



# 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 numeric 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



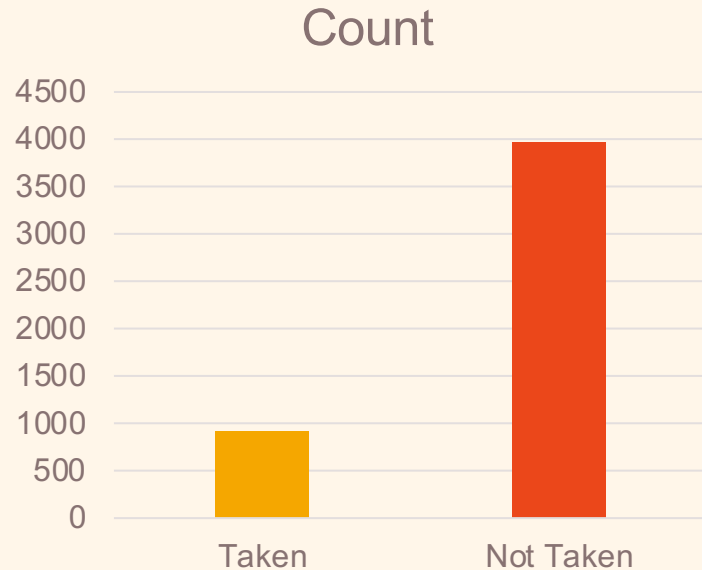


03

## **Findings**

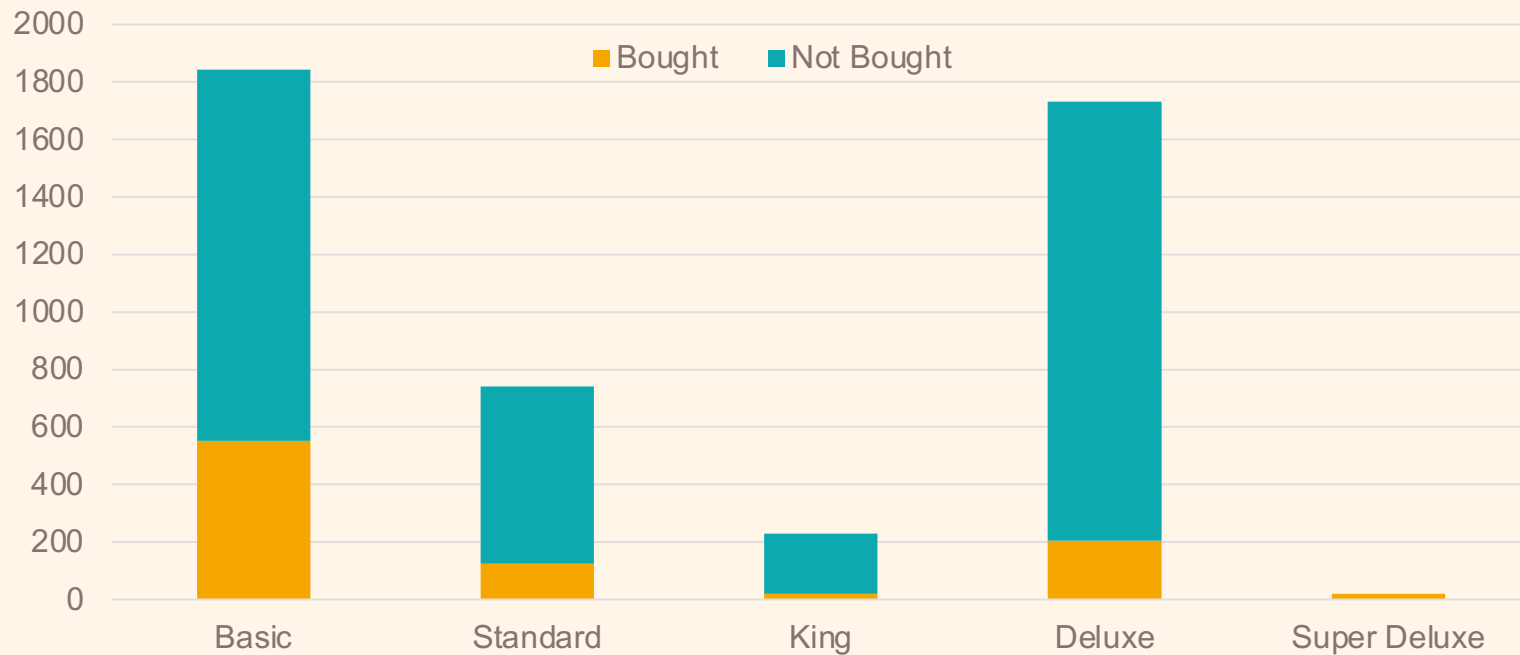


# Products Taken



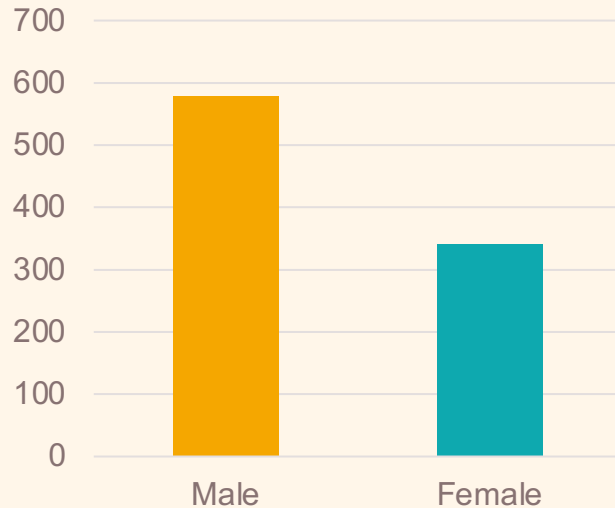
- We notice very low number of sales, indicating big amount of wasted time and resources

## Products pitched vs. Products Bought

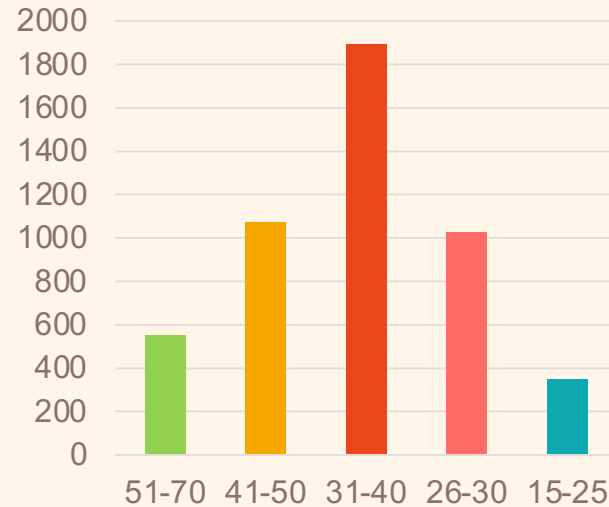


# Personal Profile of Customers

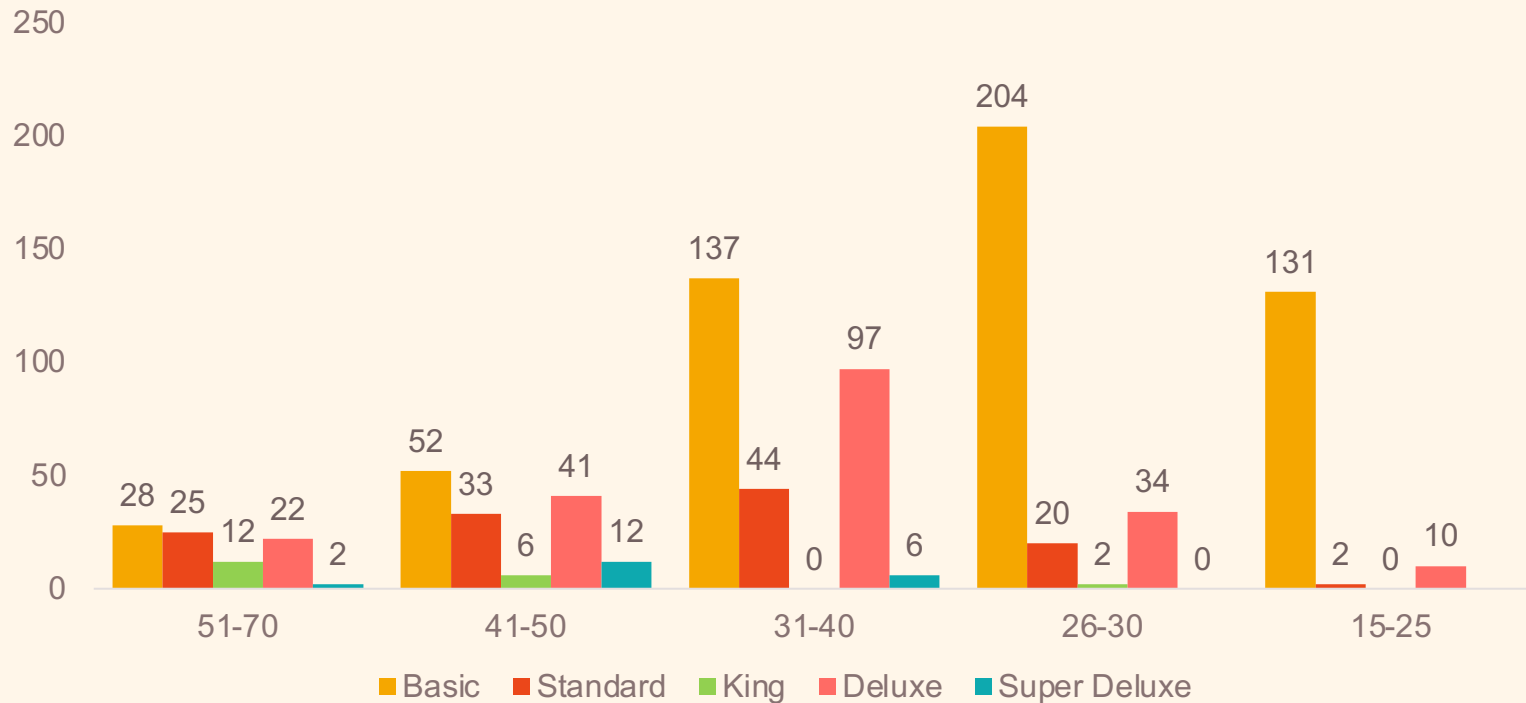
## Customer Gender



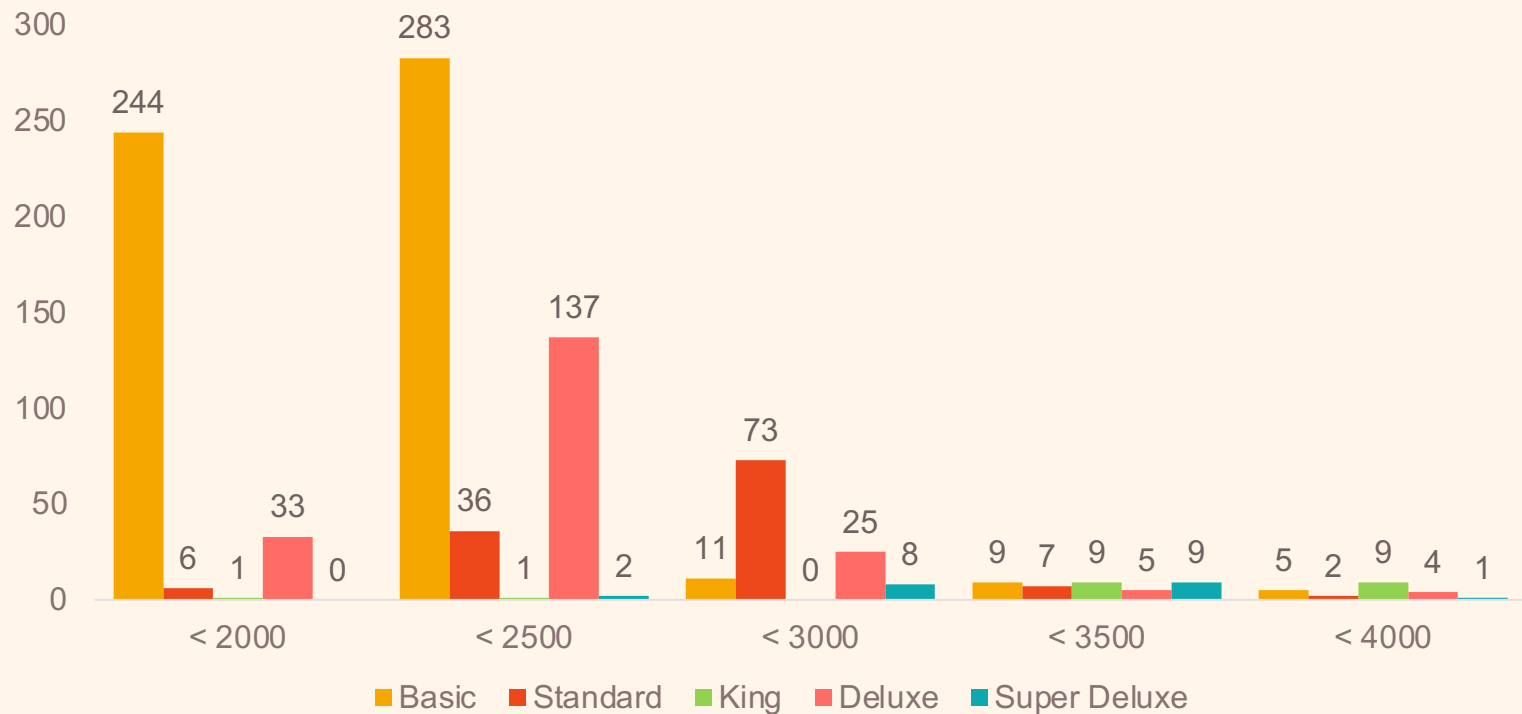
## Customers Age



## Customer age and purchased product



## Monthly Income and purchased product





04

## **Data Model**



# Testing Plan



## Data

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



## Over Sampling

- SMOTE
- RandomOver Sampler



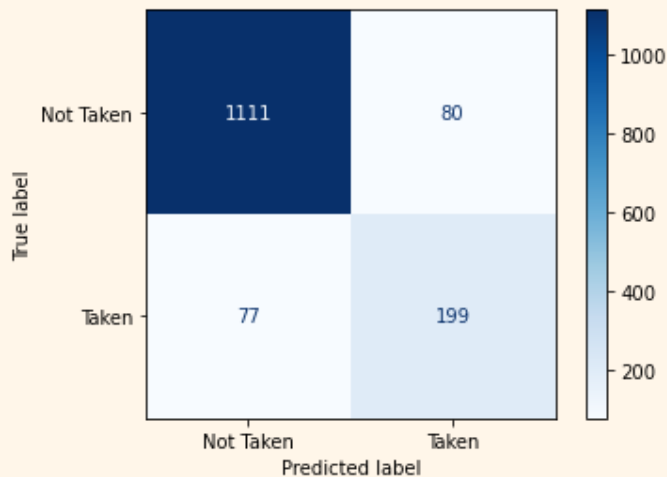
## Classification

- Random Forest
- Logistic Regression
- XGBoost
- AdaBoost

# Testing (Clean Data)



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

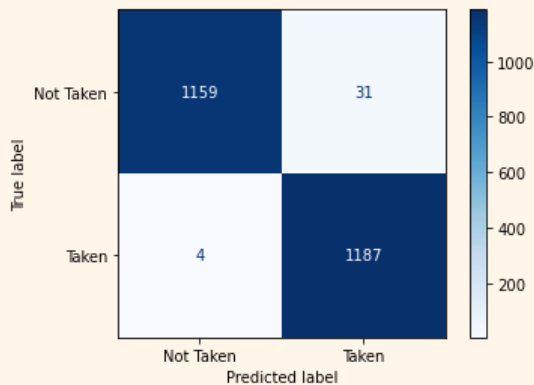




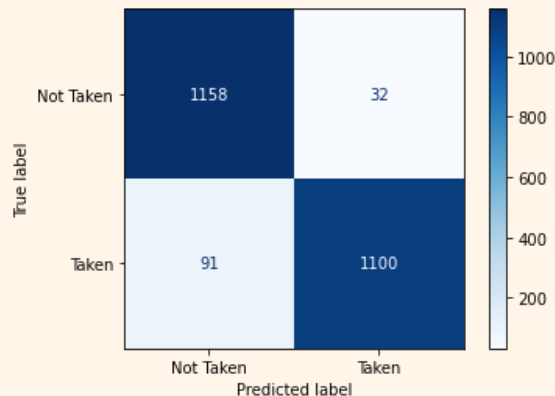
# Testing (Over Sampling)



- Mixed and matched Oversampling algorithms with different classification models
- RandomOverSampler tends to cause overfitting of the data



RandomOverSampler + Random Forest



SMOTE+ Random Forest

## RESULTS (SMOTE)



Model	Accuracy	Recall	Precision	F1
Decision Tree	90%	90%	91%	90%
Random Forest	94%	92%	97%	94%
XGBoost	90%	87%	93%	90%
AdaBoost	88%	85%	90%	87%
Logistic Regression	87%	83%	91%	87%

- Data was skewed so we needed to Over Sample
- Good results, leans toward overfitting
- Random Forest produced the highest accuracy

## RESULTS (RANDOM OVER SAMPLER)



Model	Accuracy	Recall	Precision	F1
Decision Tree	96%	98%	93%	95%
Random Forest	98%	99%	97%	98%
XGBoost	82%	82%	81%	82%
AdaBoost	76%	77%	75%	76%
Logistic Regression	73%	74%	72%	73%

- Data was skewed so we needed to Over Sample
- RandomOverSampler higher results than SMOTE
- Random Forest produced too high of a result (Overfitting)

## RESULTS (FEATURE SELECTION+SMOTE)



Model	Accuracy	Recall	Precision	F1
Decision Tree	89%	91%	88%	89%
Random Forest	88%	90%	87%	89%
XGBoost	90%	85%	93%	89%
AdaBoost	89%	85%	93%	89%
Logistic Regression	86%	83%	88%	86%

- Selected features based on customer profile only
- Numbers are lower >> more balanced
- XGBoost produced the highest accuracy

## RESULTS (FEATURE SELECTION+OverSampler)



Model	Accuracy	Recall	Precision	F1
Decision Tree	95%	99%	91%	95%
Random Forest	95%	99%	91%	95%
XGBoost	79%	78%	80%	79%
AdaBoost	74%	75%	73%	74%
Logistic Regression	72%	74%	71%	73%

- Selected features based on customer profile only
- Overfitting
- Random Forest produced the highest accuracy



**Thank you for  
Listening**