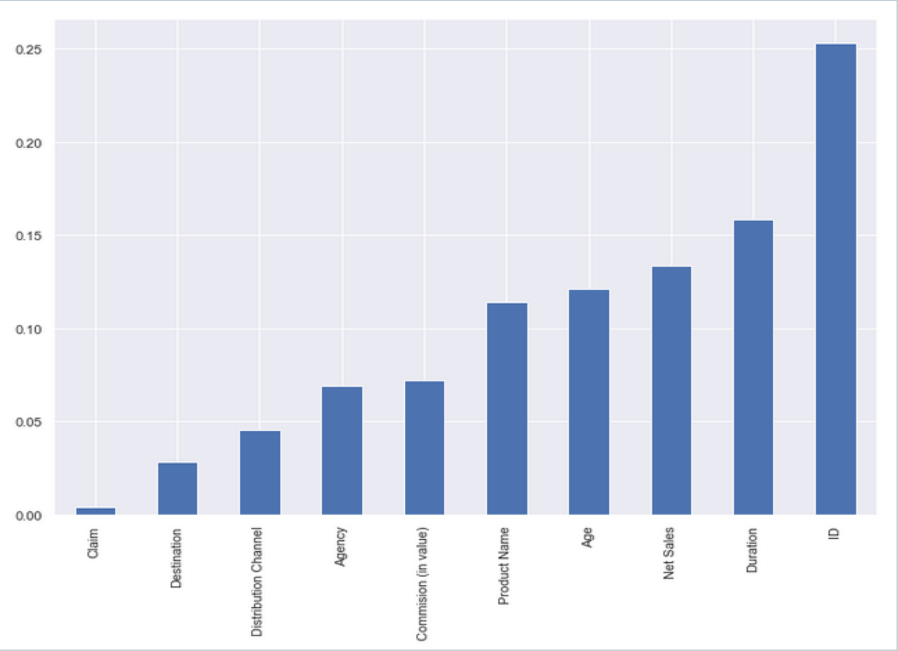
Recursive Feature Elimination

- Recursive feature elimination was performed using Random forest and Logistic Regression as the estimators. Below are the
- Important features obtained using both the methods

Logistic Regression:



Random Forest Classifier



In both methods, duration looks like important feature

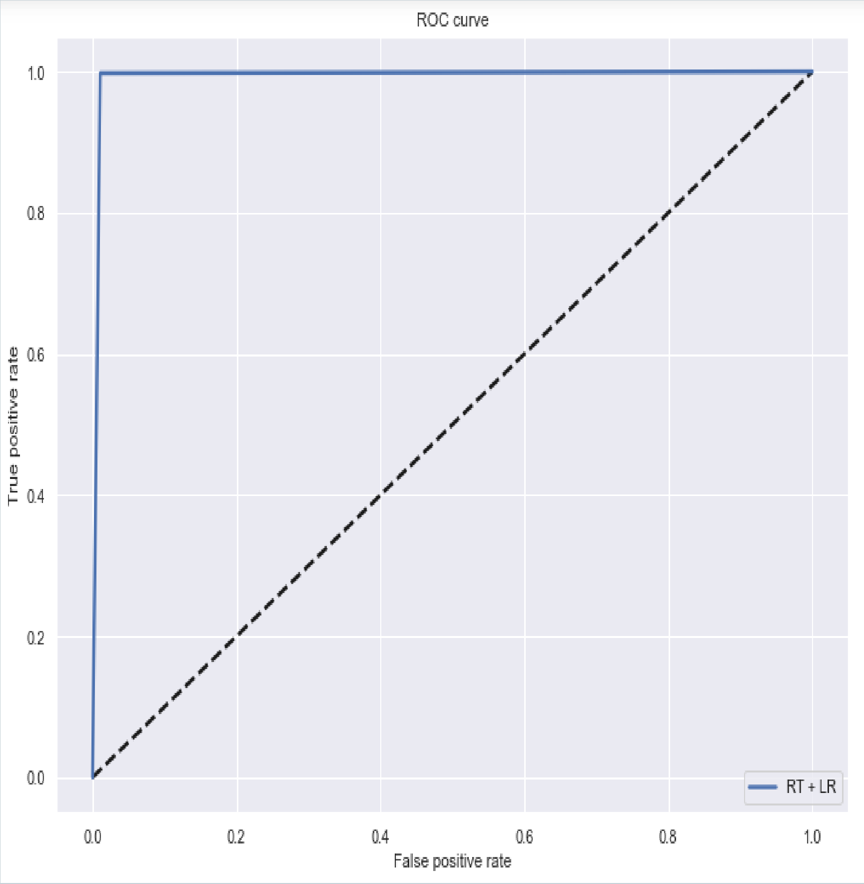
Models and Approaches

Below models were assessed without performing any hyper-parameter tuning and without treatment of class imbalance of the target. The models were
- Logistic Regression
- Random Forest Classifier

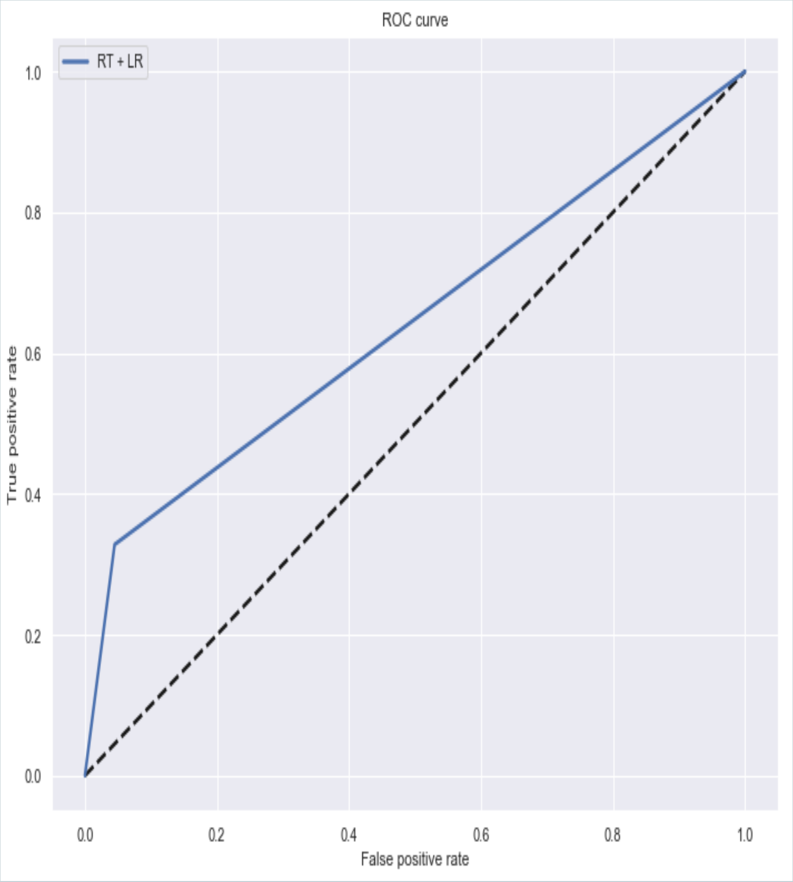
Modelling Method	Precision	Recall	F1-score
Logistic Regression	0 - 0.85 1 - 0.65	0 - 0.96 1 - 0.33	0 - 0.9 1 - 0.33
Random Forest Classifier	0 - 1.00 1 - 0.96	0 - 0.99 1 - 1	0 - 0.99 1 - 0.98

Evaluation and Results

Random Forest:



Logistic Regression:



Final Results

From the above observations and plottings it can be inferred that the best performing model was Random Forest Classifier and precision score is 99%

Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	14780	155
Actual Negative	14	3723