Will they claim it ?

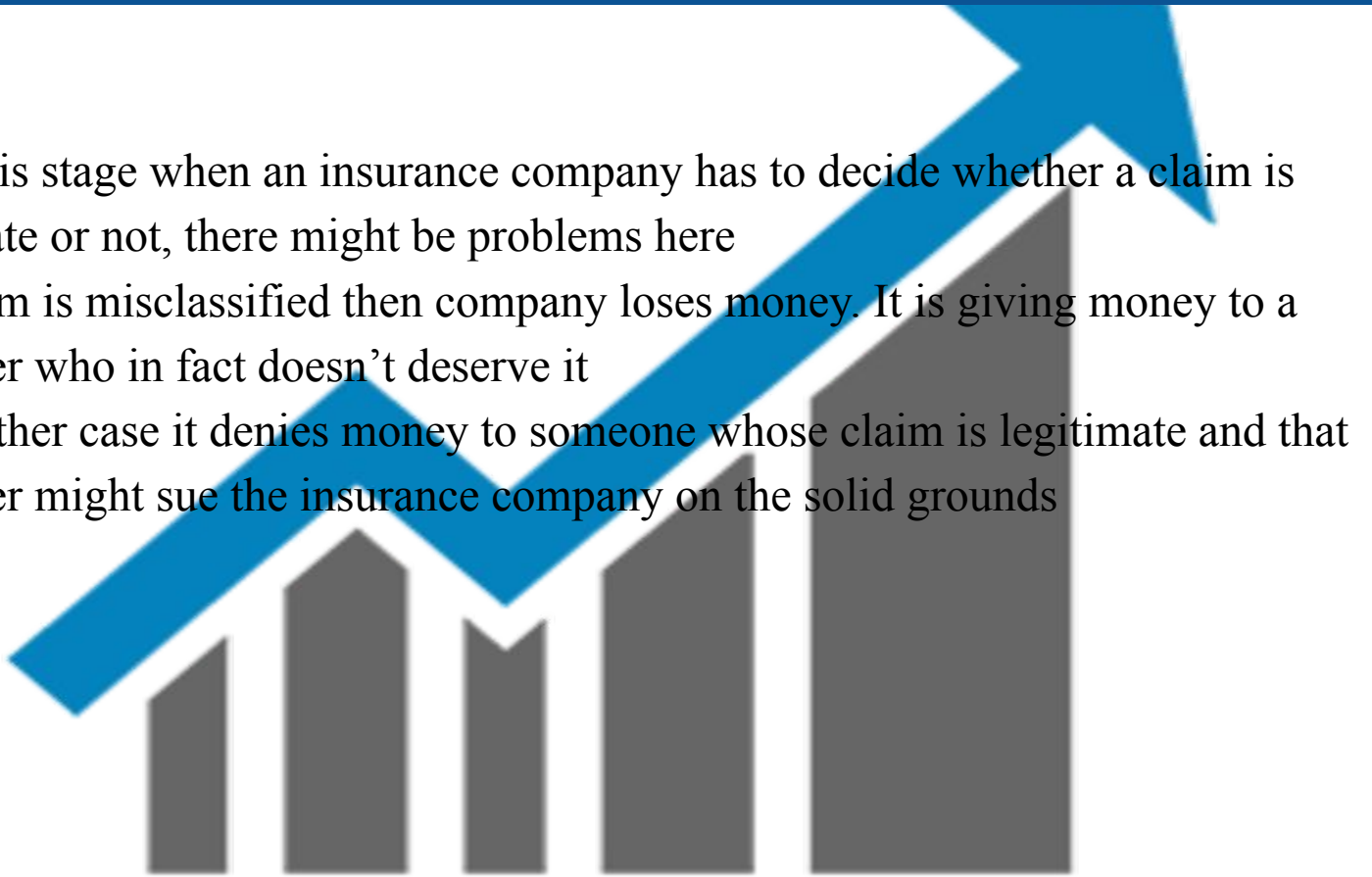
Problem Statement

- Many companies selling tickets or travel packages,give consumers the option to purchase travel insurance, also known as travelers insurance.

- Goal of this project is to predict whether insurance policies are claimed or not based on the given features.Travel insurance will have coverage for travelers concerns,including flight delays,trip cancellation or baggage loss.

- Insurance firms take risks on their customers. A very significant part of the insurance sector is risk management. To build profiles of high and low insurance risks, insurers consider almost all the quantifiable factors.

Other Related Problem



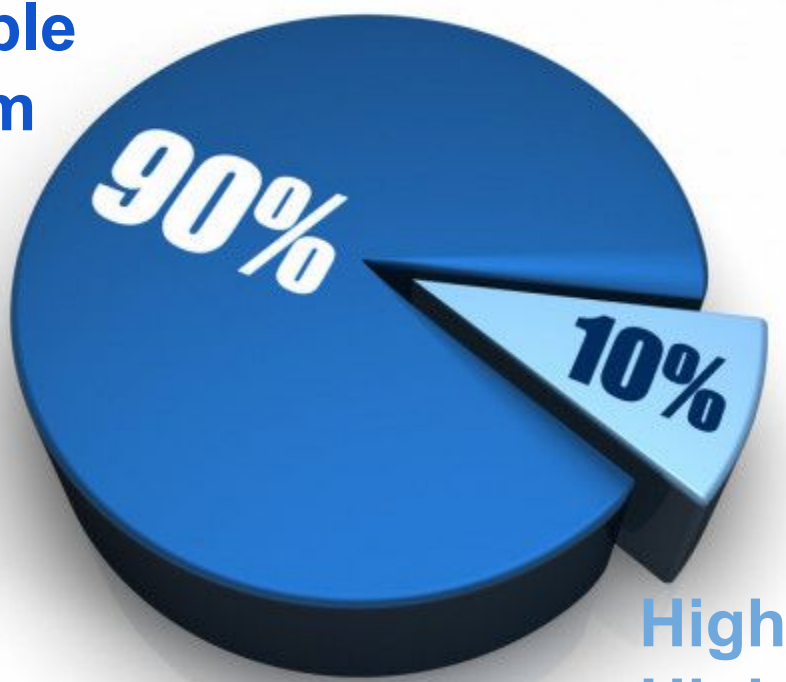
- After this stage when an insurance company has to decide whether a claim is legitimate or not, there might be problems here
- If a claim is misclassified then company loses money. It is giving money to a customer who in fact doesn't deserve it
- In the other case it denies money to someone whose claim is legitimate and that customer might sue the insurance company on the solid grounds

Who are Stakeholders? Who will benefit from the findings?

- Insurance Underwriter
- Head of claim Department
- Head of Marketing Department
- Head of Finance Department

Example

Low Risk
affordable
premium



High risk
High premium

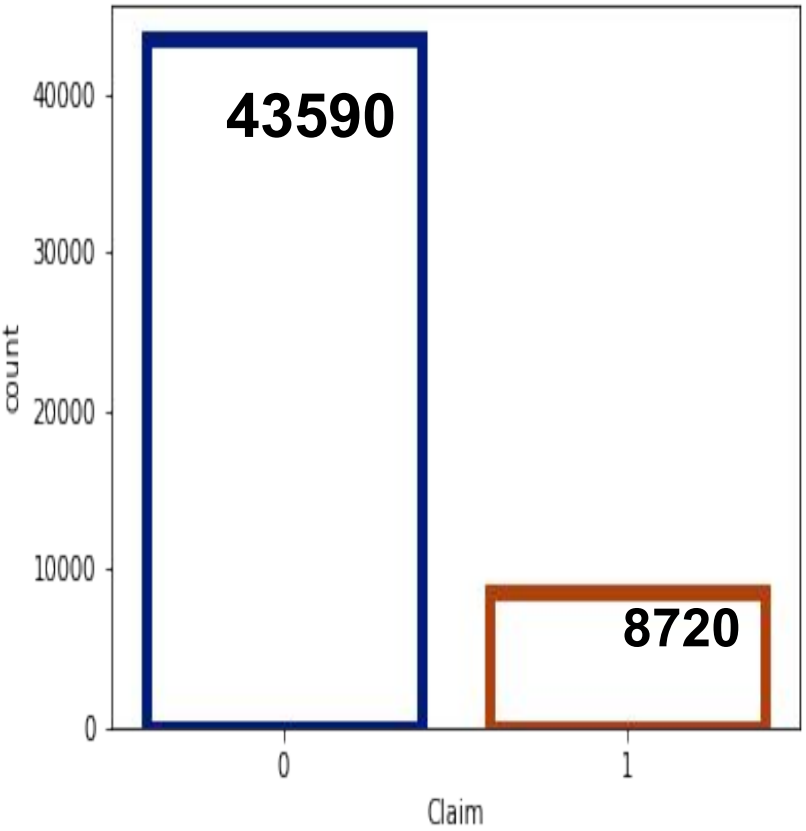
Exploratory Data Analysis

Dataset Description

The training dataset consists of data corresponding to 52310 customers and the test dataset consists of 22421 customers. Following are the features of the dataset

Feature Name	Feature Description	Feature Data Type
Agency	Name of agency	Categorical
Agency Type	Type of travel insurance agencies	Categorical
Distribution Channel	Distribution channel of travel insurance agencies	Categorical
Product Name	Name of the travel insurance products	Categorical
Duration	Duration of travel (In Days)	Continuous
Destination	Destination of travel	Categorical
Net Sales	Amount of sales of travel insurance policies	Continuous
Commission	The commission received for travel insurance agency	Continuous
Age	Age of insured	Continuous

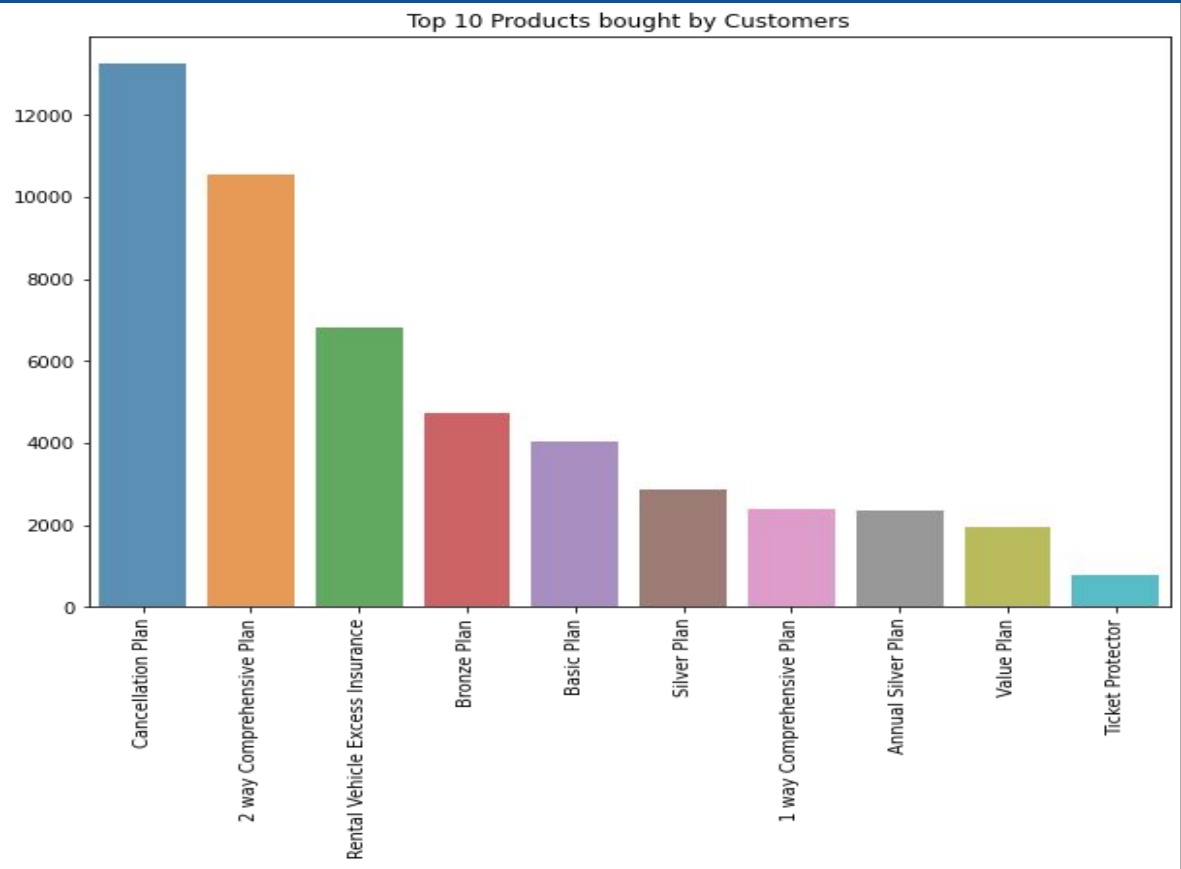
How many claims were made?



Observation

- Data is Imbalanced as the number of customers who have not made any claim is higher than the number of customers who have made claim.
- To make payment of claims more affordable for the company, we need to pool low risk claim. I.e customers who have low probability of claiming in the future.

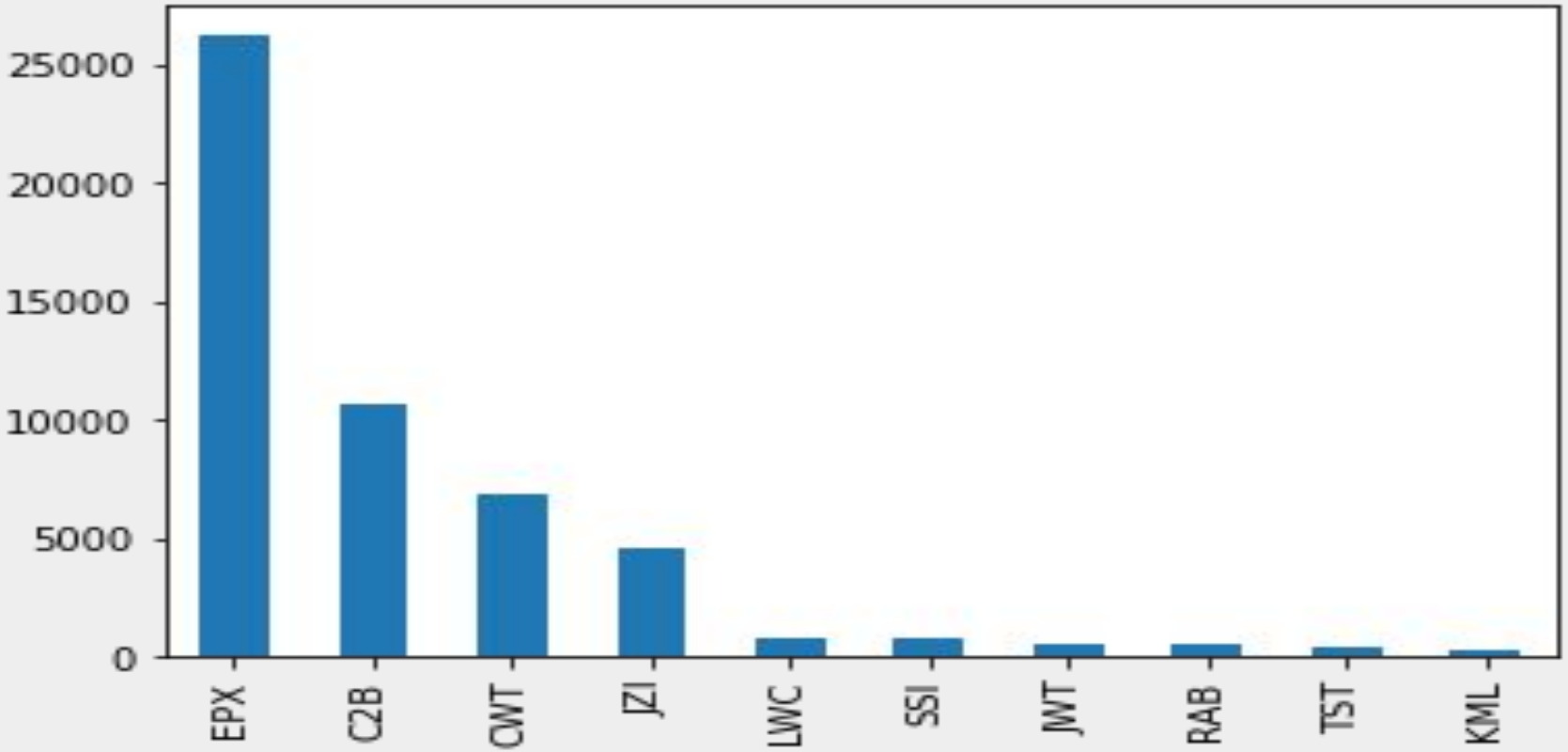
Which are the top 10 product bought by the customer?



Observation

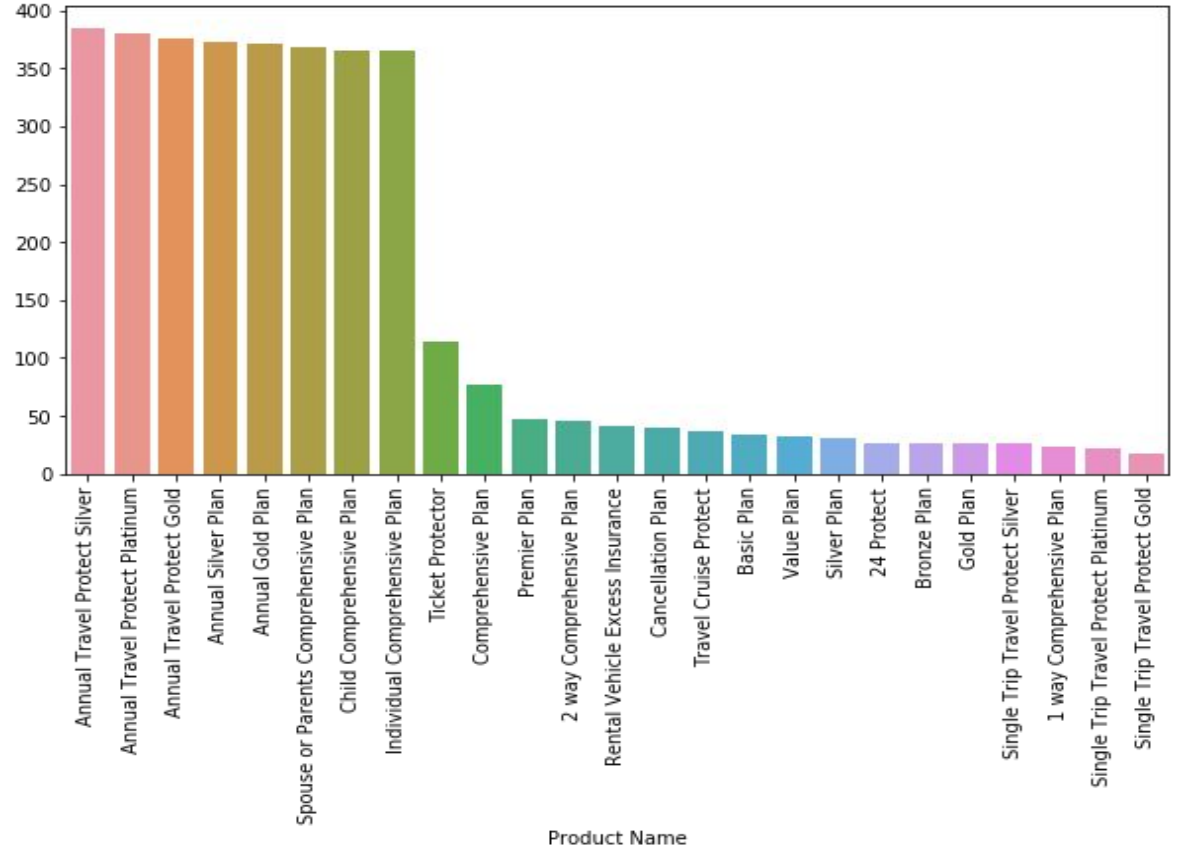
- Cancellation Plan
  13254
- 2 way Comprehensive Plan
  10555
- Rental Vehicle Excess Insurance
  6813

Which are the top agencies preferred by customers ?



EPX is the most preferred agency followed by C2B, CWT and JZI

Product bought by the customer according to Duration.

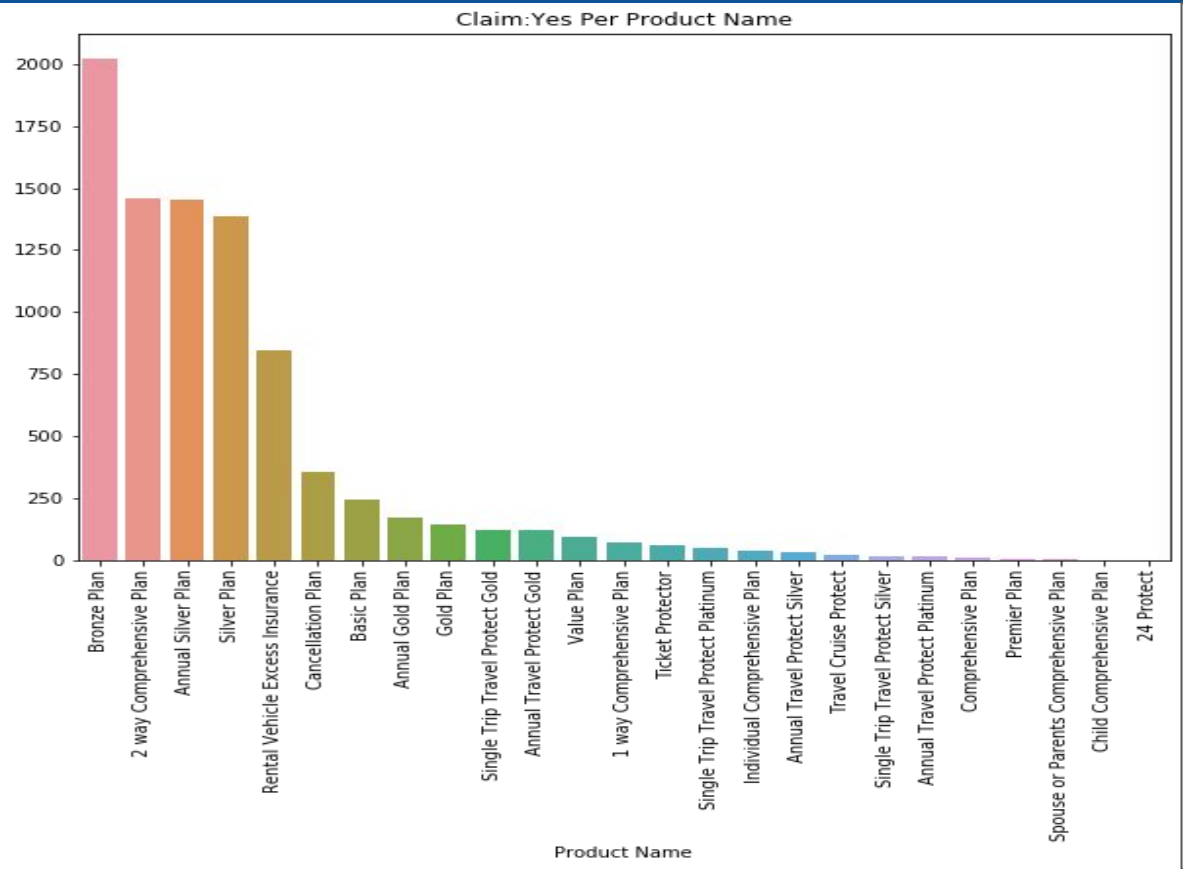


Observation

- Annual Travel Protect Silver 384.804
- Annual Travel Protect Platinum 380.26
- Annual Travel Protect Gold 375.29

According to Distribution top 8 Products are mostly uniformly distributed.

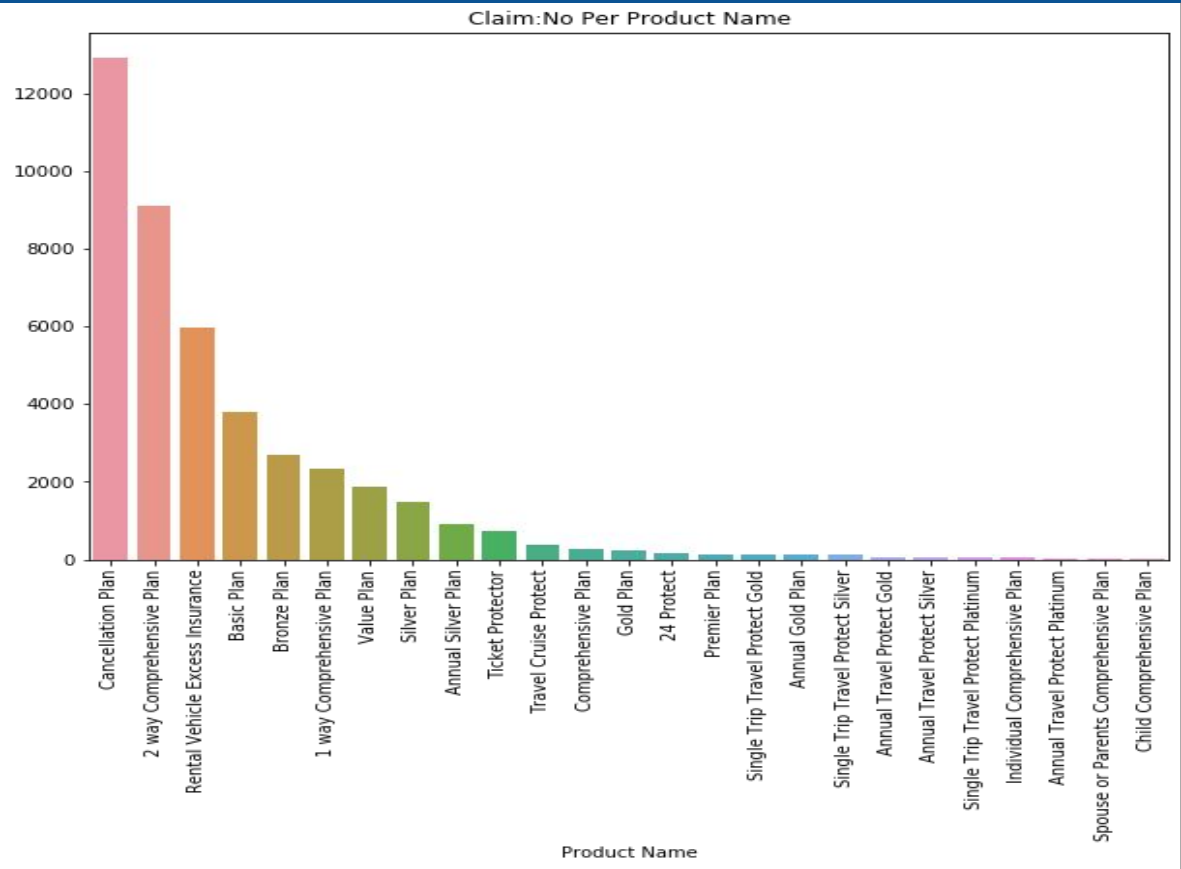
Which Product has the highest claims as Yes?



Observation

- Bronze Plan
  2020
- 2 way Comprehensive Plan
  1457
- Annual Silver Plan
  1451

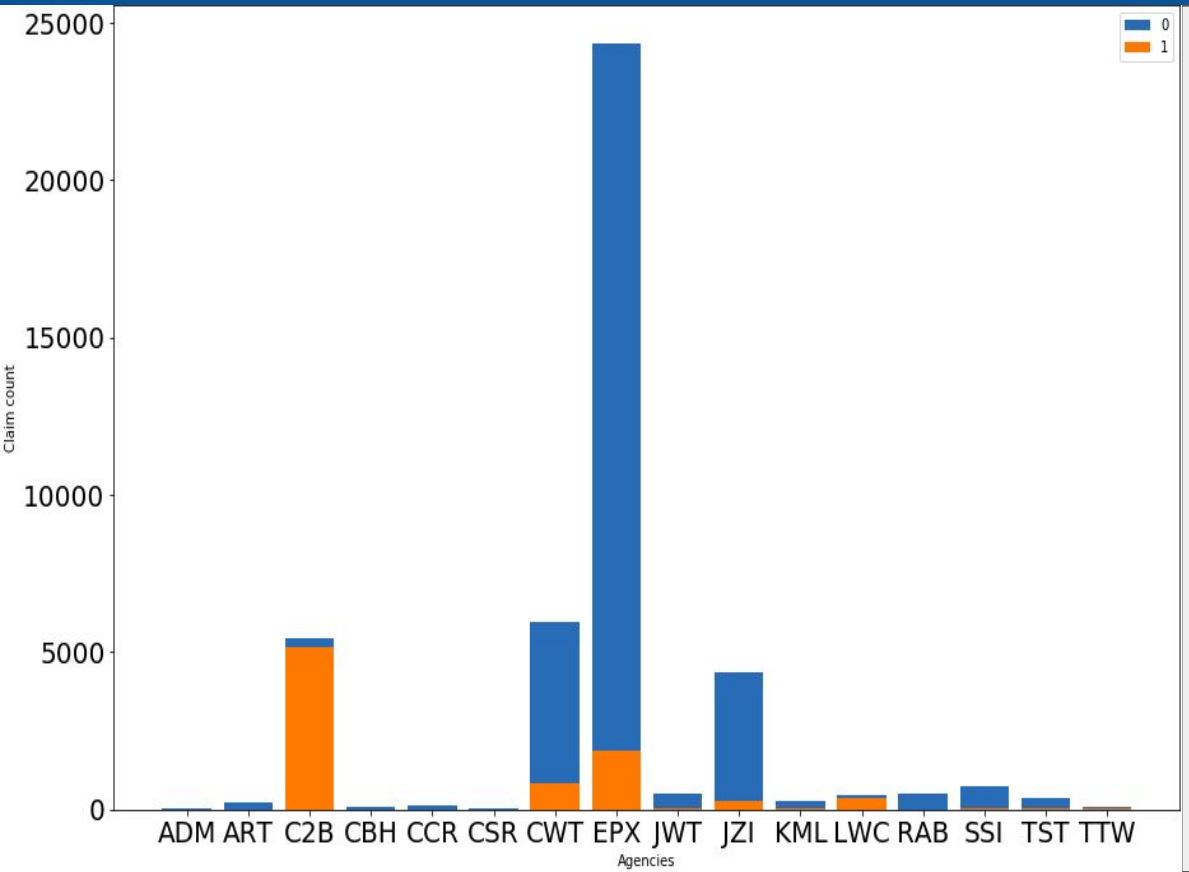
Which Product has the highest claims as No?



Observation

- Cancellation Plan
  12899
- 2 way Comprehensive Plan
  9098
- Rental Vehicle Excess Insurance
  5965

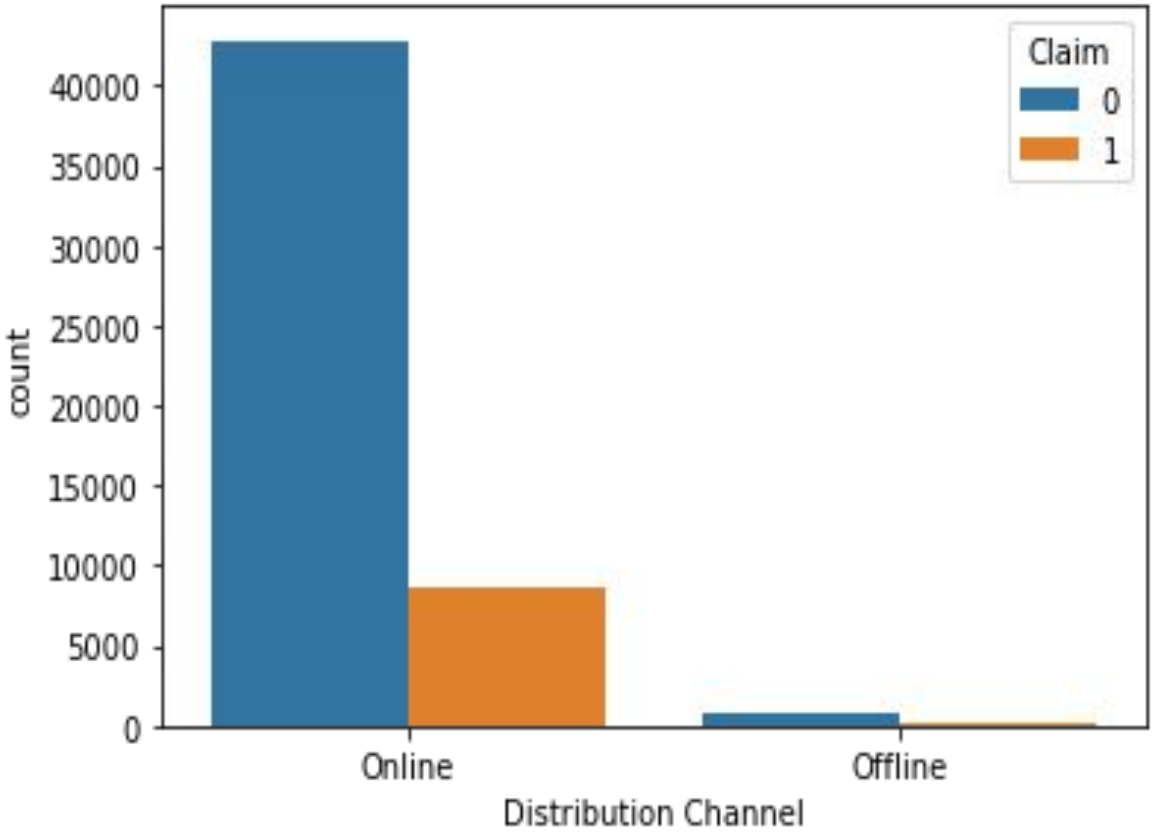
Agencies distribution according to Claim and no Claim.



Observation

- EPX agency has more biased Not Claimed data than Claimed data the same is with CWT and JZI.
- JWT, TST, ART, ADM, CBH,CSR, RAB and CCR has no data of Claimed.

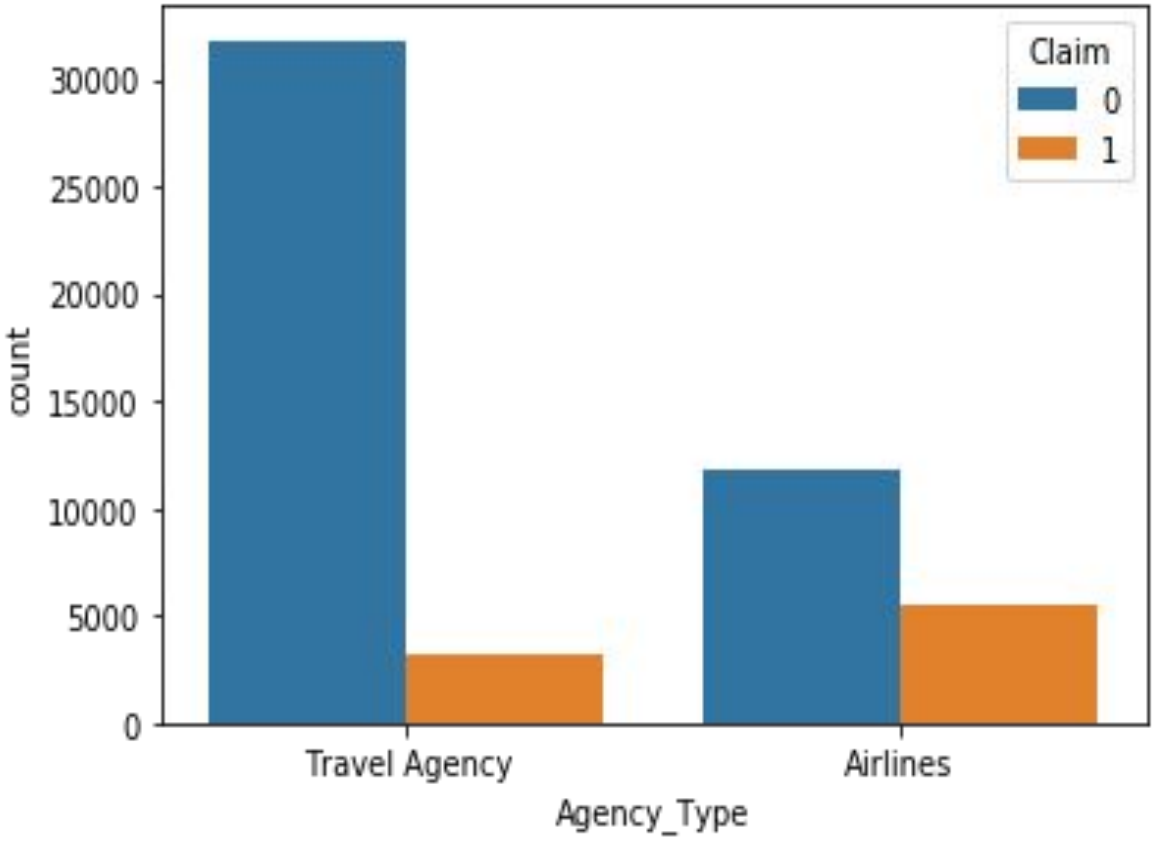
Distribution of Distribution Channel according to Claim.



Observation

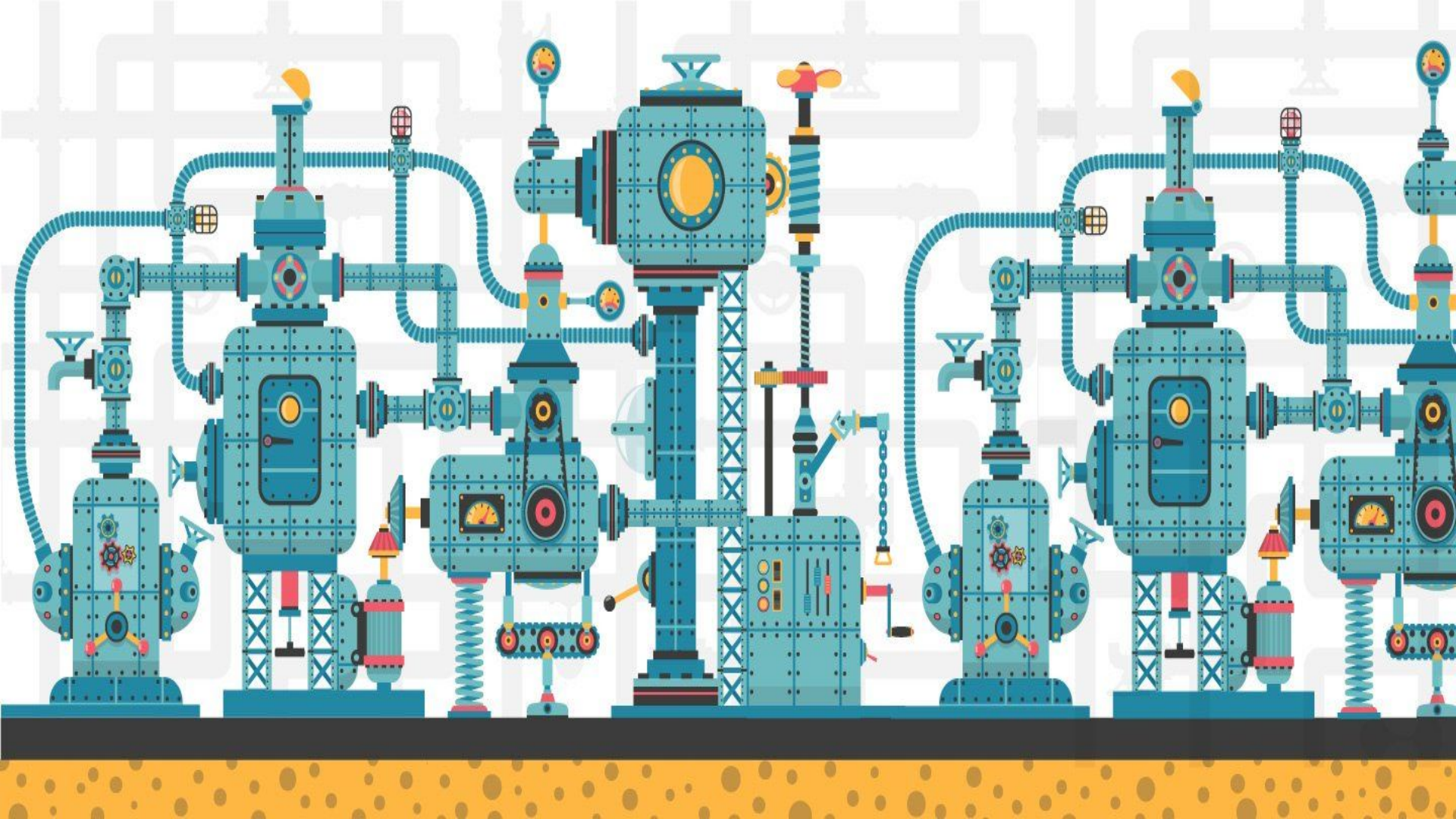
As it can be observed from the distribution that people mostly prefer Online distribution channel than offline. So company should try to make their online mode more attractive by getting feedback from the ones who are using Offline.

Distribution of Claim according to Agency type.



Observation

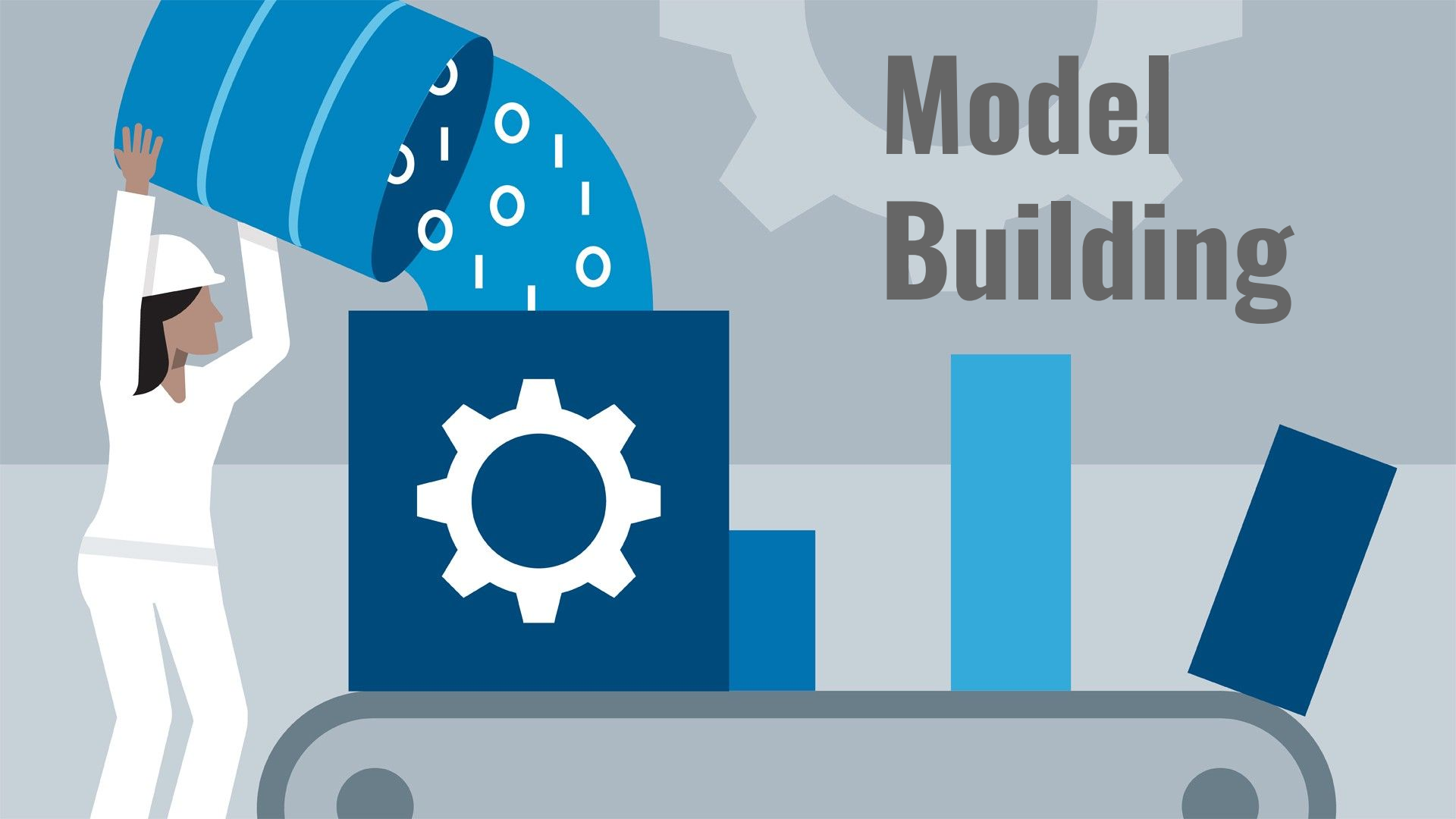
Population of people traveling through travel agency are more than that of airlines but we can see that instead of that insurance claim rate is more in airlines as compared to travel agency.



Pipeline

- Initially we have dataset with **11 features** and **52310 data points** in our training dataset and test data of around **22421 data points**.
- No missing Values, Pre-cleaned data
- As we do have **11 features** from which **Agency has 16 different categories** and **Destination has 97 different categories** so we created 2 new KPIs .
- **Sorted Agency** according to net sales. Agency with highest net sales at the top followed by others.
- **C2B,EPX,CWT,JZI & LWC** these are the major players. Rest belongs to **others**
- **Countries** clustered into **Continents**.
- There are some transcontinental countries.
- In age distribution we found that there were **2 people with age 0**.Strange!..
- 737 people with **age more than 100** ------> **mean** of the **Age column**.
- Replaced **-ve duration** ------> **0**
- Standard scaled the numeric data
- **Labeled encoding** and **One hot encoding(Product Names)** for categorical data

Model Building



Models and Approaches

We tried 6 different models:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Bagging Classifier
5. Gradient Boosting Classifier
6. XGBoosting Classifier

We split the data into training and testing data with testing data size being 30%

We used Cross Validations = 3 for almost all the hyperparameter tuning parts

Major Hyperparameters used for tuning: n_estimators, max_depth, min_leaf_sample, learning_rate, scale_pos_weight, criterion, class_weight

Models and the scores (No data Balancing):

We first trained the **untuned models** on the training data and then we trained the models by **hyperparameter tuning.** Following are the scores we got for different models we trained the data on:-

Models	Untuned			Tuned		
	Precision Score	Accuracy	F1 score	Precision score	Accuracy	F1 score
Logistic Regression	0.61	0.85	0.36	0.61	0.85	0.37
Decision Tree	0.72	0.91	0.73	0.76	0.92	0.76
Random Forest	0.81	0.93	0.77	0.81	0.93	0.78
Bagging Classifier	0.82	0.93	0.76	0.81	0.93	0.77
GBM Classifier	0.64	0.86	0.5	1	0.83	0.017
XGB Classifier	0.81	0.92	0.75	0.98	0.86	0.29

Models and Scores (Balanced Data):

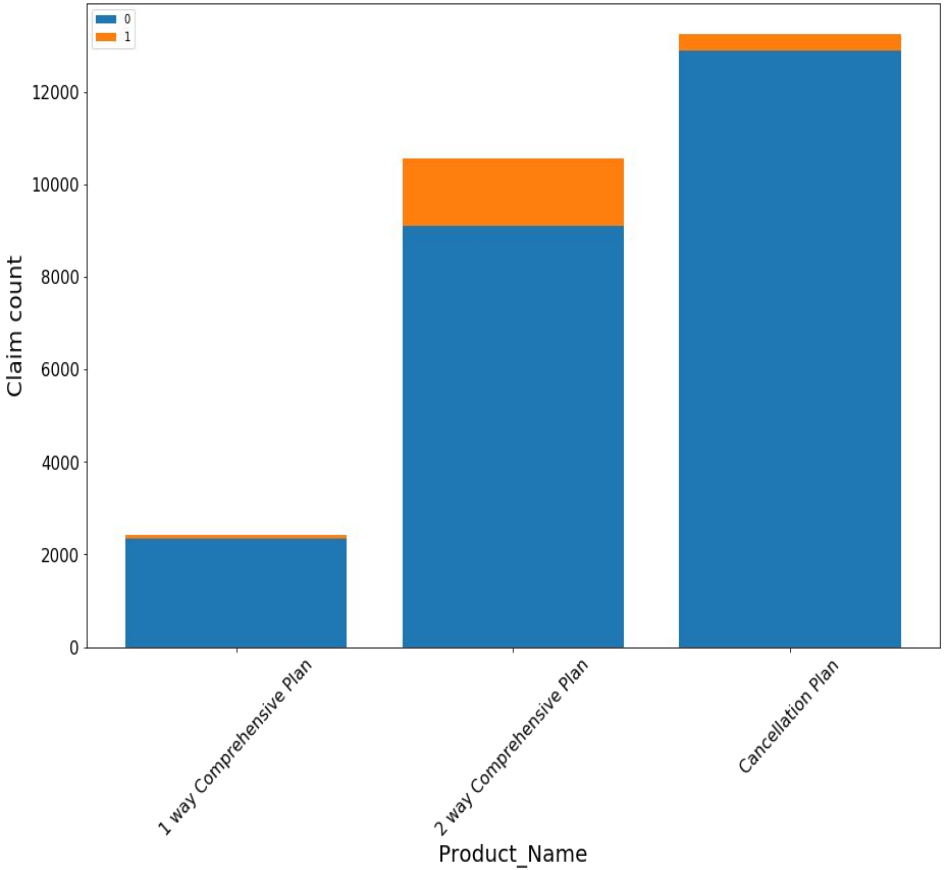
Here we tried balancing the data using three different techniques as **Random Under Sampling**, **Random Over Sampling** and **SMOTE** and here are the scores we achieved:

Models	Random Under Sampling			Random Over Sampling			SMOTE		
	Precision score	Accuracy	F1 score	Precision score	Accuracy	F1 score	Precision score	Accuracy	F1 score
Random Forest	0.89	0.91	0.91	0.95	0.97	0.97	0.95	0.95	0.95
GBM Classifier	0.9	0.91	0.91	0.96	0.98	0.98	0.96	0.96	0.96
XGB Classifier	0.91	0.91	0.91	0.96	0.98	0.98	0.96	0.97	0.97

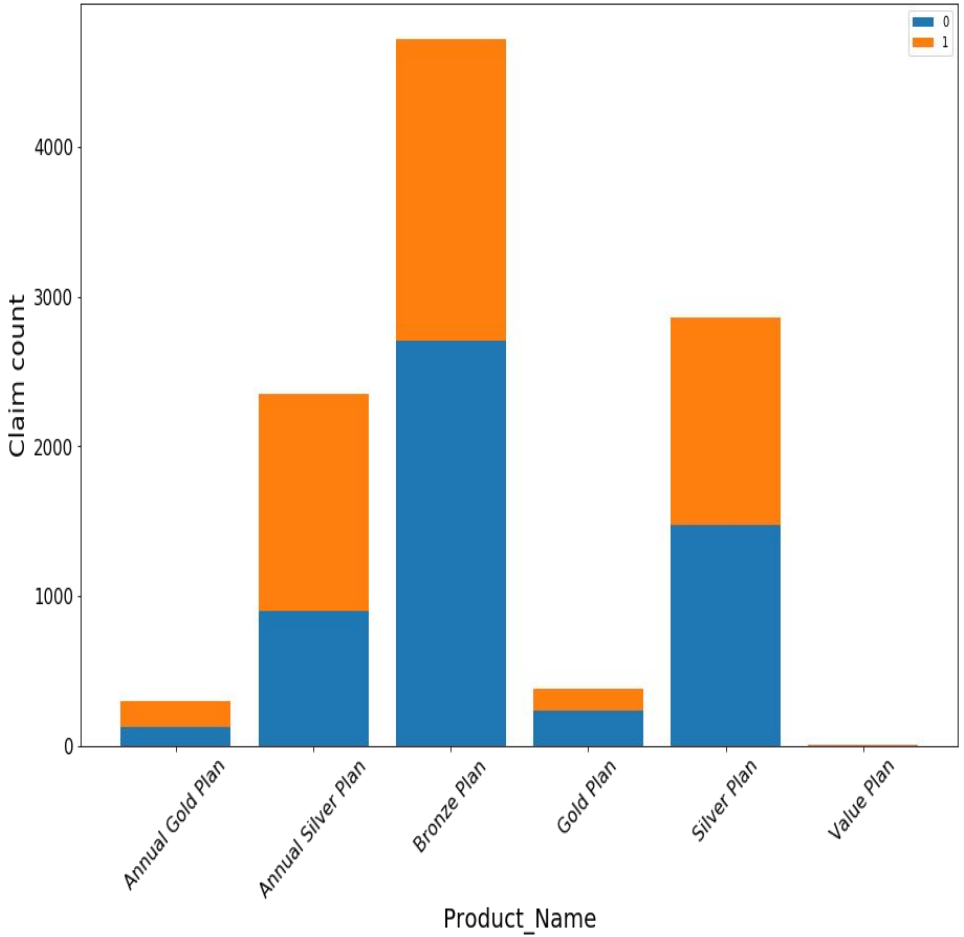
Final Results:

- **Gradient Boosting Classifier without any data balancing** and **n_estimators=80, learning_rate=0.01, max_depth=5**
- Problem - F1 score is **0.17**
- Gradient Boosting is a greedy algorithm, it overfits the data quickly and that might be the case here.
- After data balancing and XGBoost is our best bet. This is the model which supposedly going to have high F1 score and has **precision score around 0.95** after data balancing by Random Oversampling.

EPX


0
1
12000
10000
8000
6000
4000
2000
0
Claim count
1 way Comprehensive Plan
2 way Comprehensive Plan
Cancellation Plan
Product_Name

C2B


0
1
4000
3000
2000
1000
0
Claim count
Annual Gold Plan
Annual Silver Plan
Bronze Plan
Gold Plan
Silver Plan
Value Plan
Product_Name

Insights & Decisions

- The largest share of net sales for travel agency shows that they have a huge customer base when compared to airline agencies but still their average sales is lower than that of airline agency. So they should focus more on acquiring customers who opt for airline insurance.

- For both airline and travel agency only few of the products generates maximum of their revenue so either they can drop the low performing products or they need to revisit their product portfolio to do some needful changes so that it can attract more customers.

Insights & Decisions

- In spite of being the ruler of the market in terms of number of customers, EPX has 2nd best Net sales, from this we might conclude that the premium offered by EPX might be lesser as compared to the C2B which has the highest Net Sales but lesser number of customers as compared to EPX

- From the graphs of ‘product names vs claim counts’ we can observe that the major product plans of EPX have lesser claims to non claims cases but, in the case of C2B the number of claims to non claims cases is higher


Thank you!