Safe Travels Claim Prediction

- by Hustlers

Problem Statement

- Predict if the insurer should sanction the claim

Potential Business Problems

- Minimize losses by rejecting fraudulent claims
- Avoid risk of potential lawsuits arising from rejecting genuine claims
- Improve operations with faster claim settlement process
- Modify product features if some products have a higher claim percentage
- Improve brand equity and goodwill with faster claim sanctions processing

Stakeholders

- Head of Claims department
- Head of products
- Finance Department
- CEO
- Underwriters

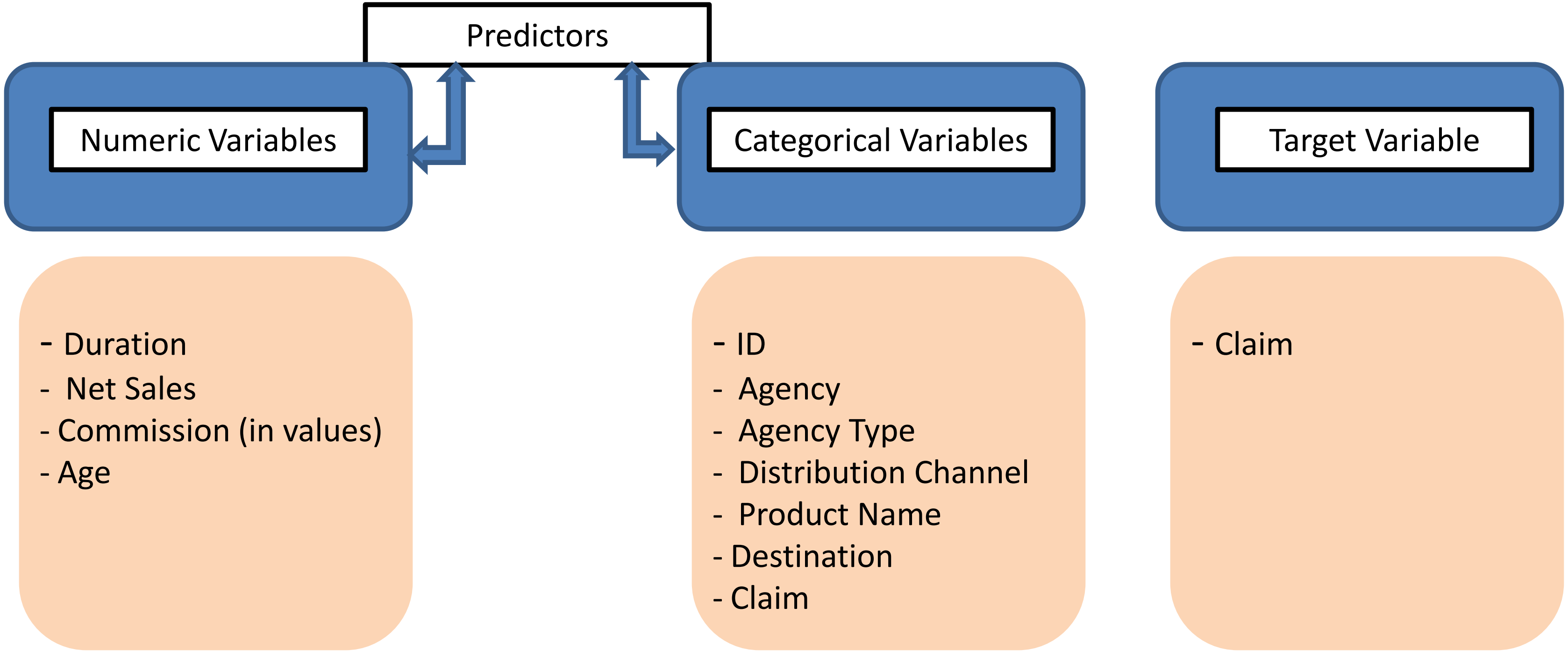
Why solve this problem?

Business Impact

- Improve prediction -> **identify genuine claims->** Sanction genuine claims and avoid losses
- Improve prediction -> **identify products with higher claim percentage**> add exclusions to the policy or change product features/ increase Premium charges
- Improve prediction-> **identify products with lower claims and higher Net Sales-**> will there be future losses and how much. Reserve funds accordingly

Data

Dataset Information : The data consists of records of roughly 52310 clients and 10 features and 1 target that describes whether the claim was sanctioned or not.



Predictors

Numeric Variables

- Duration
- Net Sales
- Commission (in values)
- Age

Categorical Variables

- ID
- Agency
- Agency Type
- Distribution Channel
- Product Name
- Destination
- Claim

Target Variable

- Claim

Evaluation Metric

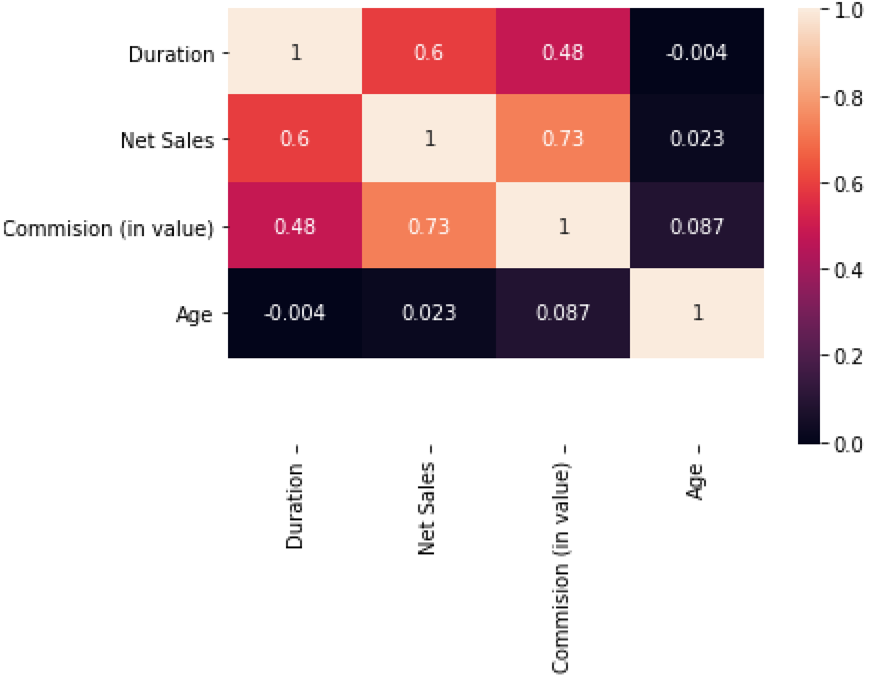
The evaluation metric for this project is **precision_score**

False Positive – predicted to sanction Claim when claim was not genuine

False Negative - predicted to reject Claim, when claim was genuine

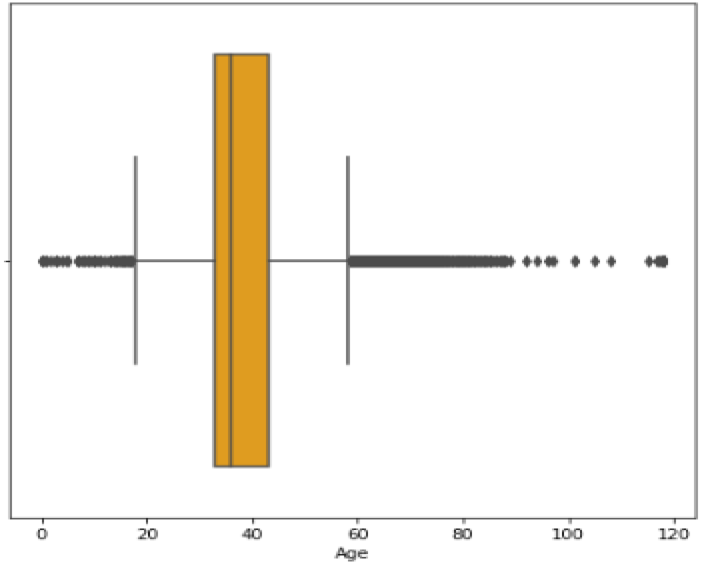
For the use case, reduce False Positives to ensure better **Precision score**

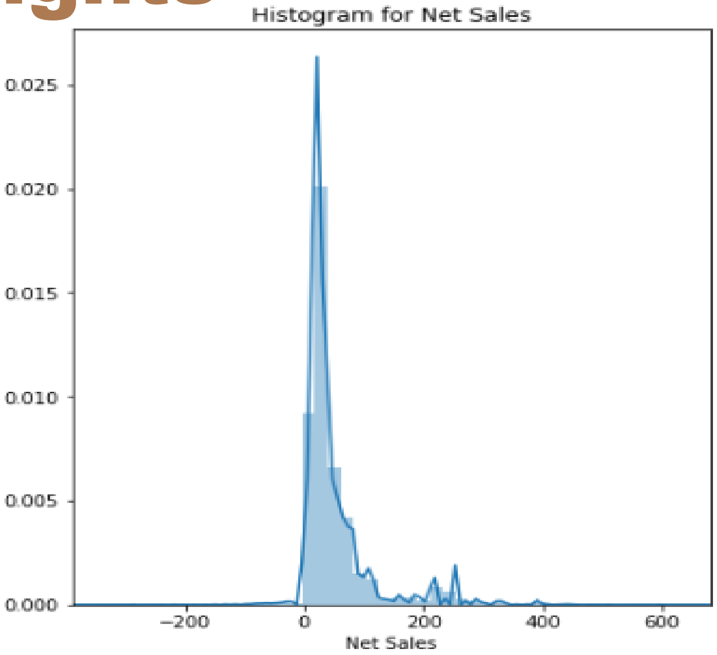
First steps - EDA



The seems some relation between Net Sales, Commission and Duration. Will require further analysis

EDA - continuous - Age, Net Sales - Bring out Key Insights





Age – with a median of 36 most data points are concentrated in 30-45 age group. There is a lot of variance in the data

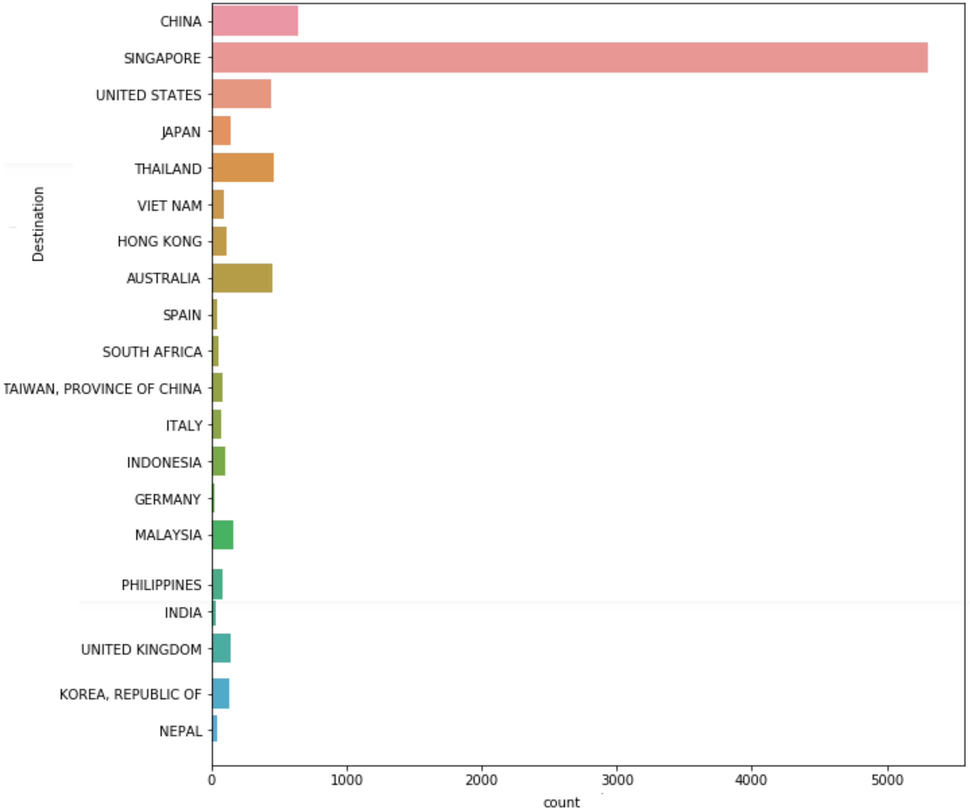
Net Sales- This feature has a few negative values and the data is right skewed

EDA – categorical features

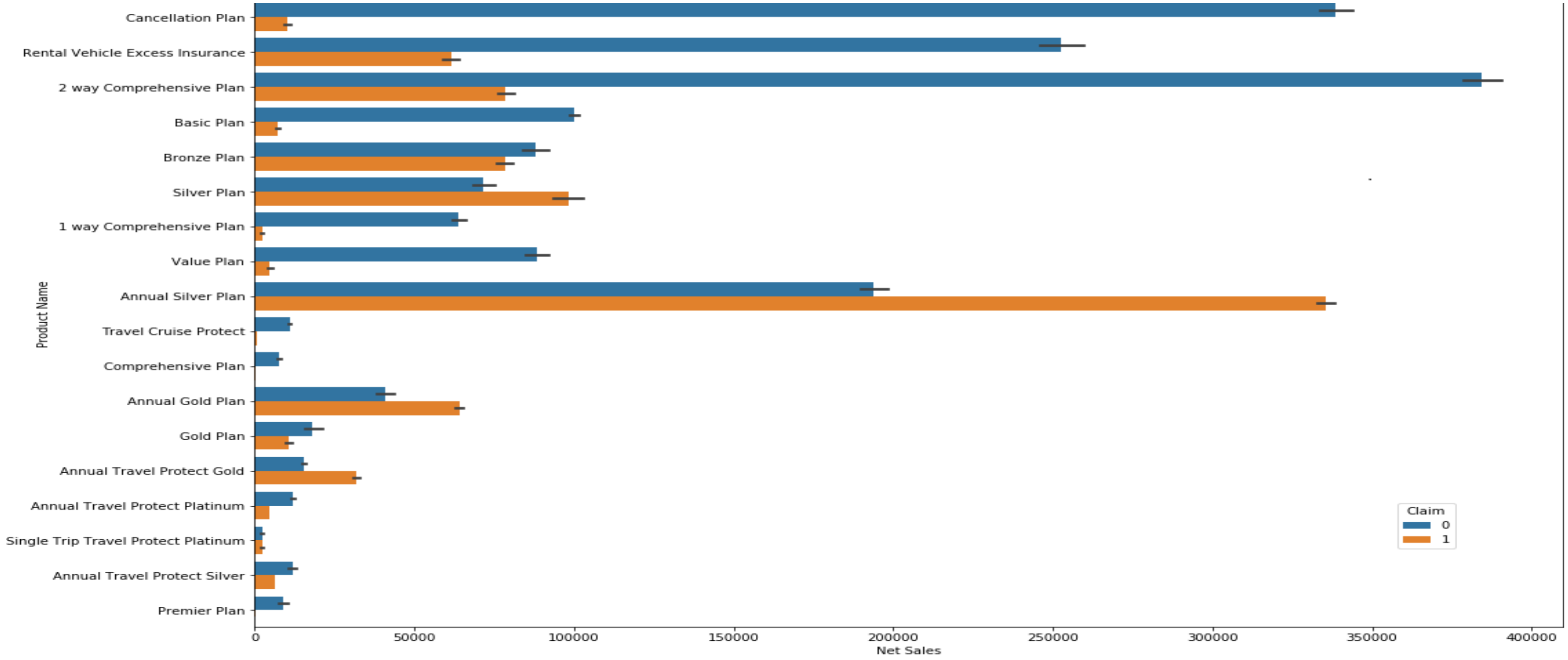
Singapore has the highest number of claims sanctioned at 5306 followed by China at 642 claims sanctioned.

The variance is too high within this feature

After EDA, Destinations with <50 counts were grouped as “Others”



EDA - Multivariate



Annual Silver Plan has the highest Net Sales and 65% of its Claims sanctioned.

Bronze Plan has the most number of Claims sanctioned and 5th highest Net sales.

There are a few products with lower Net Sales but higher numbers of claims sanctioned.

Pipeline

Outlier treatment :

The Outliers in the continuous features were detected and later treated using a method called **Winsorization.**

Columns	Values before Winsorization	Result
Duration	45327	5484
Net Sales	34494	0
Commission	49154	6394
Age	3578	0

Pipeline

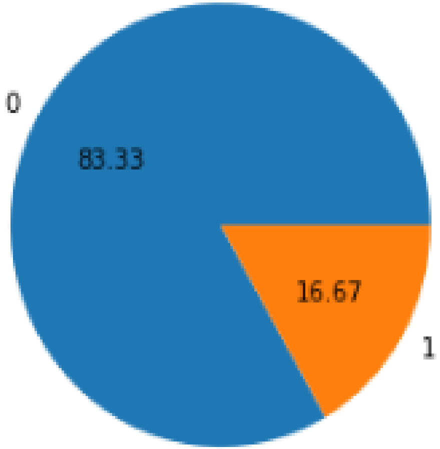
Missing Values :

1. There were no missing values in the continuous features
2. The categorical features also didn’t have any missing or garbage values.

Class Imbalance :

The distribution of the target below shows a clear imbalance in the two classes namely

0- Claim Rejected
1- Claim Sanctioned



Models and Approaches

Three vanilla models were assessed without performing any hyper parameter tuning and without treatment of class imbalance of the target. The models were

-LogisticRegression
-RandomForestClassifier
-ExtraTreesClassifier
-XGBoost Classifier
-GradientBoostingClassifier
-XGBClassifier
-VotingClassifier (hard and soft)

RandomForest, VotingClassifier and ExtraTrees gave good precision score compared to others.

This called for performing hyper parameter tuning using RandomizedSearchCV and also treatment of class imbalance using RandomOverSampler for further improvement of the precision_score.

Models and Approaches

Models Assessed : The vanilla models used yielded the following results below.

Modelling Method	Precision score
LogisticRegression	0.610130
DecisionTreeClassifier	0.736162
RandomForestClassifier	0.823813
GradientBoostingClassifier	0.640496
XGBClassifier	0.787495

Handling data imbalance

Techniques used:

- SMOTETomek

-RandomOverSampler

Observations:

-Not major difference was noticed in both up sampling techniques

- RandomOverSampler sampling Technique on DecisionTreeClassifier, RandomForestClassifier and VotingClassifier(Hard) models generated better results

Pipeline

Feature Selection :

Following methods were used for feature selection :

-Correlation
-SelectKBest- mutual_info_classif
-RFE

After estimating Pearson Correlation coefficients between continuous features Commision was dropped

Pipeline

Feature Selection :

- SelectKBest model with mutual_info_classif was used to select the best features. From 88 features it selected 64 best features

- Recursive feature elimination (RFE) was performed using Logistic Regression as the estimators. It improved the score slightly; however compared to other ensembling models the score was not good hence we chose to not implement this technique.

Model Tuning

After performing hyper parameter tuning using RandomizedSearchCV, treating imbalanced classes it was observed that **Hard Voting Classifier** gave the best precision score and is our chosen model.

The precision score is 0.79

Final Results

From the above observations it can be inferred that the best performing model was Hard Voting Classifier giving an a precision_score of 79.6%.

Confusion Matrix :

	Predicted Positive	Predicted Negative
Actual Positive	1574	401
Actual Negative	1040	12678

Insights &Decisions

Products- Cancellation Plan and 2 way Comprehensive Plan are the most selling products and this is reflected in their Net Sales too. These products also have a lower number of claims sanctioned. These are revenue generators and could be marketed more

Destination- Asia is the dominant market for Safe Travel and Singapore is the leading destination with over 14000 products sold

Risk- Although Annual Silver Plan, Bronze Plan, Annual Gold Plan, Annual Travel Protect Gold have high Net Sales, they have a higher claims percentage ranging from 40-65%. With such high claims rate these products could be impacting the revenues of Safe Travels.

Next Steps????

If time permitted, could have tried the following :

- Better feature engineering
- An ensemble of different models
- Create Api and implement on Azure, Google cloud and AWS

Thank You!