Tour Packages Sales Predicting Model

Ola Barradh

Problem Description

The travel agency is suffering from low number of sales despite the resources directed on pitch sales.

This presentation aims to:

- Explore any patterns regarding the success or fail of a sale pitch
- Build a predictive model for future customers

01

Tools

02

Dataset

03

Findings

04

Data Model

01

Tools

Tools

Jupyter Notebook

Data Processing

Numpy, Pandas

Vizualization

Matplotlib and Seaborn

Modeling

imblearn and sickit-learn

02

Dataset

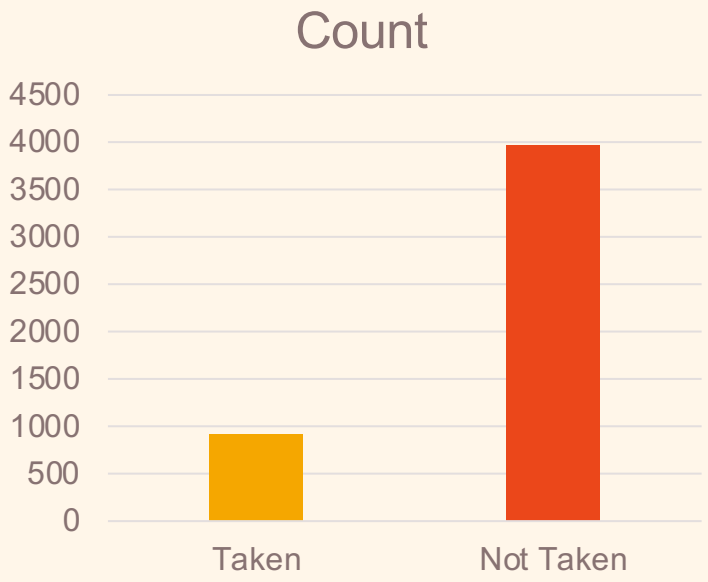
- The data set is made of 4888 rows and 20 columns.
- The dataset consists of different datatypes and is a mix of ordinal, categorical and continuous data
- The data covers information both about the customer and the sale pitch interaction

```
 #   Column
---  ------
 0   CustomerID
 1   ProdTaken
 2   Age
 3   TypeofContact
 4   CityTier
 5   DurationOfPitch
 6   Occupation
 7   Gender
 8   NumberOfPersonVisiting
 9   NumberOfFollowups
 10  ProductPitched
 11  PreferredPropertyStar
 12  MaritalStatus
 13  NumberOfTrips
 14  Passport
 15  PitchSatisfactionScore
 16  OwnCar
 17  NumberOfChildrenVisiting
 18  Designation
 19  MonthlyIncome
dtypes: float64(7), int64(7),
memory usage: 763.9+ KB
```

03

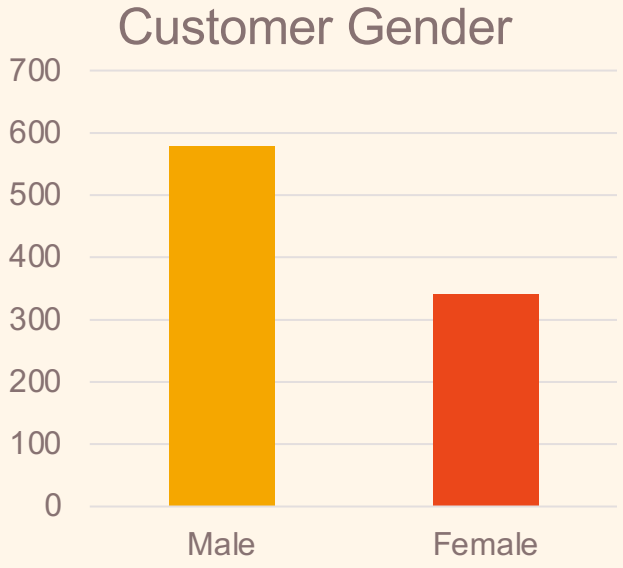
Findings

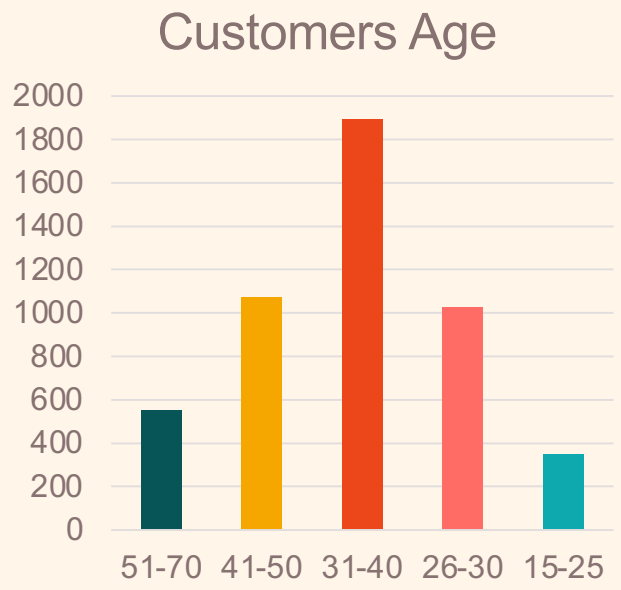
Products Taken



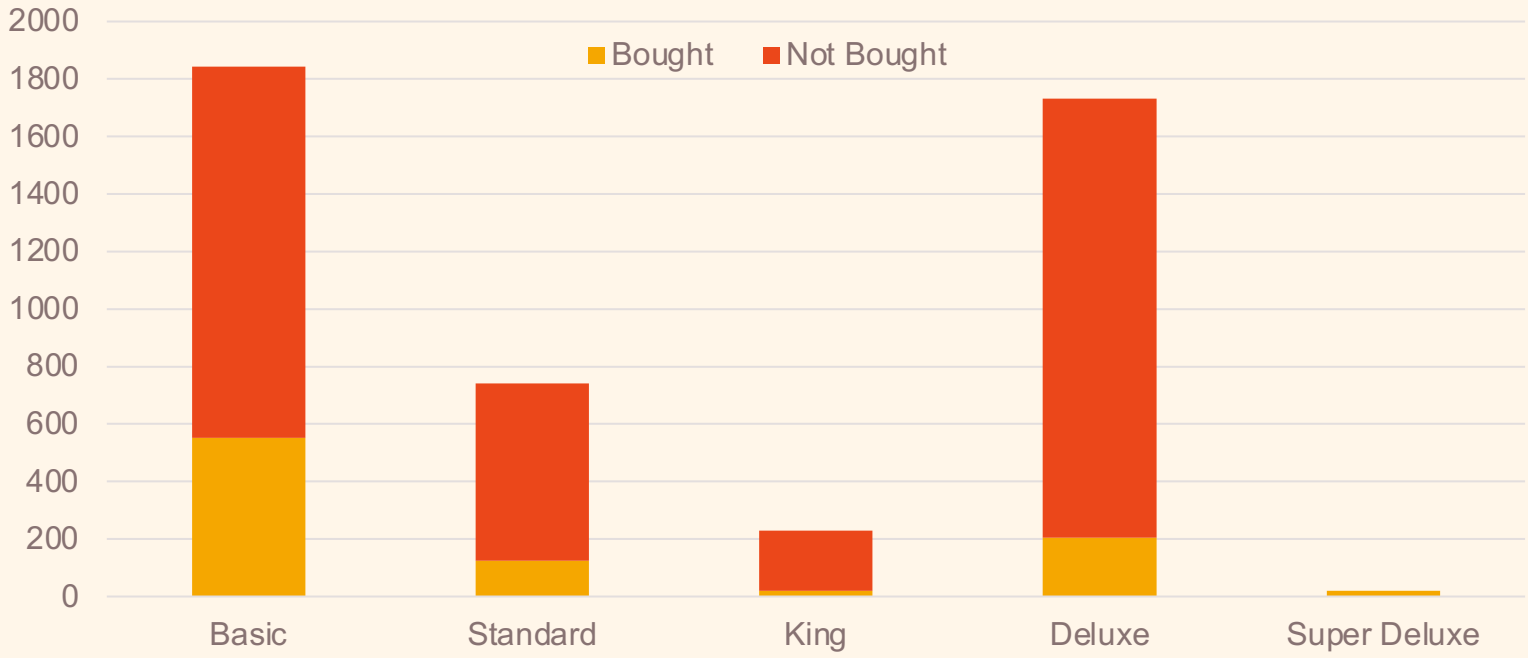
- We notice very low number of sales, indicating big amount of wasted time and resources

Personal Profile of Customers





Products pitched vs. Products Bought



04

Data Model

Testing Plan

Data

- Clean data
- Clean data + Oversampling
- Clean data +oversampling + feature selection

Over Sampling

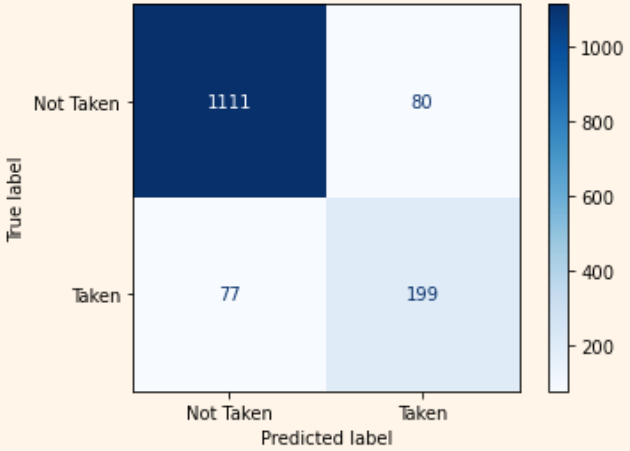
- SMOTE
- RandomOver Sampler

Classification

- Random Forest
- Logistic Regression
- XGBoost
- AdaBoost

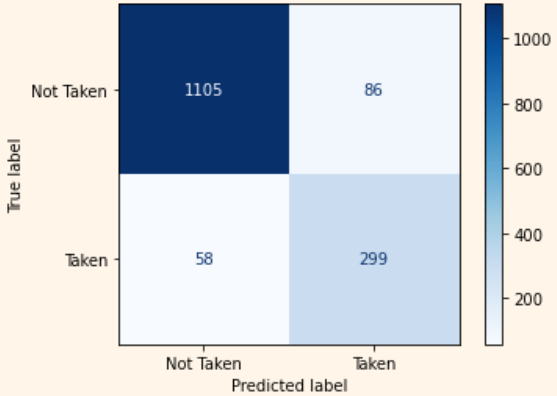
Testing (Clean Data)

- Testing on Clean Data resulted in high accuracy 90% . The Recall was at 72%.

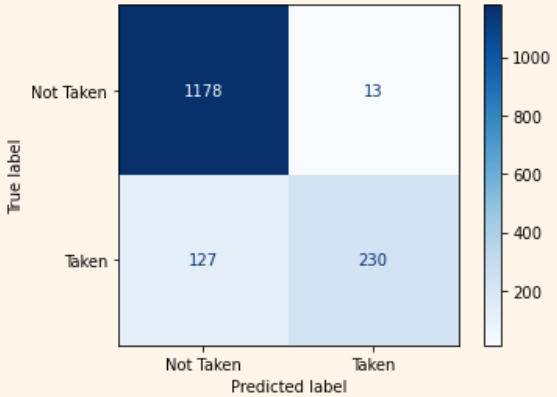


Testing (Over Sampling)

- Mixed and matched Oversampling algorithms with different classification models
- RandomOverSampler performed better than SMOTE, with higher recall, accuracy and lower False negatives and false positives



RandomOverSampler + Random Forest



SMOTE+ Random Forest

RESULTS

Model	Acuuracy	Recall	Percision	F1
Decision Tree	91%	84%	77%	80%
Random Forest	91%	73%	93%	82%
XGBoost	86%	55%	81%	60%
AdaBoost	83%	45%	73%	55%
Logistic Regression	80%	13%	86%	23%

- Data was skewed so we needed to Over Sample
- RandomOverSampler outperformed SMOTE
- Random Forest produced the highest accuracy

Thank you for Listening