DATA SCIENCE PORTFOLIO

Widya Ayuningtyas

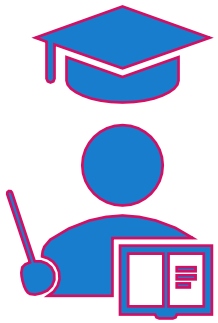
June
2022

Hello, my name is

Widya Ayuningtyas.

I have completed my Master degree of Management and finished my Data Science Bootcamp from dibimbing.id.

EDUCATION



Widyatama University

Graduated : 2021

Parahyangan Catholic University

Graduated : 2017

WORKING EXPERIENCES

PT EWINDO (2017 – 2018)

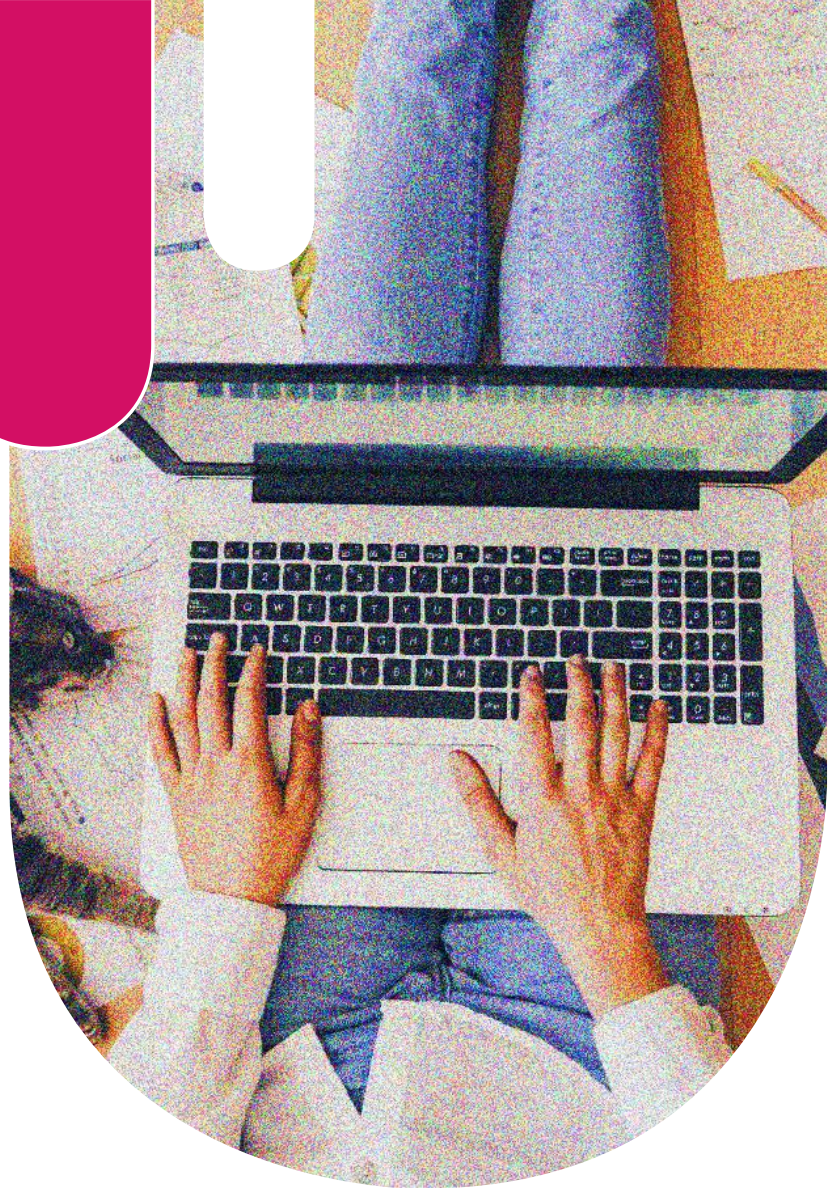
Production Planning and Inventory Control Staff

- Created forecasting plans and communicated deadlines to ensure scheduled materials order were completed on time.
- Completed paperwork, recognizing discrepancies and promptly addressing for resolution in team meeting.
- Monitored company inventory to keep stock levels and databases updated.

PT DIAN MEGAH INDO PERKASA (2014)

Research Intern

- Gathered, arranged and corrected research data to create representative graphs and charts highlighting results for presentations.



SKILLS AND PROFICIENCY

- PYTHON CODING
- DATA VISUALIZATION
- MACHINE LEARNING
- SQL DATABASE

CERTIFICATION

DQLAB

Python Fundamental

Dibimbing.id

DATA SCIENCE

Kindly check all my certificates at :
https://drive.google.com/drive/folders/11YNelQvxv6CB4MpEPpoZVtxzhSublUmZ?usp=sharing

DATA SCIENCE PORTFOLIO

- Beginner Data Science Projects

(Data Cleaning, Data Wrangling, Exploratory Data Analysis, Data Visualization, SQL Exercises)

- Customer Personality Analysis

https://github.com/widyaayuningtyas7/EDA-Customer-Personality-Analysis

- **Holiday Package Purchase Prediction**

https://github.com/widyaayuningtyas7/HOLIDAY-PACKAGE-PURCHASE-PREDICTION

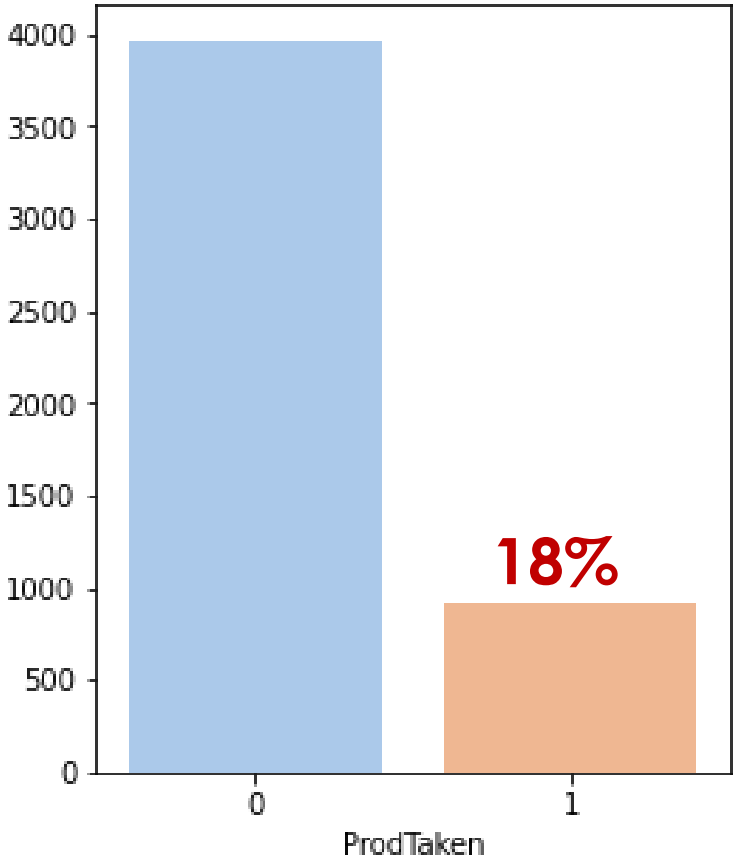
HOLIDAY PACKAGE PURCHASE PREDICTION

OUTLINE

- BACKGROUND
- DATA UNDERSTANDING
- EXPLORATORY DATA ANALYSIS
- MODELLING AND EVALUATION
- BUSINESS INSIGHT AND RECOMMENDATION

BACKGROUND

Last year sales, only 18% of offered product are taken by customers.



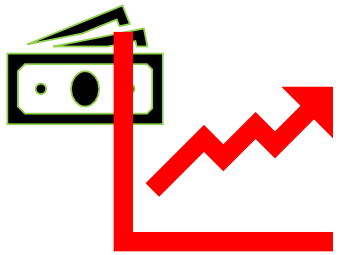
total customer data 4888 entries

EXPAND CUSTOMER BASE

Source : [https://www.kaggle.com/datasets/susant4learning/holiday-package-purchase-prediction]

BACKGROUND

Customers were **contacted at random** without looking at the available information.



Company wants to to make the **marketing expenditure more efficient**.

OBJECTIVES

- Model to predict customer's purchase
- Feature importance
- Business insight

DATA UNDERSTANDING

4888 entries, total 20 columns

CUSTOMER INFORMATION

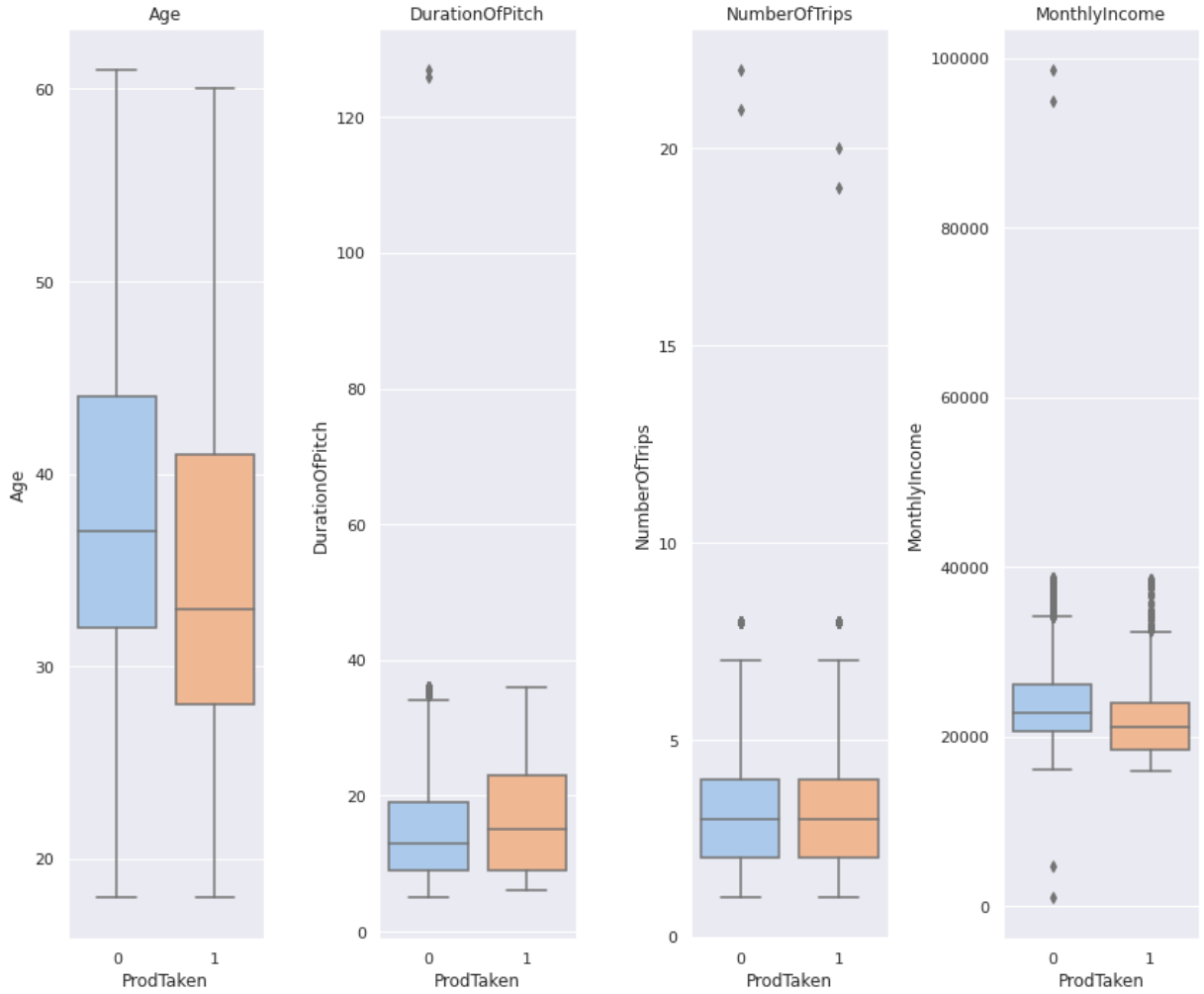
- ID
- **ProdTaken (as a target)**
- TypeofContact
- CityTier (Tier 1 > Tier 2 > Tier 3)
- Occupation
- Gender
- NumberOfPersonVisiting
- PreferredPropertyStar
- MaritalStatus
- NumberOfTrips
- Passport
- OwnCarNumberOfChildrenVisiting
- Designation

CUSTOMER INTERACTION

- DurationOfPitch
- NumberOfFollowups
- ProductPitched
- PitchSatisfactionScore

EXPLORATORY DATA ANALYSIS

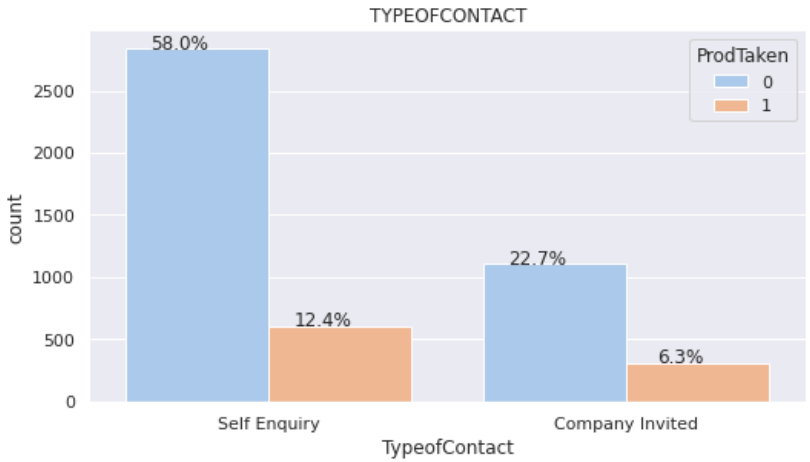
Features distribution with their relation to **Product Taken**

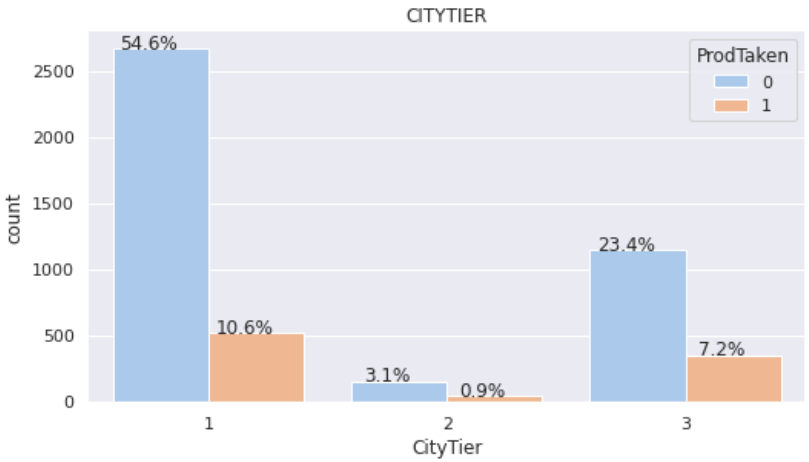


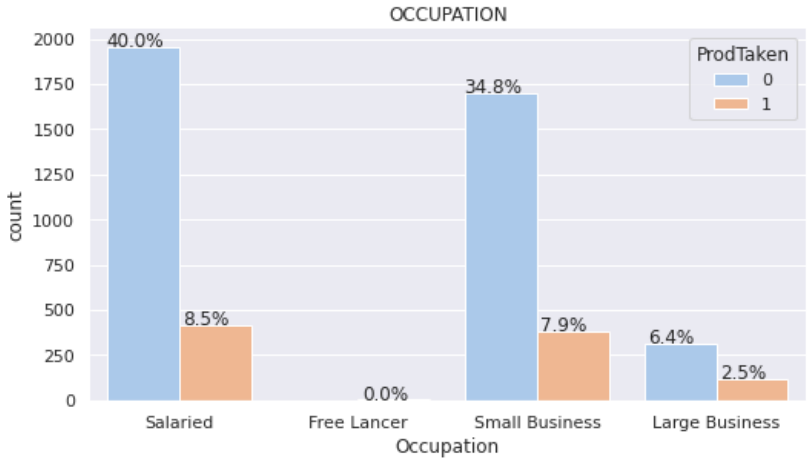
- Product taken mostly by customer at age between 30 to 40, so **the younger customer group (young adult)** are likely to take the product.
- Despite the outlier, customer with **longer duration of pitch** are most lakely to take the product (between 15-20 min)
- **Number of trips are not seen to really affecting** whether customer take the product or not.
- Monthly income are affecting product taken, **customers with lower monthly income are more likely to take the product**, the relation quite similar with age.

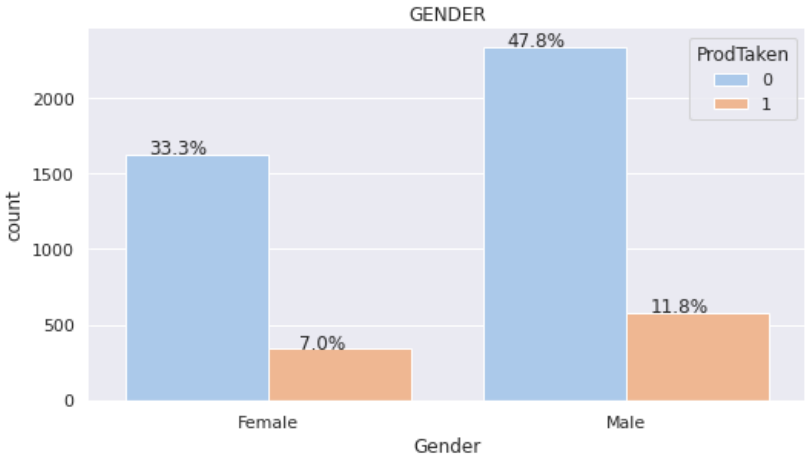
EXPLORATORY DATA ANALYSIS

Features distribution with their relation to **Product Taken**





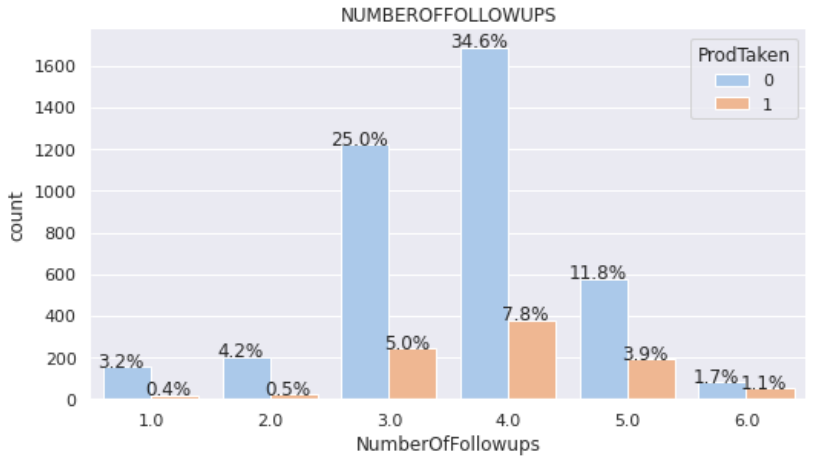


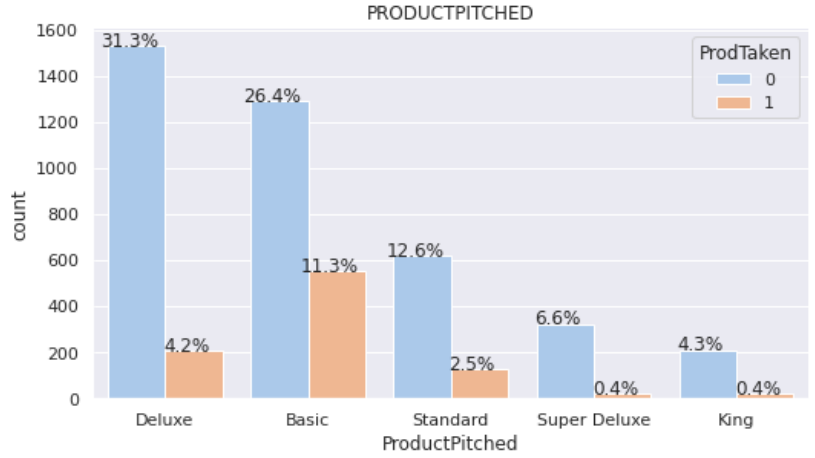


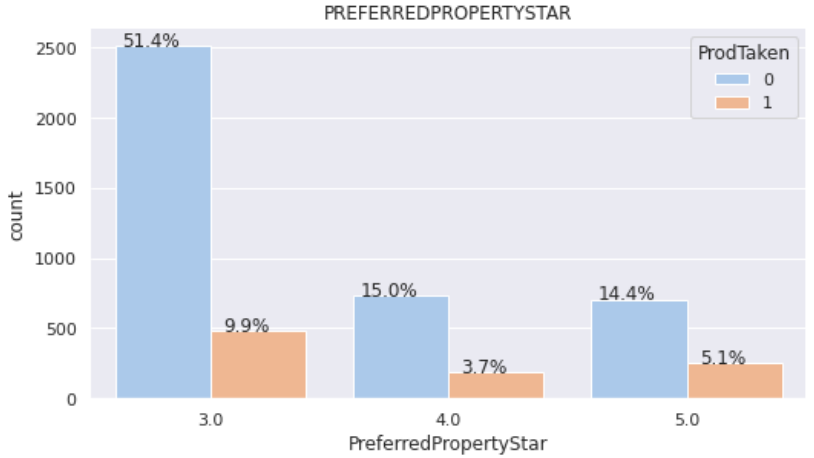
- Product are highly taken by customers with **type of contact self inquiry**.
- Most customers are coming from **city tier 1**.
- Customers who take the product **mostly are Salaried Workers and customers with Small Business**.
- **Female customers** are more likely to take the product.

EXPLORATORY DATA ANALYSIS

Features distribution with their relation to **Product Taken**



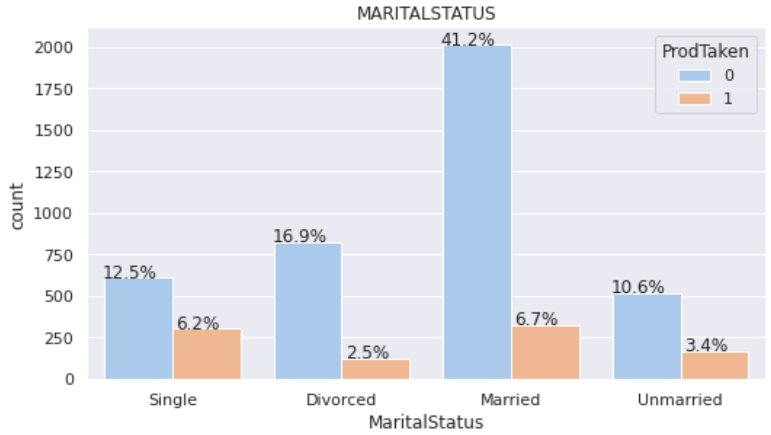


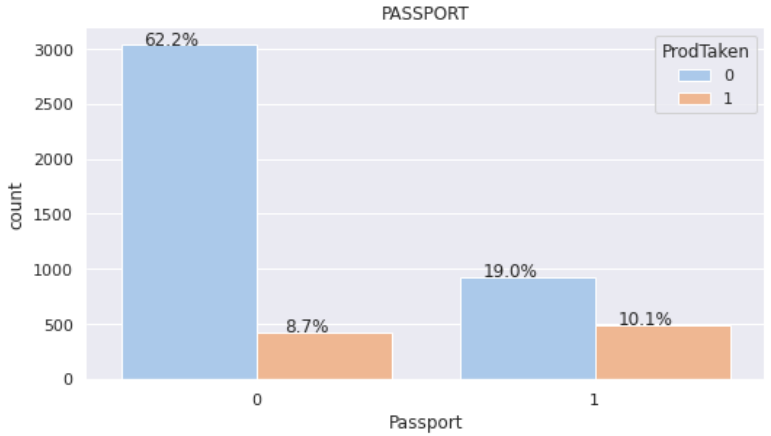


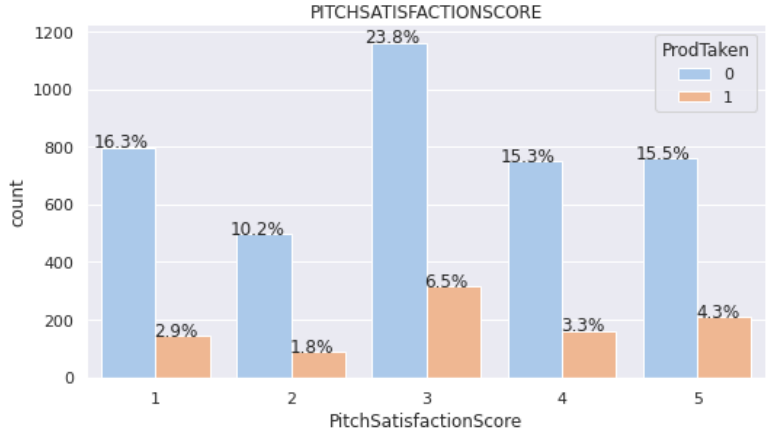
- Most product taken by **customers accompanied by 1-2 person**
- Customers are more likely to take the product after about **3-5 follow ups**
- Most product pitched by the company are deluxe, basic and standard but **the most taken are standard**.
- Customer prefer **property with star rate 3**, and followed next with star rate 5.

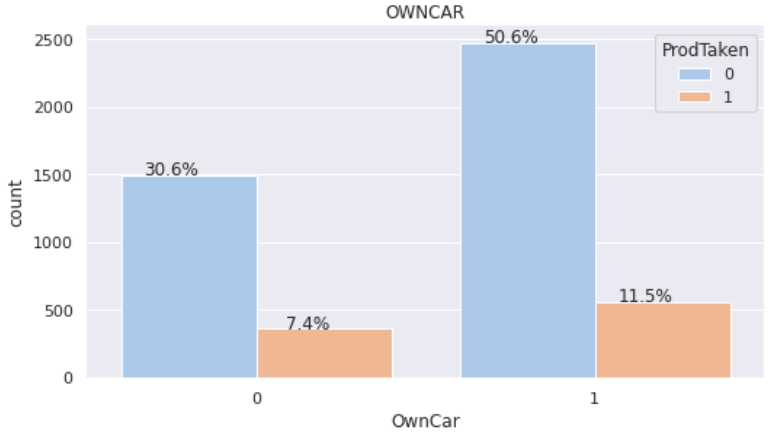
EXPLORATORY DATA ANALYSIS

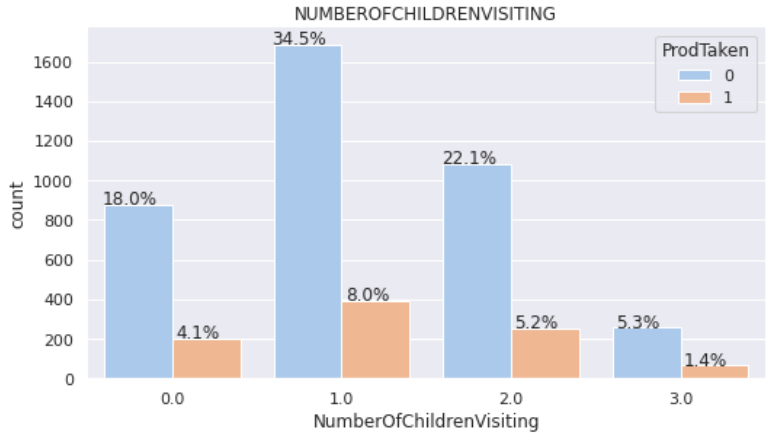
Features distribution with their relation to **Product Taken**

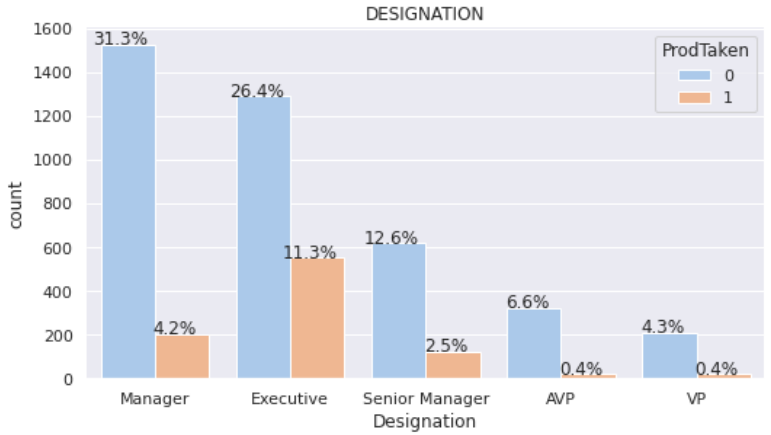












- Most product taken by customers with **marital status single and married**.
- Customers **who have passport** are more likely to take the product
- Most customers take the product with **pitch satisfaction score 3**
- Customers **who own a car** are more likely to take the product.
- Most customers who take the product **visit destination with 1 children**, it's related with customer with married status and number of person visiting 3 (most customers are family with 1 child)
- Customers who take the product mostly are **Executives**.

EXPLORATORY DATA ANALYSIS

Multicollinearity check

- There are no significant or too high correlation from each features.
- Relation between number of person visiting with number of children are explained before.
- Relation between age and monthly income are also explained before
- It is safe to assume we can progress for modelling

Multicollinearity tresshold > 0.8

DATA PREPROCESSING

MISSING VALUES

- **Age**, **MonthlyIncome** : are highly correlated with monthly income, and monthly income are depending on designation of occupation, so we will impute missing values by median with designation as a basis.
- **TypeOfContact** : since Self Inquiry occurring the most, so we will impute missing values with Self Inquiry
- **DurationOfPitch** : it could be there are no pitching done so we will impute missing values with 0
- **NumberOfFollowups**, **PrefferedPropertyStar** : will be imputed by median values of the collumn
- **NumberofTrip** : have relation with age so we will impute this by median with designation as a basis
- **NumberOfChildrenVisiting** : the cause of missing values are probably because there are no children visiting so we will impute missing values with 0

4888 entries, total 20 columns

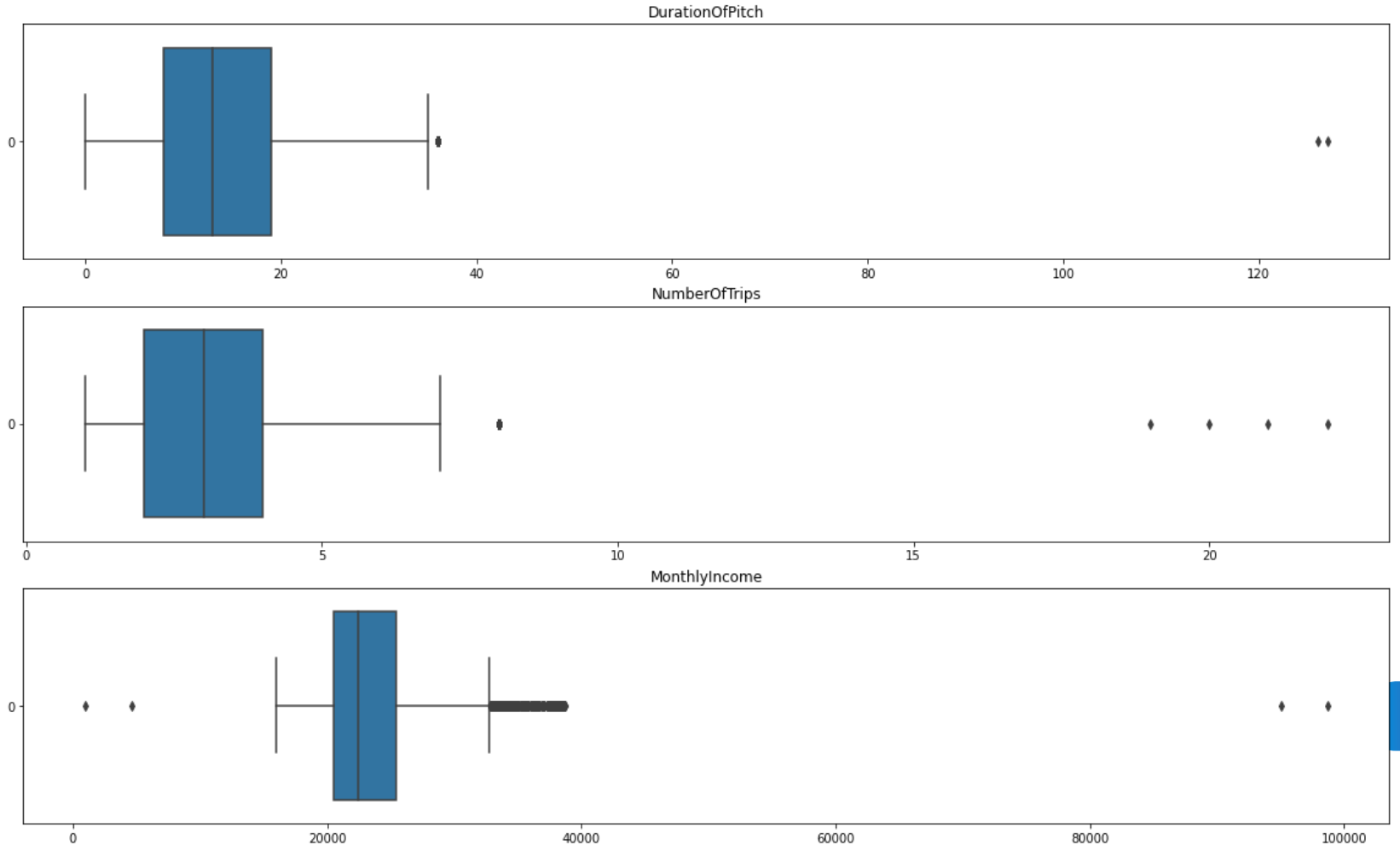
Attributes	Missing Value Count
ProdTaken	0
Age	226
TypeofContact	25
CityTier	0
DurationOfPitch	251
Occupation	0
Gender	0
NumberOfPersonVisiting	0
NumberOfFollowups	45
ProductPitched	0
PreferredPropertyStar	26
MaritalStatus	0
NumberOfTrips	140
Passport	0
PitchSatisfactionScore	0
OwnCar	0
NumberOfChildrenVisiting	66
Designation	0
MonthlyIncome	233

DATA PREPROCESSING

4888 entries, total 20 columns

OUTLIER HANDLING

- Dropping entries with duration of pitch greater than 36 (2 entries)
- Dropping entries with number of trips greater than 8 (4 entries)
- Dropping entries with monthly income less than 13000 or greater than 40000 (4 entries)



MODELLING with

TREE BASED ALGORITHM

- The data type of decision tree can handle any type of data whether it is numerical or categorical
- **adapts quickly to the dataset**
- can handle large datasets efficiently
- **robust to outliers**

METRIC USED

F1-Score : to minimize false positive and false negative

AUC : metric for imbalanced dataset

RESULT

Weighted Decision Tree	Recall	F1-Score	AUC
Before Tuning	0.63	0.68	0.78
After Tuning	0.74	0.54	0.75

Weighted Random Forest	Recall	F1-Score	AUC
Before Tuning	0.53	0.67	0.76
After Tuning	0.61	0.53	0.72

Gradient Boosting	Recall	F1-Score	AUC
Before Tuning	0.35	0.48	0.66
After Tuning	0.44	0.55	0.70

STEPS

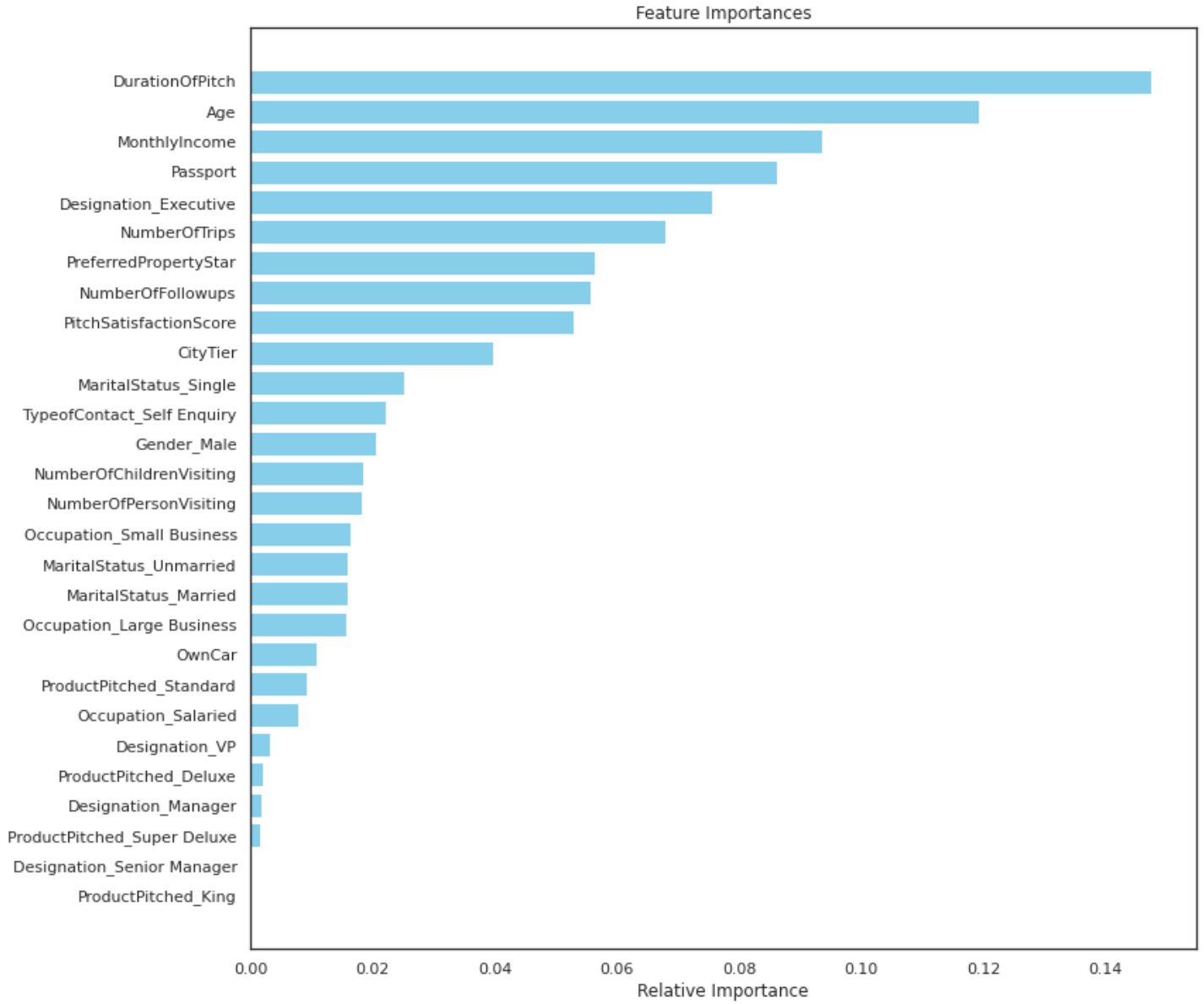
1. Encoding
2. Split Data
3. Modeling
4. Evaluation
5. Hyperparameter Tuning
6. Modeling
7. Evaluation

Feature Importance

Decision Tree

Top 5 Feature Importance

- Duration of Pitch
- Age
- Monthly Income
- Passport
- Designation (Executives)



CONCLUSION

From Data Analysis

- Basic and Deluxe are the most popular packages.
- Number of Children and Number of People visiting does not seem to impact the performance of model a lot.
- Most customers are family with 1 child
- Age and Income have a correlation and we see that young adult age groups and slightly lower Monthly Income are more likely to take the product.
- The Duration of Pitch needs to be more effective and concise.
- Customers with Designation as Executive should be the target customers for the company

RECOMMENDATION

- The marketing team can offer the individual packages to the specific business designation
- The marketing team can create product and customer segment specific sale pitch to reduce the Duration Of Pitch and optimizing number of Follow Ups
- The Wellness Tourism Package should be offered with considering the features importance
- The company can run various campaigns and offers for customers with family to increase sales
- The data shows customers with passport has higher buying ratio and business can curate international packages for such customers, also offering to help customers to make passport could help increasing interest of customers
- Specific packages can be created for different income groups
- The data collection process can be enhanced to capture additional information like package price, which product are actually taken, and satisfaction rate after tour

CONTACT ME

Widya Ayuningtyas

PHONE NUMBER
(+62) 87822945965

EMAIL
widya.ayuningtyas7@gmail.com

LINKEDIN
https://www.linkedin.com/in/widyaayuningtyas7/