Will they claim it?

Greyatom Hackathon
by Wanderlust Data

Problem Overview

For an insurance company, the forecasting of claims is central to a successful operation. If the claims can be forecasted accurately, **premiums** can be adjusted accordingly, creating the opportunity to be one step ahead of the competitors. Charging a lower premium than the **competitors**, while maintaining a sufficient buffer to make profit to stay in business, will lead to more customers, which in turn leads to more **profit**.

Who are the Stakeholders? Who will benefit from the findings?

Head of Marketing Department

Head of Finance Department

Head of Claims Department

Processing Department

Insurance Underwriter

Data Science Problem

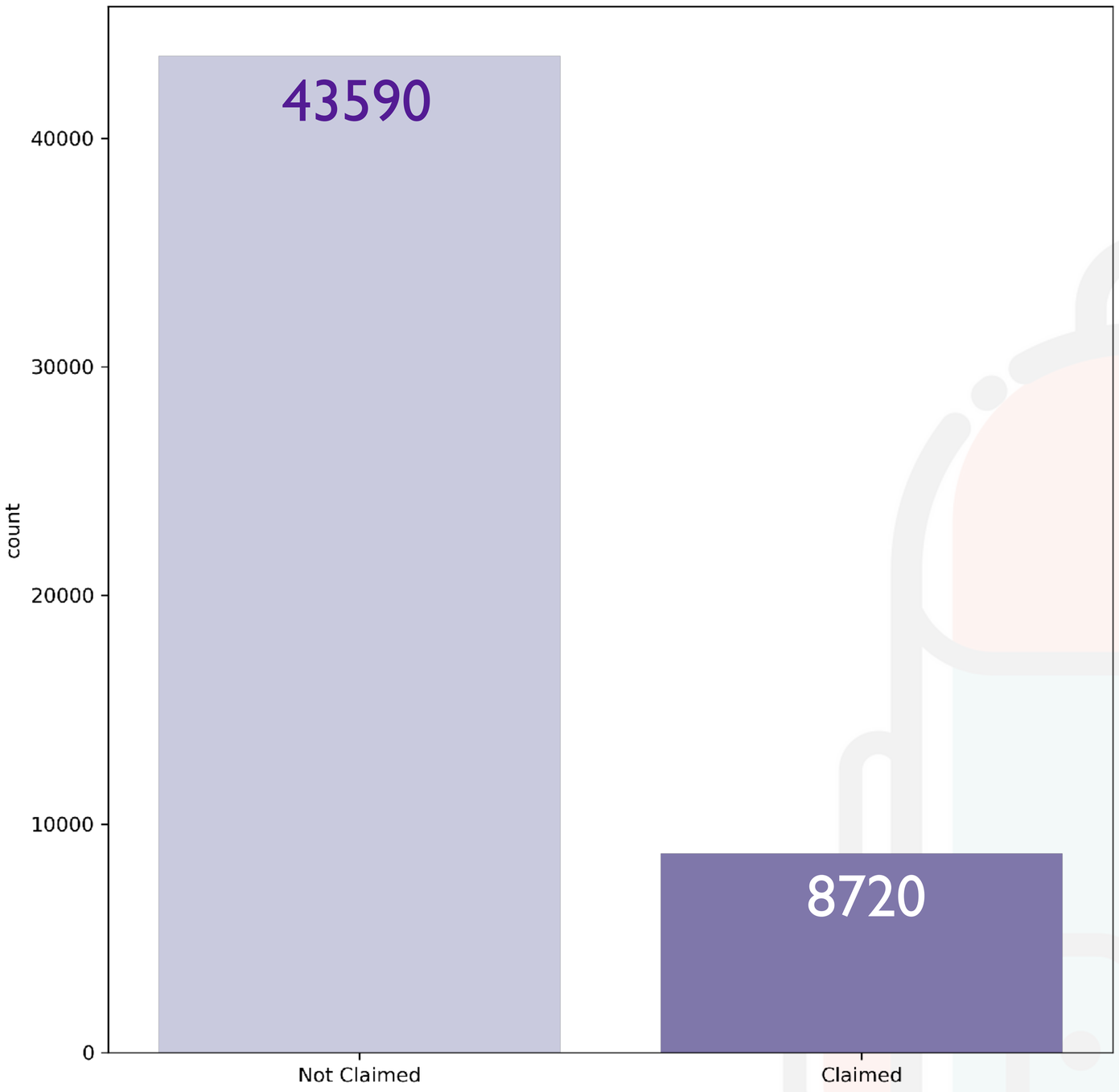
To predict if a new buyer will make the claim or not based on the given data.

Business Metric

To correctly identify high and low risk customers by looking at the claim pattern, such that we increase the customer onboarding rate by 25%

Exploratory Data Analysis

How many claims were made?



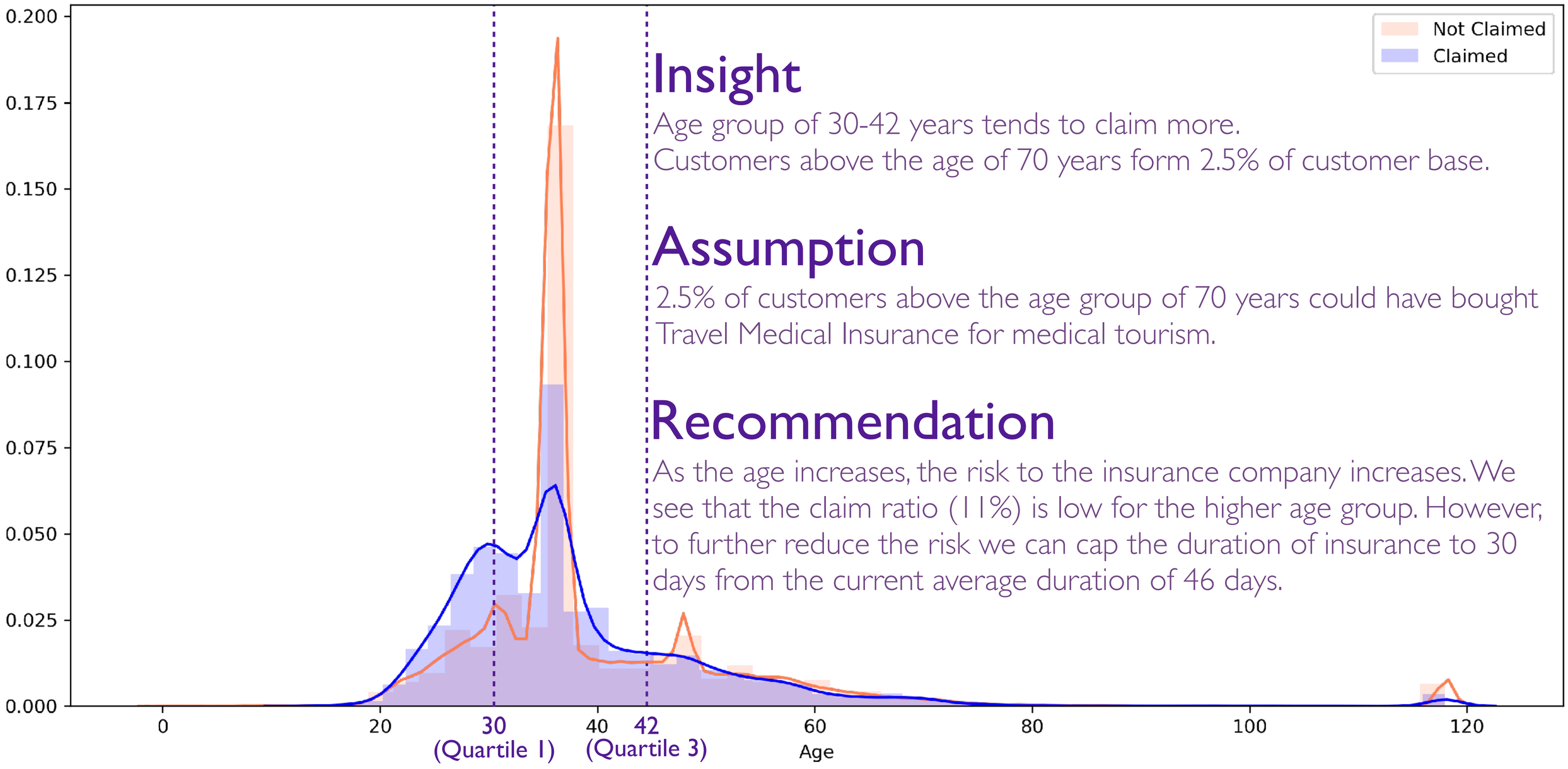
Observation

Data is Imbalanced as the number of customers who have not made any claim is **5 times** the number of customers who have made the claim.

Recommendation

To make payment of claims more affordable for the company, we need to **pool low risk clients** i.e. customers who have low probability of claiming in the future.

Does age effect the Claim likelihood?



Insight

Age group of 30-42 years tends to claim more.
Customers above the age of 70 years form 2.5% of customer base.

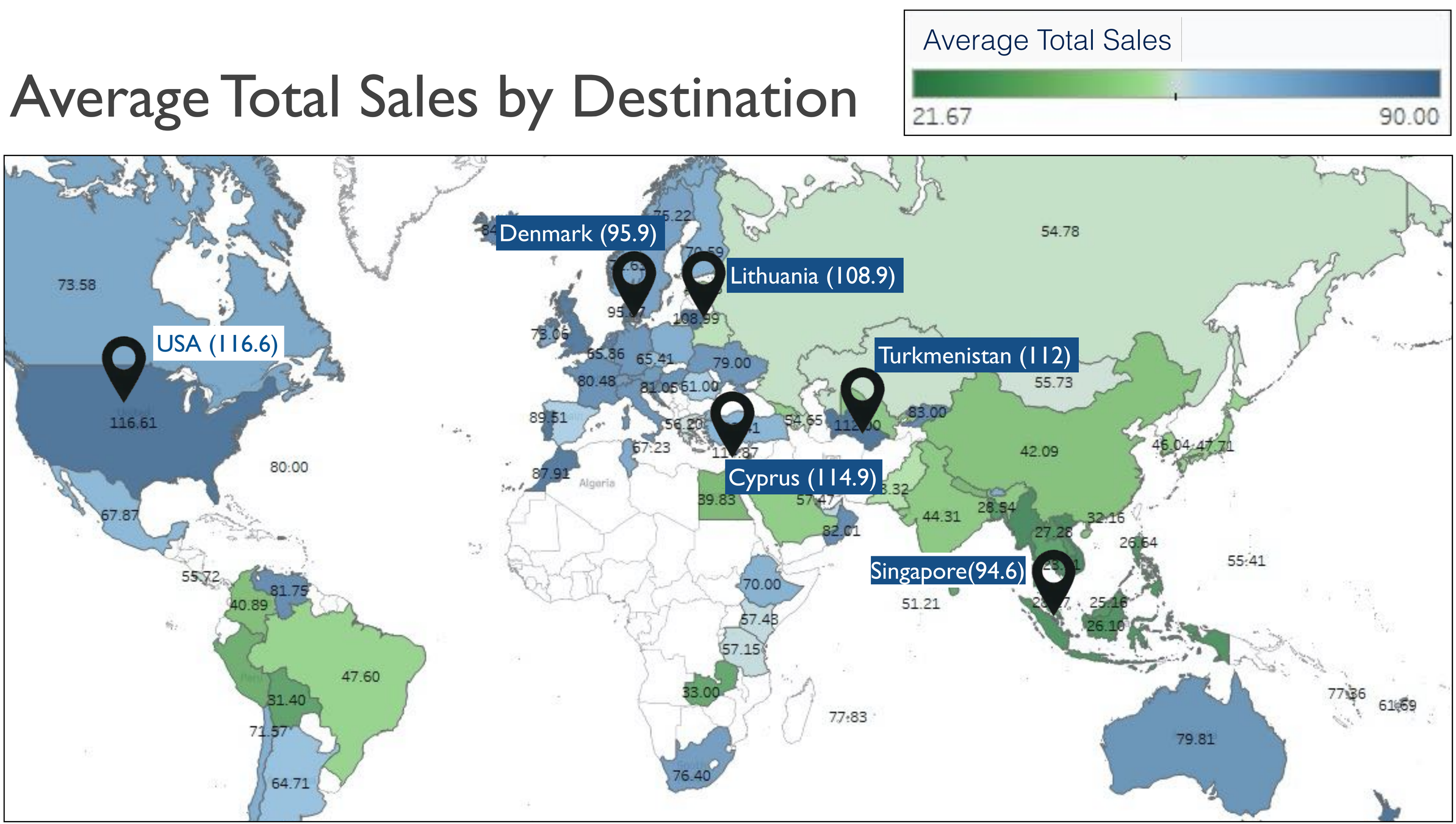
Assumption

2.5% of customers above the age group of 70 years could have bought Travel Medical Insurance for medical tourism.

Recommendation

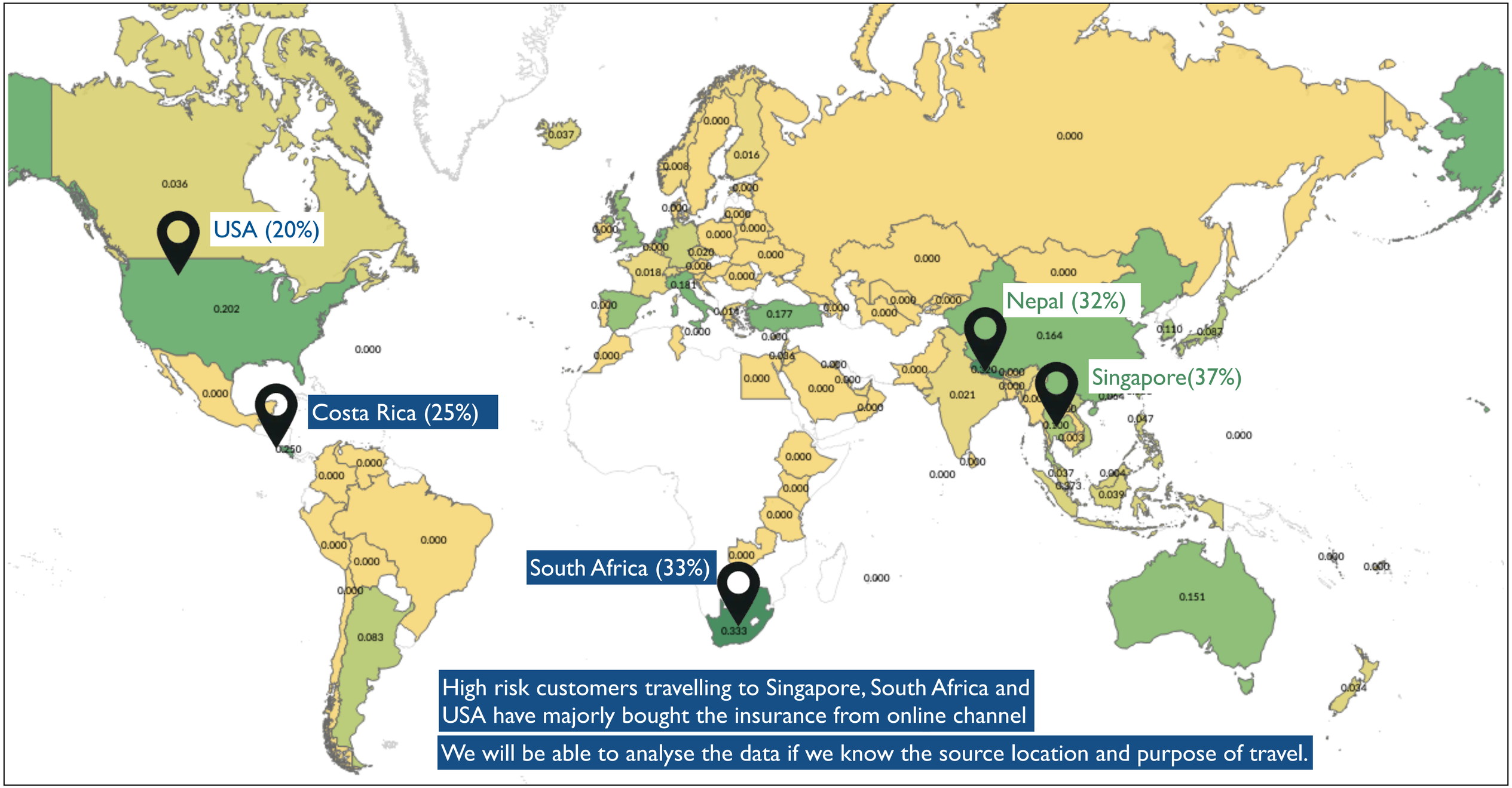
As the age increases, the risk to the insurance company increases. We see that the claim ratio (11%) is low for the higher age group. However, to further reduce the risk we can cap the duration of insurance to 30 days from the current average duration of 46 days.

Average Total Sales by Destination

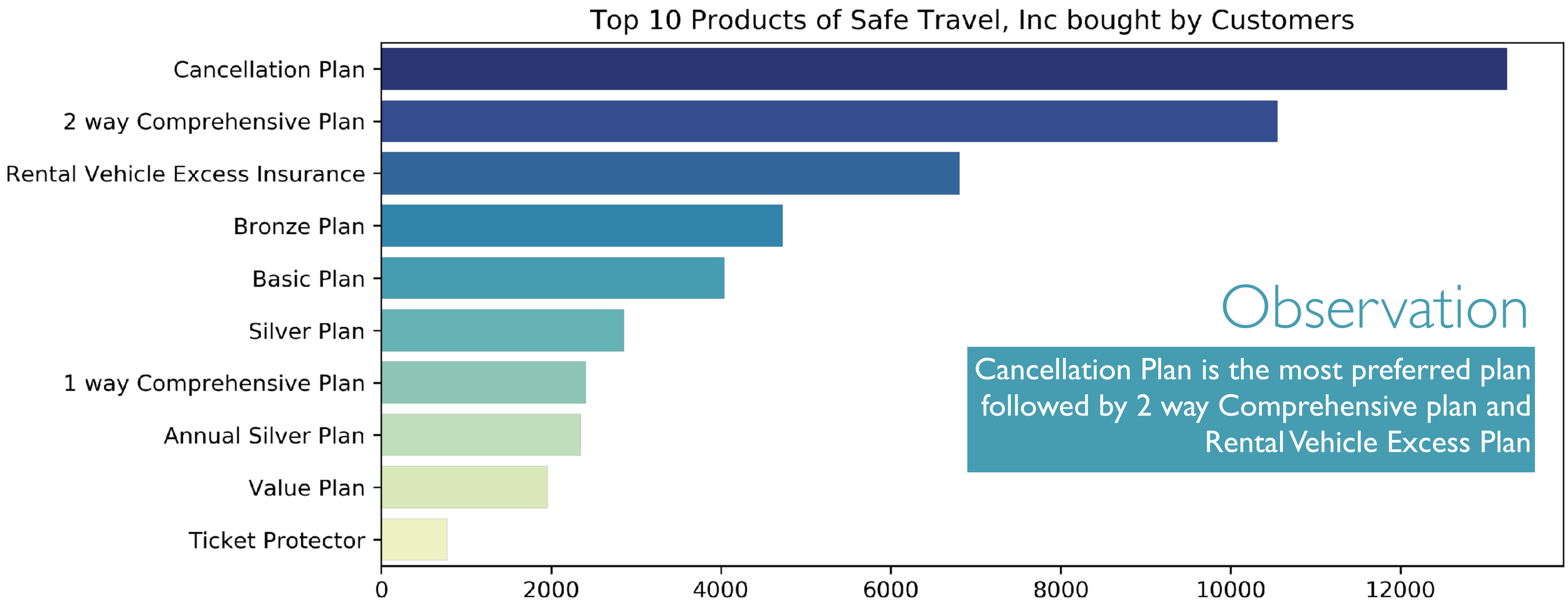


*Total Sales = Net Sales + Commission

Claim Rates by Destination



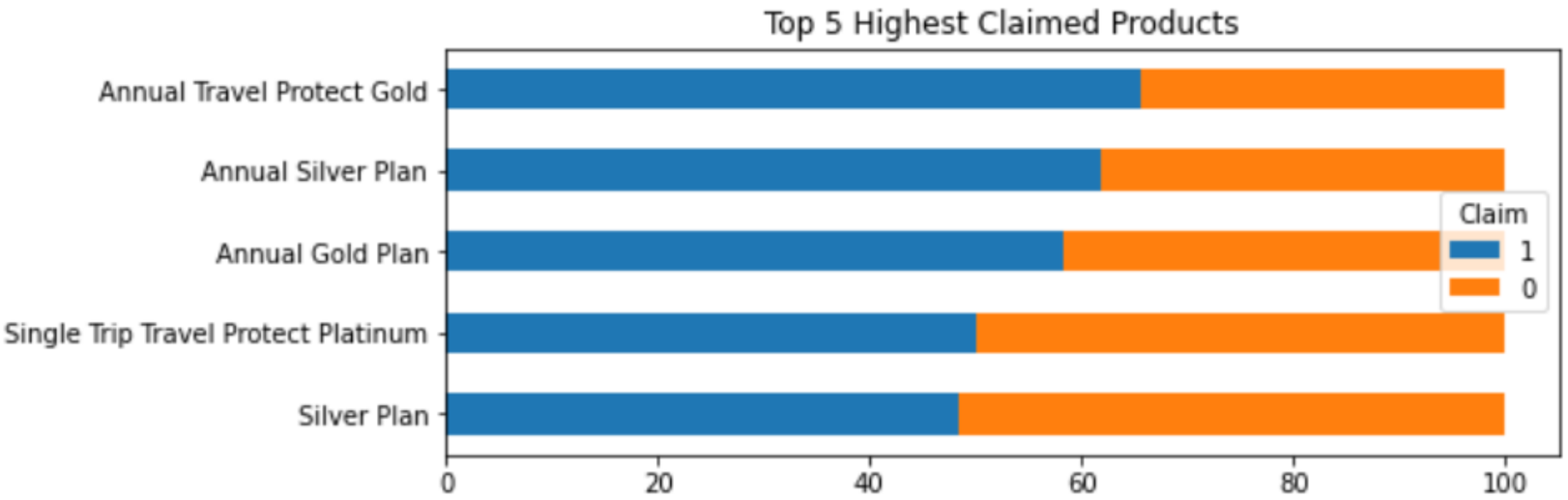
Which are the top 10 products of Safe Travel, Inc bought by the customer?



Observation

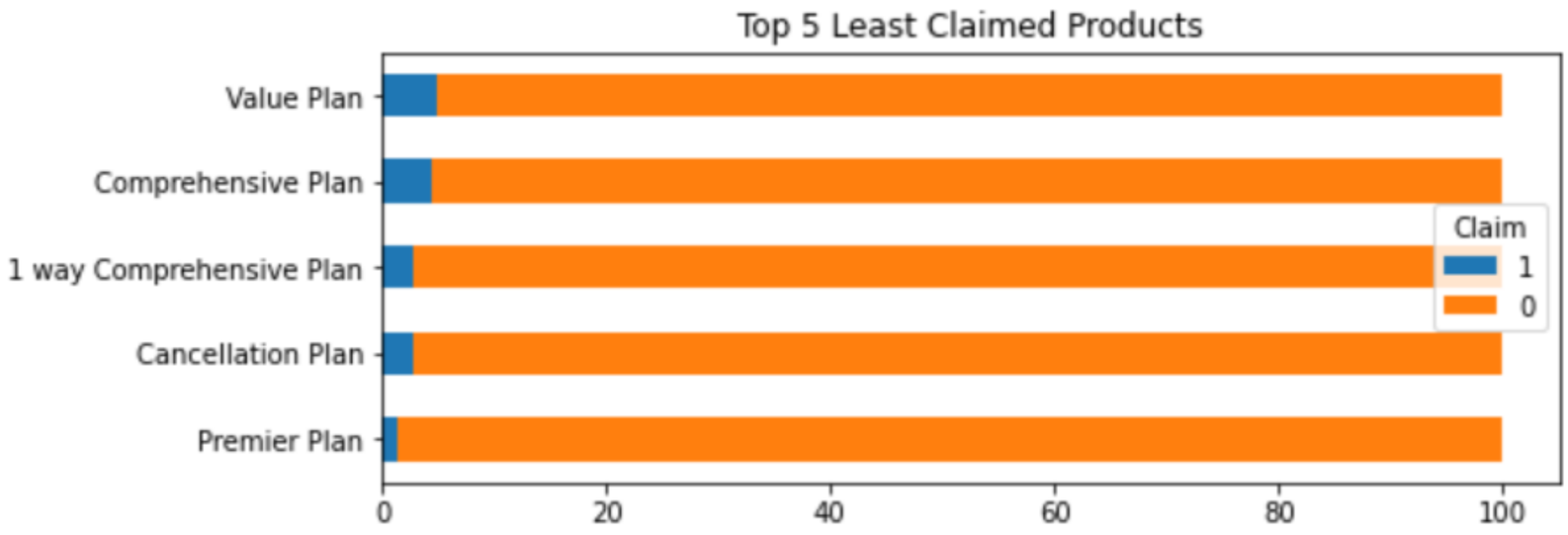
Cancellation Plan is the most preferred plan followed by 2 way Comprehensive plan and Rental Vehicle Excess Plan

Which products have received highest and lowest claims?



Observation

Annual Travel Protect Gold, Annual Silver Plan and Annual Gold Plan have less buyers and highest claims.
The above plans have claims in the range of 48-66%



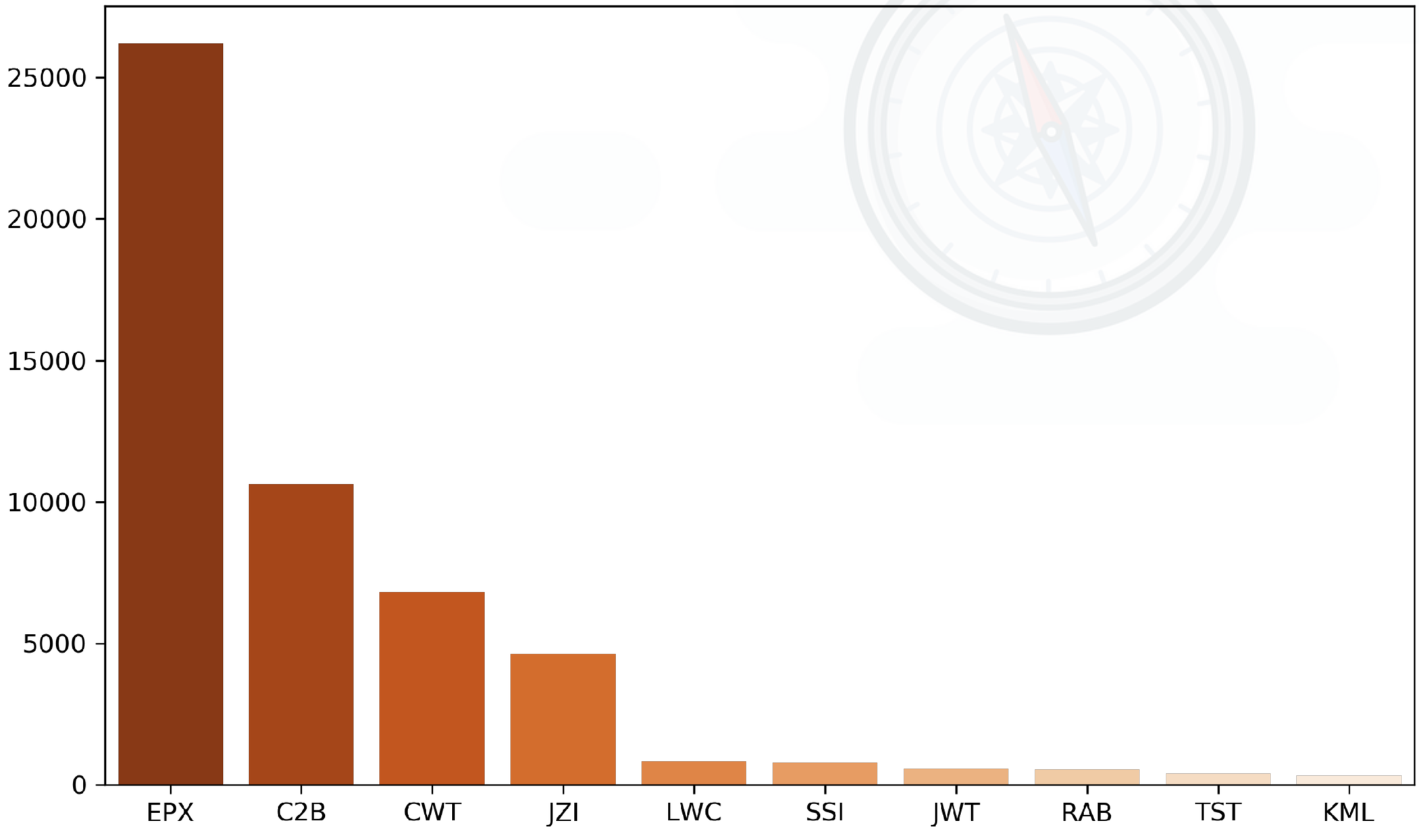
Observation

- Cancellation Plan, 1 Way Comprehensive Plan, Value Plan are amongst the top 10 most preferred products
- The above plans have claims in the range of 1.5-5%

What is their Claim Pattern?

	Top Claimed Products	Least Claimed Products
Average Age of a customer who make Claims	37 years	45 years
Destinations most visited by people to make Claims	Singapore(94%)	Thailand(25%), USA(11%), China(10%)
Where did they buy Insurance from?	From Airlines (94%) through Online channel (100%)	From Airlines (91%) through Online channel (93%)
Avg. Net Sales and Commission charged	Net Sales: 167.53 Commission: 46.23	Net Sales: 33.08 Commission: 3.8
Median Duration of Travel	364 Days	28 Days

Which are the top agencies preferred by the customers of Safe Travel?

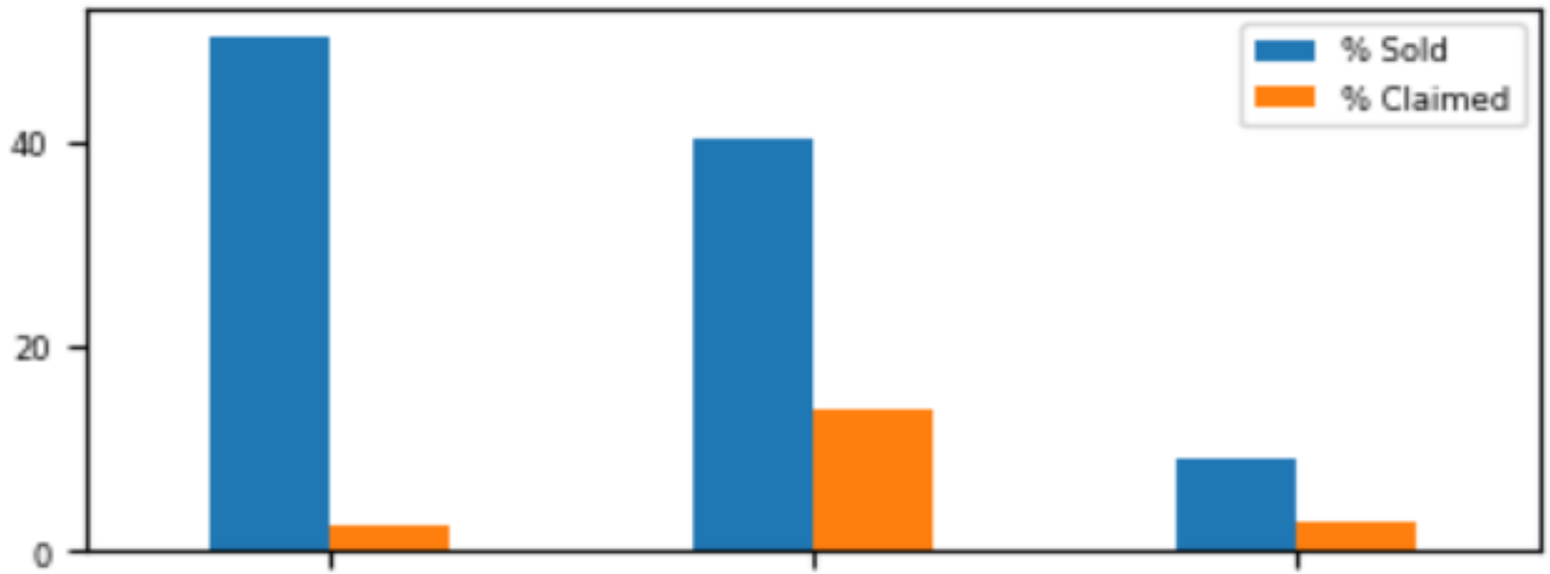


EPX | Which are the products sold by them? What is their claim rate?

EPX is a Travel Agency which is into low risk products with the highest claim rate being 13.8% for 2 way comprehensive plan.

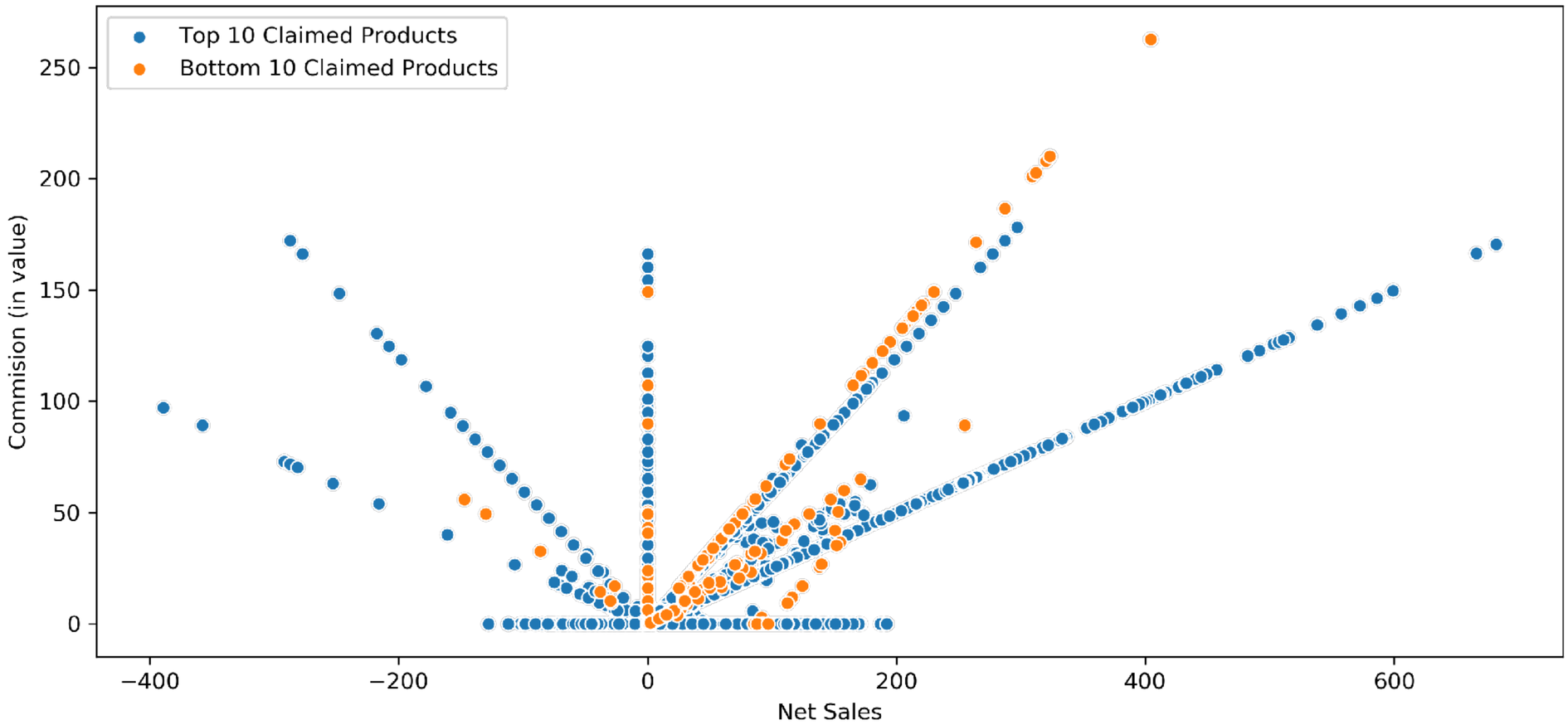
EPX is preferred for travel to China, Thailand and US amongst the higher age range.

1 way comprehensive plan preferred by the older age group.



	Cancellation Plan	2 way Comprehensive Plan	1 way Comprehensive Plan
Average Net Sales	28.59	53.95	33.97
Average Commission	0.28	0.19	0.64
Distribution Channel	Online(100%)	Online(98.6%)	Online(66.17%) Offline(34%)
Destination	Thailand (33%), US (11%), China (10%)	China (28%), US (17%), Thailand (12%)	China (29%), US (21%), Singapore (18%), Thailand (13%)
Average Age	34 years	37.5 years	62 years

What't the relationship between Net Sales and Commission?



Insight

High claimed products have high commission for high Net Sales and for products that were sold at a high negative sales. However, it has low commission for products that were sold at a low negative sales with majority of the policies being sold at 0 commission.

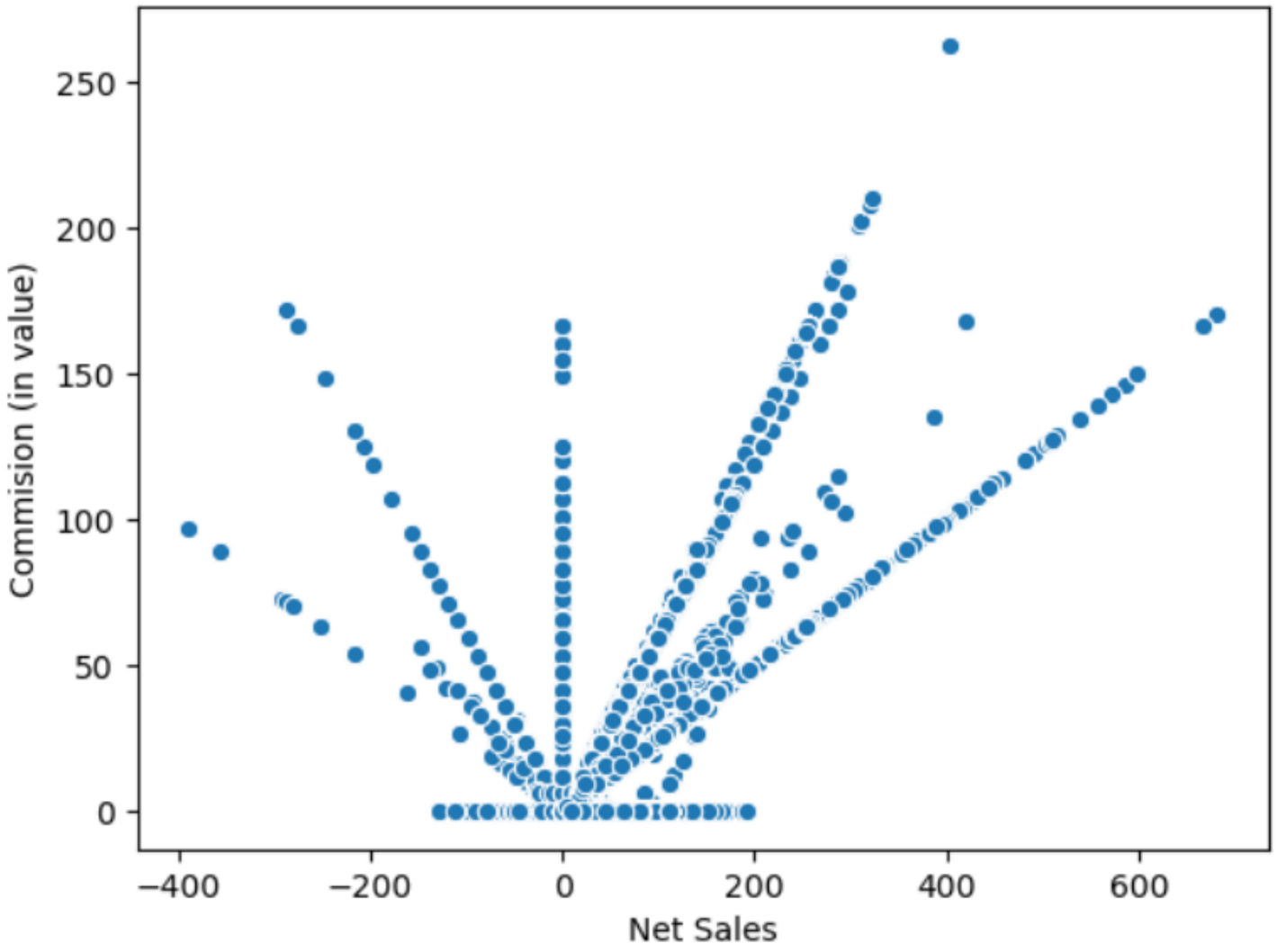
Average commission for high claimed products is 12.26.

For products that have low claim rate, the net sales is in the range of -200 to 400 with the commission lying in the range of 0-50

Recommendation

Low risk profile customers should be charged a low commission to ensure that we have more people buying our products.

Are the negative values an error in data entry?

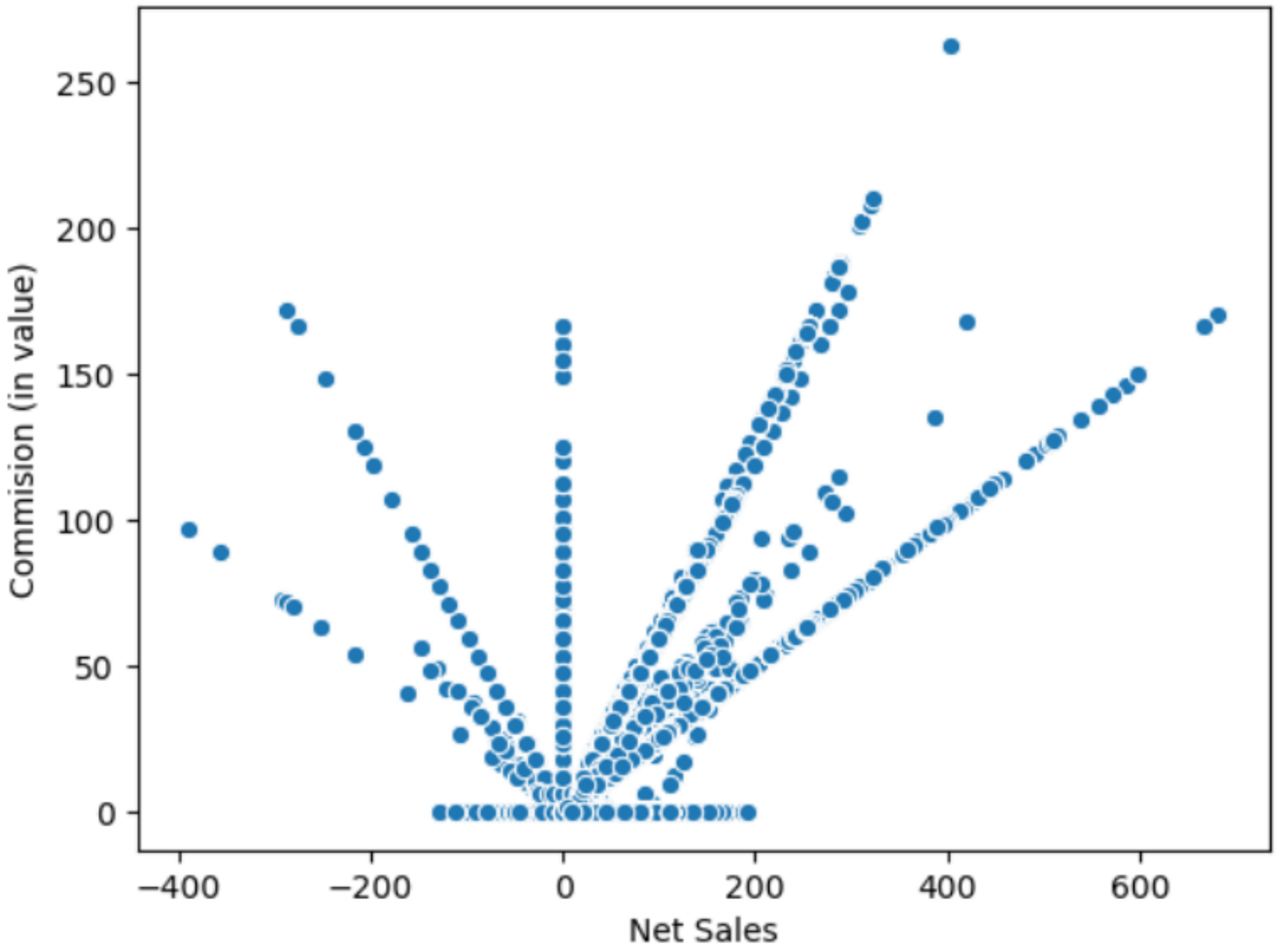


Observation

On first inspection, we observe that there is a symmetry between the positive and negative sales value when plotted against commission

Our Assumption: If there is no difference between commission charged for positive net sales value and negative net sales value, the negative net sales value are error in data entry and we can impute them with absolute values

Are the negative values an error in data entry?



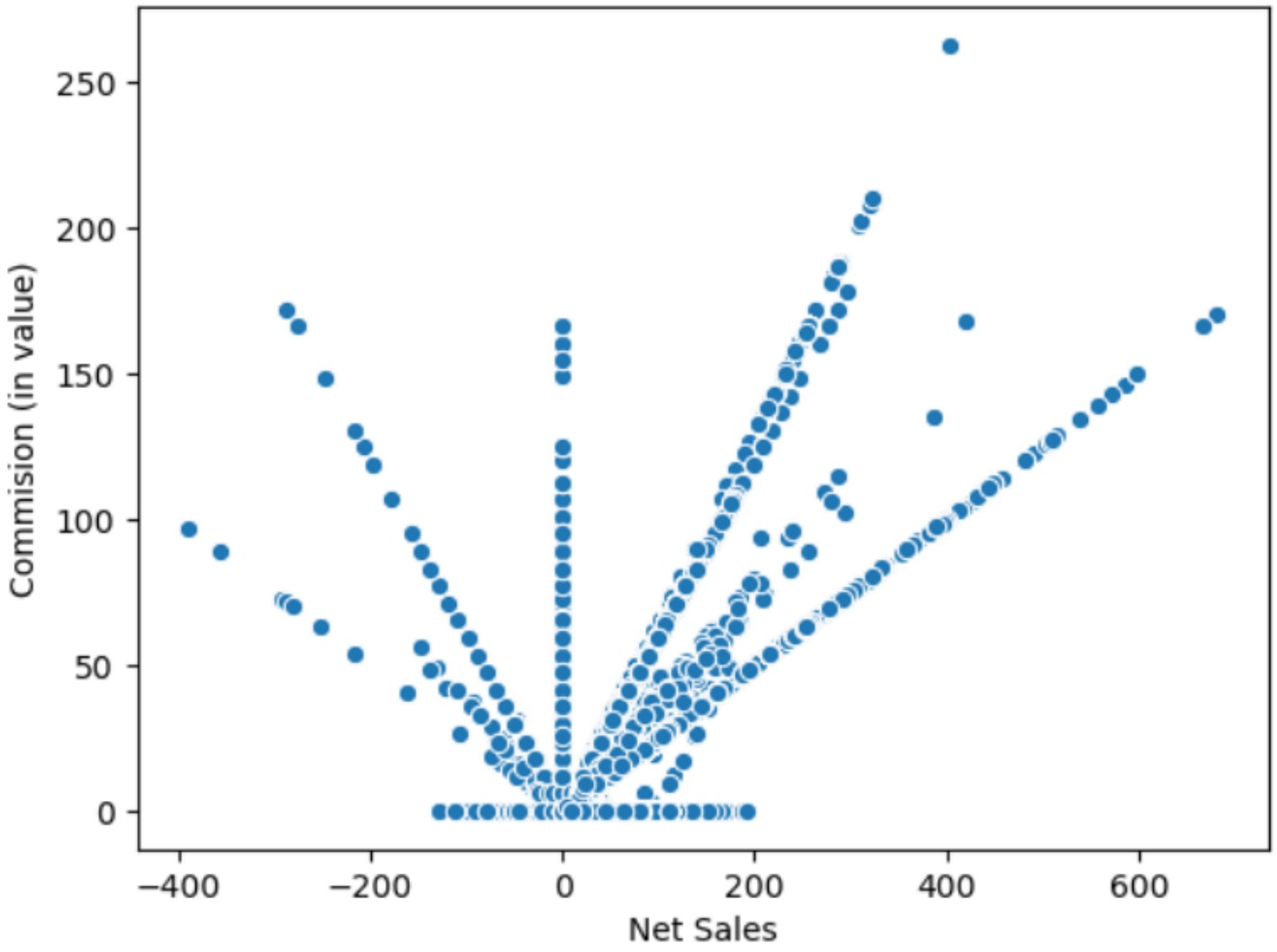
Hypothesis

Null Hypothesis: There is no difference in the commission charged at Net Sales>0 and Net Sales<0

Alternate Hypothesis: There is a difference in the commission charged at Net Sales>0 and Net Sales<0

We conducted two-tailed t-test and found a p-value of 2.71

Are the negative values an error in data entry?



Conclusion

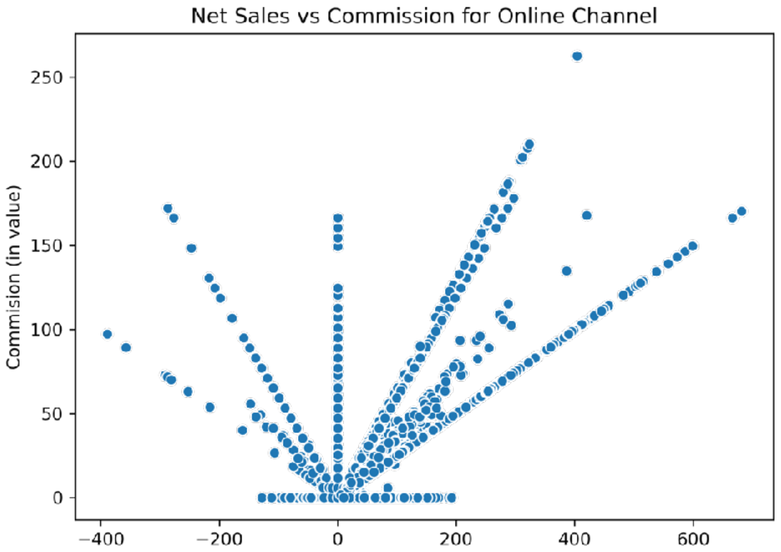
We rejected the Null Hypothesis - There is a significant difference in the commission charged at Net Sales>0 and Net Sales<0.

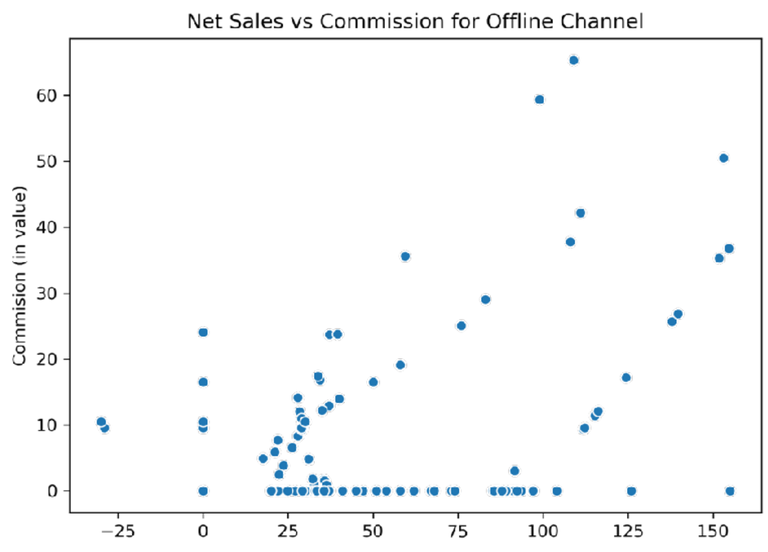
We also found that the average commission charged for negative sales was higher than at positive sales.

Negative sales could be a result of discounts/ promotional offers

How does Distribution Channel and Travel Agency affect Sales and Commission?

Distribution Channel



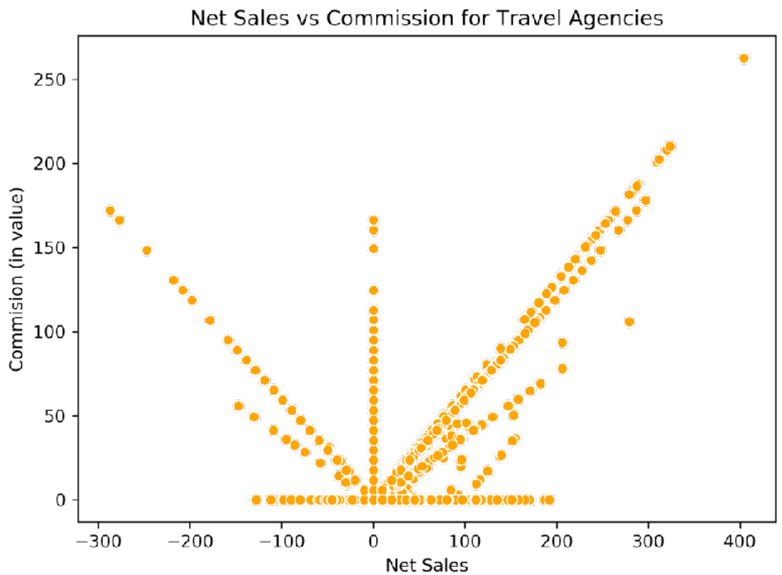


Insight

We see that the policies sold online and through travel agency and airlines have negative net sales.

Since the negative net sales value is very high, we are assuming that online channel must have given heavy discounts (something like buy 1, get 1) to attract more customers.

Agency Type



We also observe that Airlines don't work with 0 commission; while the travel agencies do work with 0 commission.

Assumption: Some Travel Agencies also work as tour operators and could be including the commission in the tour package instead of charging it exclusively.

Insight

Model Building

Model Building Stages

Pre-processing

Feature Engineering

Feature Selection

Baseline Model

Oversampling Techniques

Threshold Tuning

Approach

Strategy 1

Strategy 2

Pre-processing

Feature Engineering

Feature Selection

Baseline Model

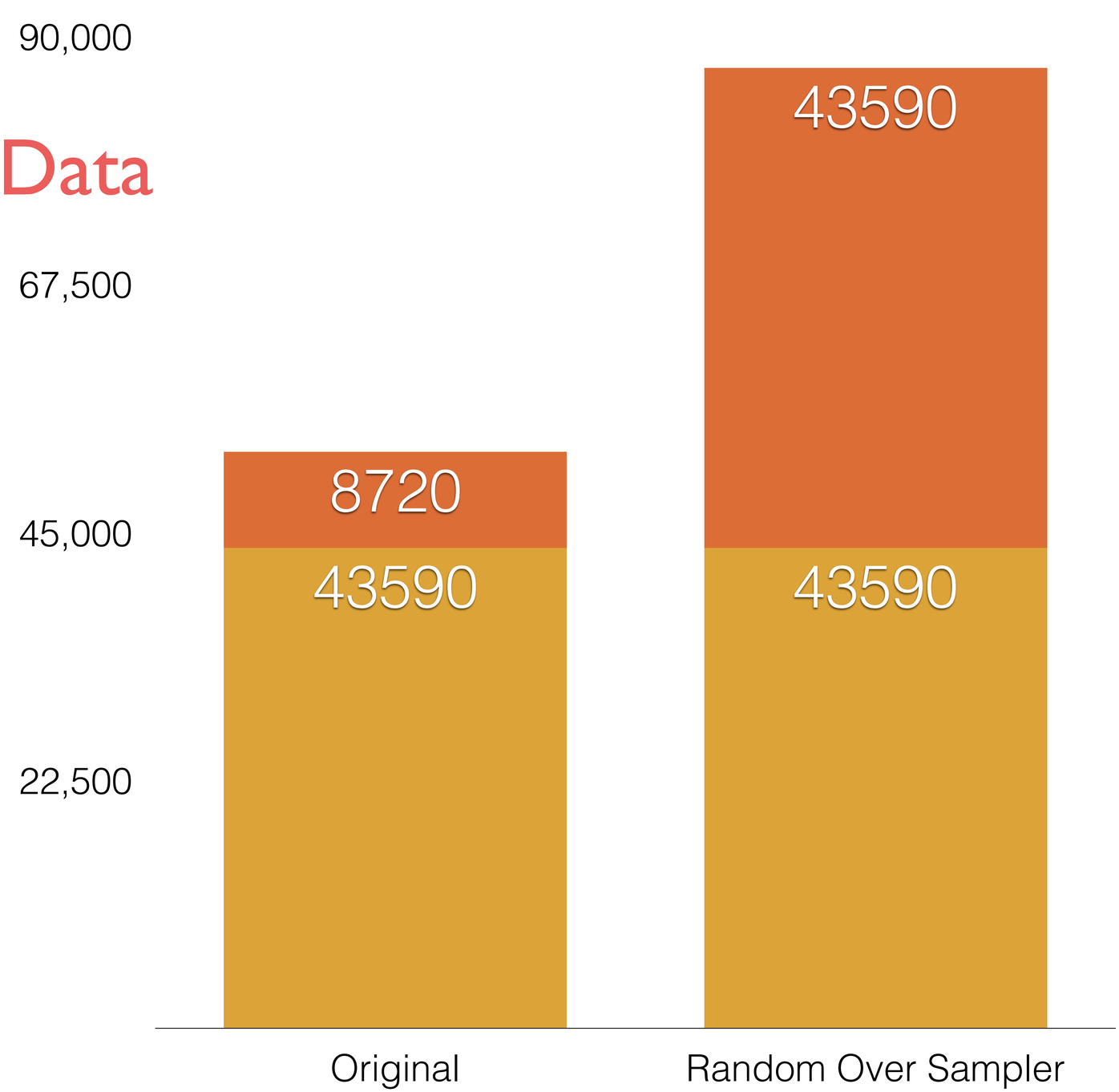
Oversampling Techniques

Threshold Tuning

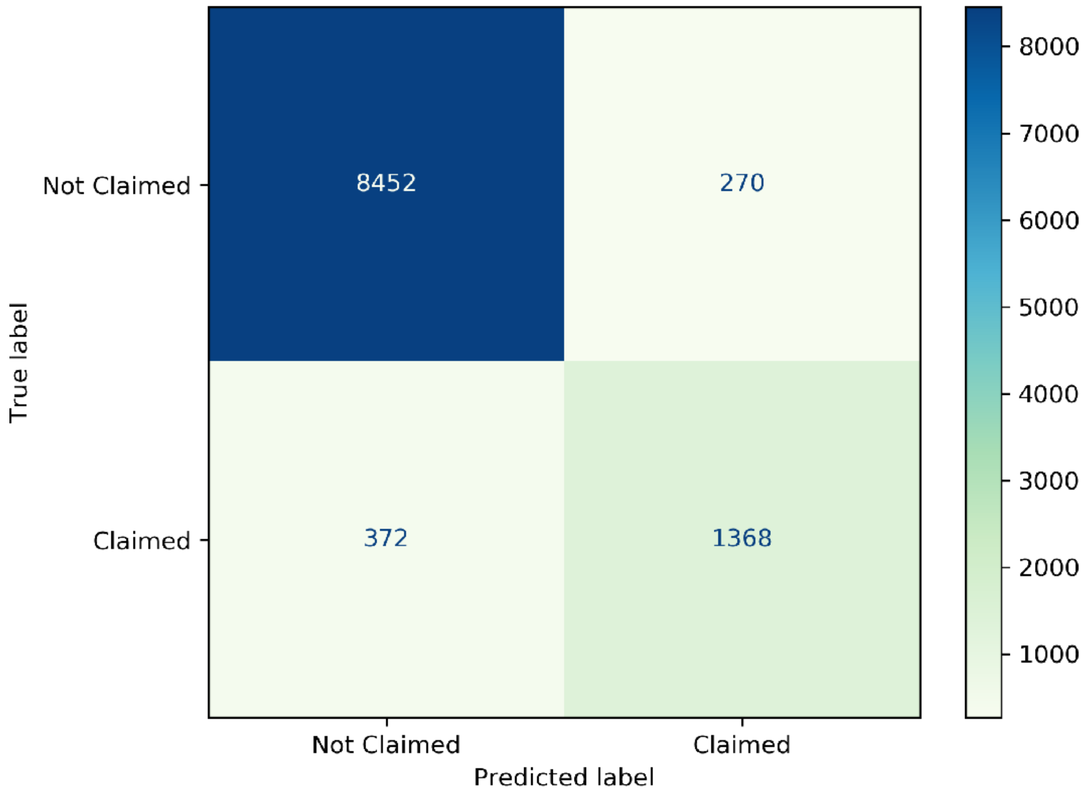
	Strategy 1	Strategy 2
Pre-processing	Duration Column: Replace -1, -2 by 0	Duration Column: Replace -1, -2 by 0
Feature Engineering	- OHE all categorical features, Sqrt. transformation of cont. features, Standard Scaler	- What's Different: OHE countries into top 10 and 'Others'. OHE rest all categorical features
Feature Selection	- Drop ID	- Drop ID, Combine Net Sales and Commission Features (SUM).
Baseline Model	- Ensemble Models (RF highest performing)	- Ensemble Models (RF highest performing)
Oversampling Techniques	- Done. Random Over Sampler	- Done. Random Over Sampler
Threshold Tuning	Probability of predicting 1 fine-tuned to optimise presicion.	Probability of predicting 1 fine-tuned to optimise presicion.

Balancing Data

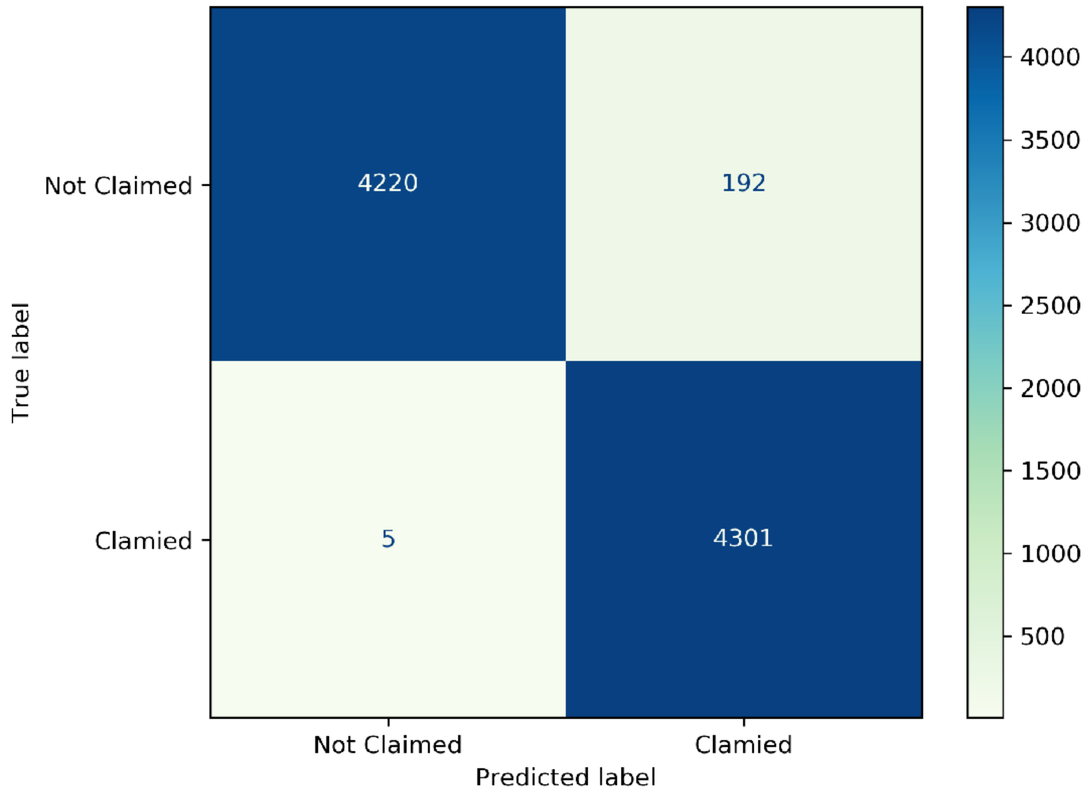
Before and After Over Sampling



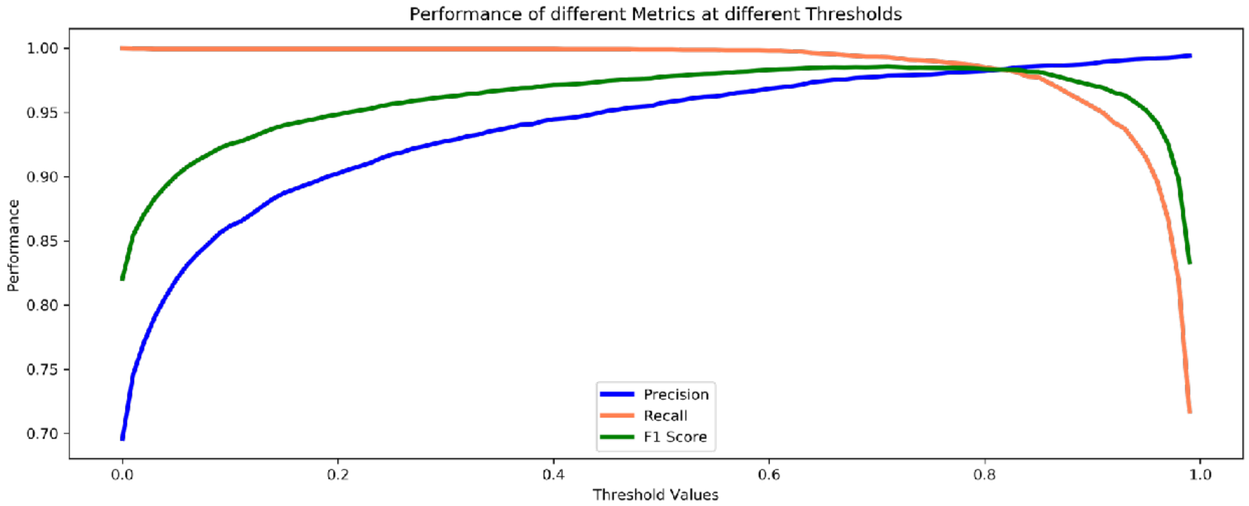
Confusion Matrix



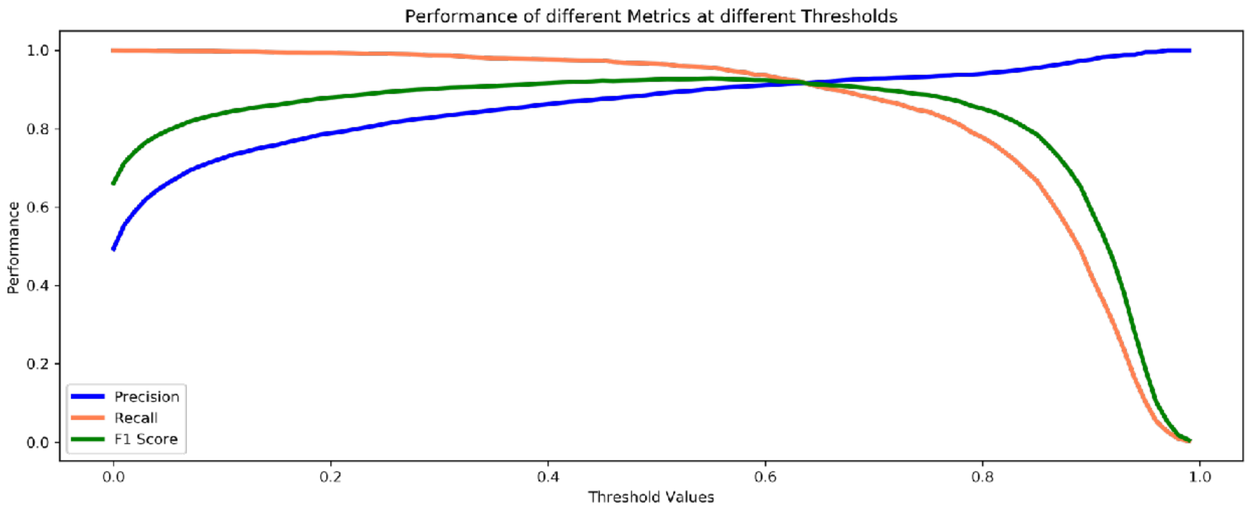
Before OverSampling



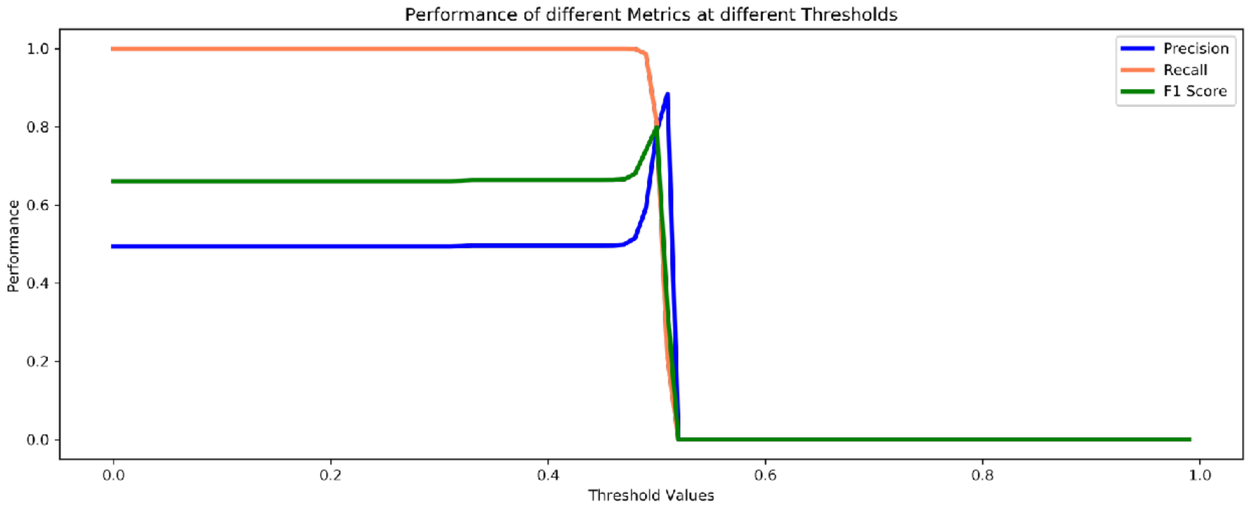
After OverSampling



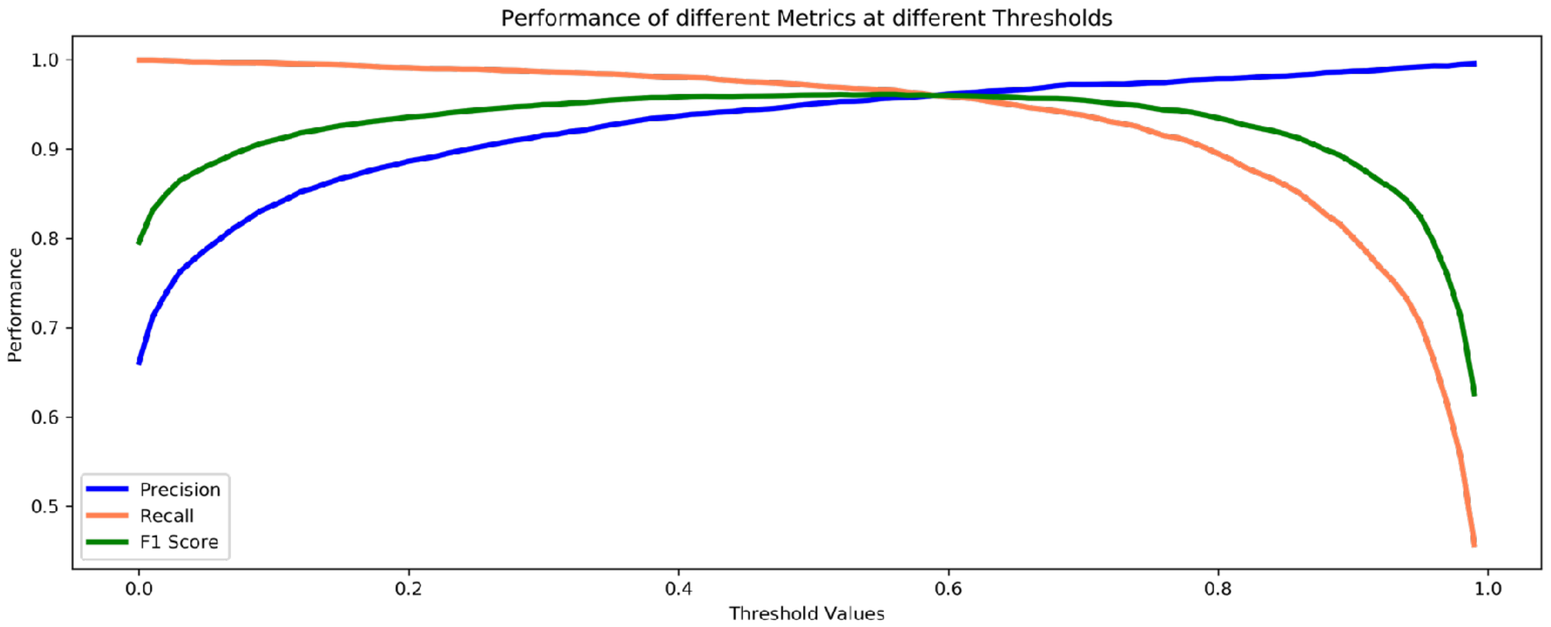
Random Forest Classifier



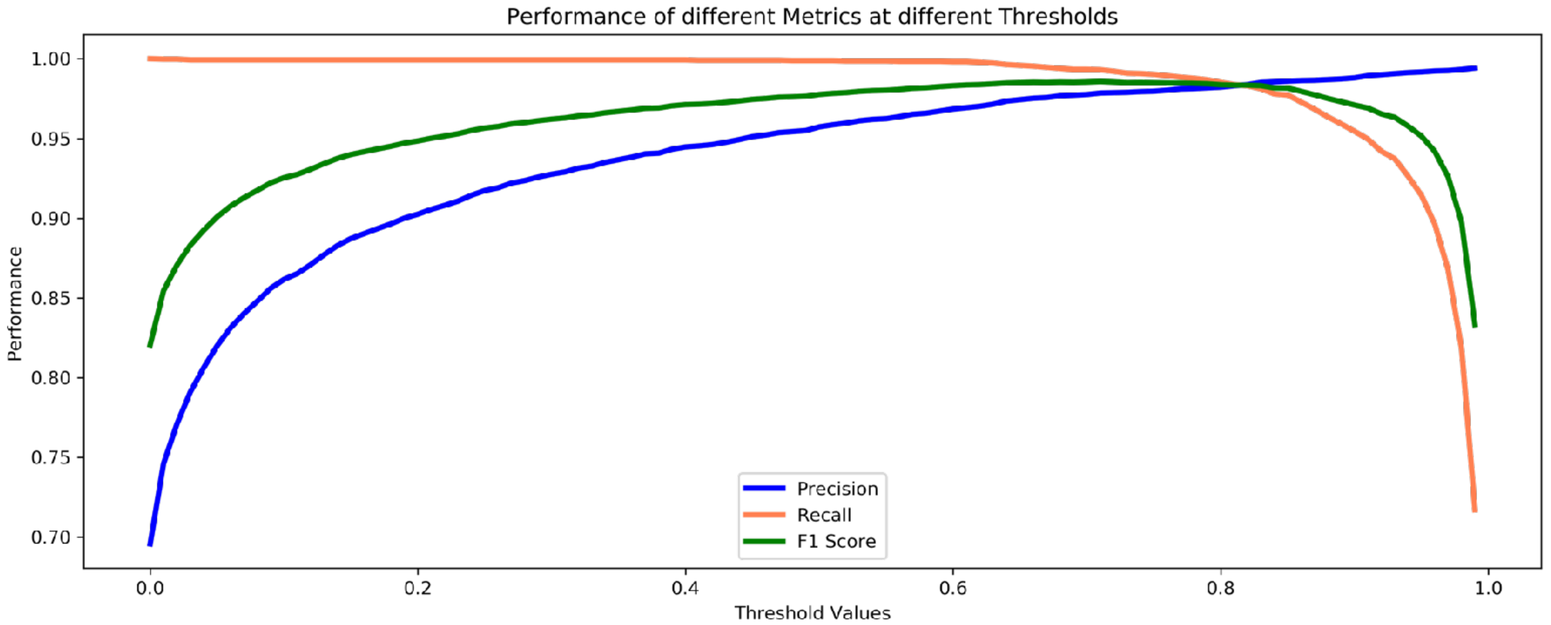
XGBoost Classifier



AdaBoost Classifier

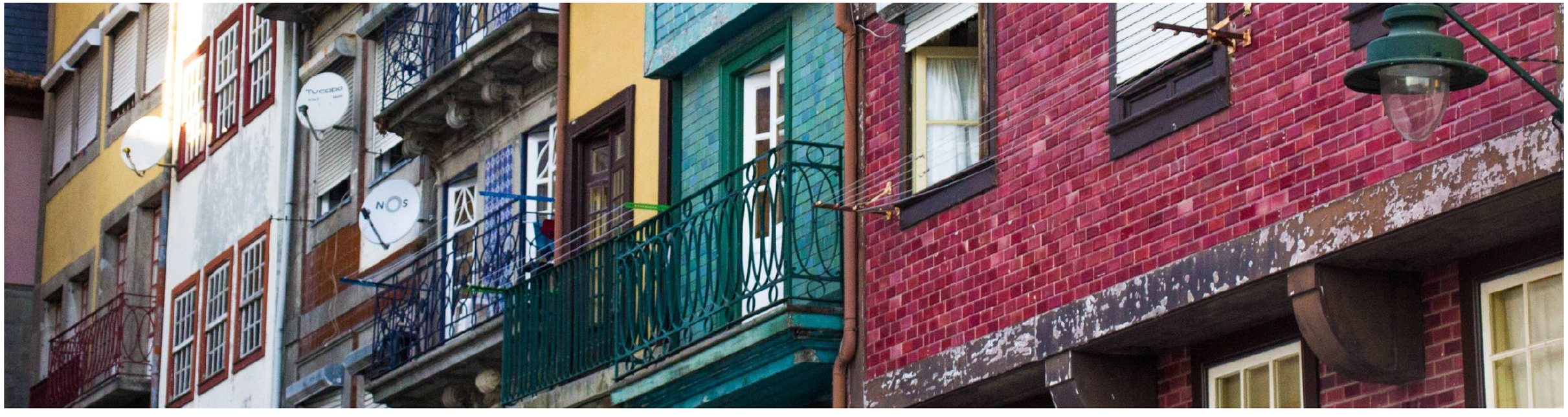


Random Forest Classifier (SMOTE)



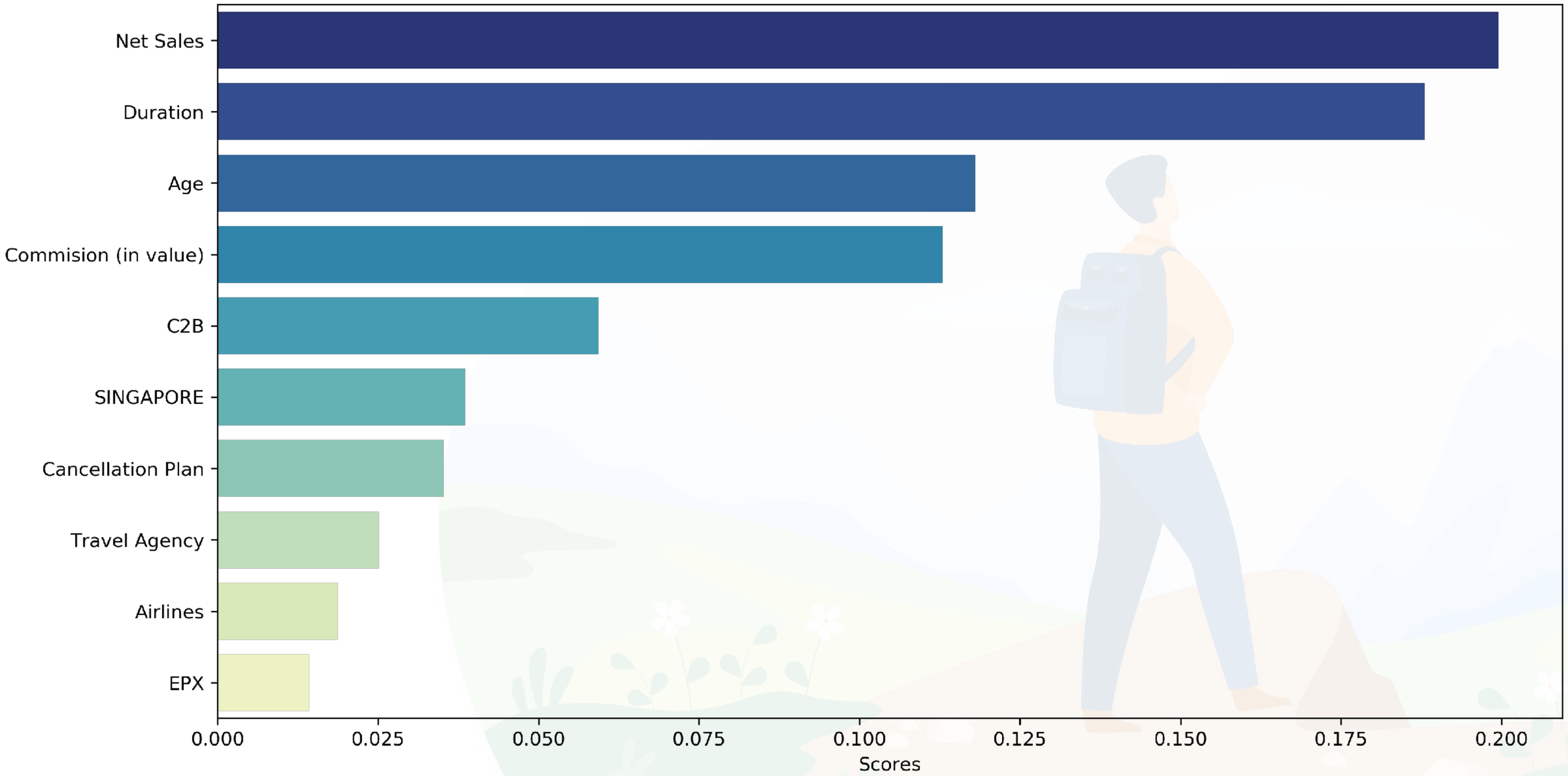
Random Forest Classifier (OverSampling)

SMOTE is computational heavy and less stable over thresholds.
SMOTE gives higher precision of 0.96 on test data.



		Accuracy	Recall	F1 Score	Presicion (Train)	Precision (Test)	Computation Time
Strategy I	Before Oversampling	93.86	78.62	80.99	83.51	84.21	12.07 sec
	After Oversampling (threshold = 0.98)	90.75	81.89	89.74	99.32	93.73	17.82 sec
Strategy II	Before Oversampling	93.1	77.52	78.86	80.24	84.1	5.4 sec
	After Oversampling (threshold = 0.98)	90.26	80.86	89.13	99.28	93.35	12,1 sec

Feature Importance



Conclusion

Learnings

To make payments more affordable we need to have customers with low risk profile. Those will tends to claim less and hence the business can make more revenue/profit accordingly.

The risk to the insurance company increases based on the age of the customer. The claim ratio for the higher age group is 11%, so as to further reduce the risk to the insurance companies we need to cap the duration of insurance to 30 days instead of current average of 46 days.

As per the data customer trvalleing to Singapore, USA and South Africa have claimed the most. Thus the insurance company should charge the higher premium.

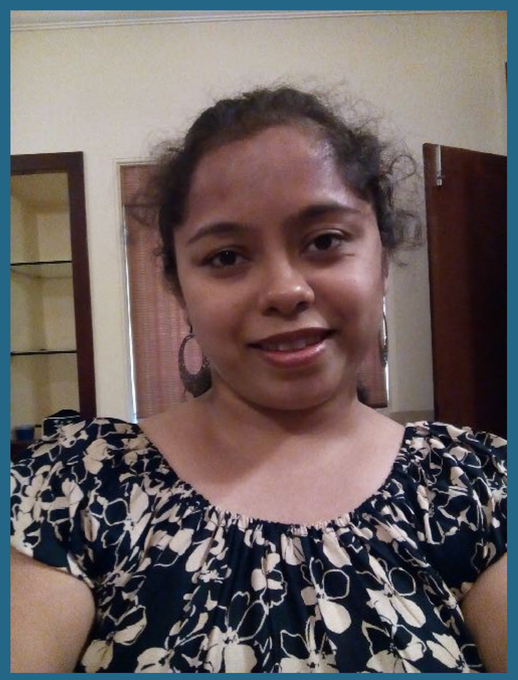
Low risk profile customers should be charged with a low premium of policy in order to ensure that we have more people buying up policies.

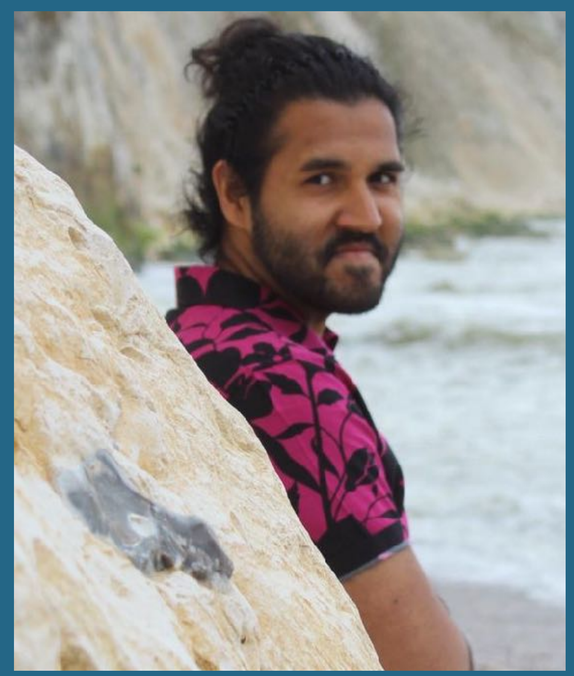
Further Data than would be beneficial: Purpose of travel and source country, Amount Claimed and Amount Sanctioned.

Further Data than would be beneficial: Customer profession, education, income.

		Threshold Values					
		0.5	0.7	0.9	0.95	0.98	Computation Time
Strategy I	Presicion	95.7	97.62	98.89	99.22	99.32	17.82 secs
	Recall	99.88	99.3	95.63	91.68	81.79	
	F1 Score	97.74	98.45	97.23	95.3	89.7	
Strategy II	Presicion	95.21	97.38	98.86	99.18	99.28	12.1 secs
	Recall	99.9	99.34	95.07	90.47	80.86	
	F1 Score	97.5	98.35	96.93	94.63	89.13	

Wanderlust Data





Questions?