**Capstone Project**

**Predict Holiday Package Sales**



|  |  |
| --- | --- |
| Name | : Tee Liang Yuan |
| Cohort | : 19th Jul 2021 |
| Submission Date | : 12th Jan 2022 |

**Contents**

[Background 3](#_Toc93103254)

[Process overview 3](#_Toc93103255)

[Problem statement 3](#_Toc93103256)

[Stakeholders 4](#_Toc93103257)

[Business question 4](#_Toc93103258)

[Data question 4](#_Toc93103259)

[Data 4](#_Toc93103260)

[Data science process 4](#_Toc93103261)

[Data analysis 4](#_Toc93103262)

[Modelling 5](#_Toc93103263)

[Outcomes 5](#_Toc93103264)

[Implementation 5](#_Toc93103265)

[Data answer 6](#_Toc93103266)

[Business answer 6](#_Toc93103267)

[Response to stakeholders 6](#_Toc93103268)

[End-to-end solution 6](#_Toc93103269)

[References 7](#_Toc93103270)

# Background

A travel agency ‘Travel.com’ encountered poor sales performance in the past year, therefore the sales & marketing manager need to address the issue by rolling out a new marketing campaign. The goal is to increase the sales by promoting a new holiday package and at the same time expanding their customer base. The manager likes to know which customer group should be targeted for their new marketing campaign, and how likely the customer will purchase their package. This project aims to find out the prospective customer group and create a sales prediction model.

# Process overview

# Problem statement

* Based on the past record the company had spent 1200 hours on sales pitch but the result is only 18%.
* To address the poor marketing performance, the company hope to spend the resource more on the prospective customer based on analysis.
* Once the right customer group is determined, the company likes to know how the sales performance is?

# Stakeholders

**Sales & marketing manager** who need to plan and execute the marketing campaign.

# Business question

* Which group of customers should be targeted to improve the market efficiency?
* Will the sales increase if we engaged to the targeted customer group?

# Data question

1. What are the variables that affect the product taken the most?
2. Which product is the bestselling?

# Data

* Source: [<https://www.kaggle.com/susant4learning/holiday-package-purchase-prediction>](https://www.kaggle.com/susant4learning/holiday-package-purchase-prediction)
* Size: 4888 rows x 20 columns
* Size after cleaning & feature engineering: 4853 rows x 21 columns

# Data science process

## Data analysis

* The data processing tools are pandas, numpy, and dataprep.
* Additional 2 columns are created which are age group and income group. This is to help categorized the customer for analysis.
* All the nulls’ values are replaced by the means.
* In EDA, there are 5 variables selected which are age group, income group, passport, product pitch, and marital status. The reason these variables had been selected is because they have the smallest different between customer taken or not taken the product. This shows that the success rate of getting a sale closed among this group of customers is higher.

## Modelling

* The target value is ‘Product Taken’.
* There are 5 important features from analysis which are age group, income group, passport, product pitch, and marital status.
* The features that are high correlation to each other is removed from modelling, for example age group and age, income group and monthly income.
* The features that are 0 correlation to target value ‘Product Taken’ is removed from modelling.
* All the non-numeric value is converted to numeric value for imputation.
* The machine learning models that are used is KNN, Random Forest, AdaBoost, XGBoost, LR, and Ensemble Model – Stacking.
* The model took less than 30s to train.
* The model final performance (Ensemble - Stacking) has accuracy 87.6%.
* Ensemble – Stacking model is selected as it has highest accuracy.
* A/B Testing conducted to check if the analysis is valid.

## Outcomes

* From the result of analysis, the group of customers that is more willing to purchase the product has propensity to a younger age, single man/ woman, and income between 10k~20k.
* The basic package is most welcomed by customer.
* The chances to get a deal closed is much higher for customer with a passport.

## Implementation

* More features data should be included like travel destination, travel date, sales date and etc.
* The model should be further developed for sales forecasting.

# Data answer

1. There are 5 important variables which are age group, income group, passport, product pitch, and marital status.
2. Basic package is the most selling product.
3. The customer group is targeted and verified with A/B Testing.

# Business answer

* The sales & marketing performance could be increased by targeting the prospective customer group from analysis.

# Response to stakeholders

* Based on the analysis the prospective customer group to basic package are younger age customer.
* Suggest planning for a family friendly package to customer who has children, and price range between basic and standard package.

# End-to-end solution

* The model could used to as sales forecasting tool to predict future sales, customer interest, and popular destination.

# References

* Data source: https://www.kaggle.com/susant4learning/holiday-package-purchase-prediction
* Code reference:
  + IOD Lab 8.3 Stacking
  + Kaggle\_Dorian: <https://www.kaggle.com/dorianvoydie/eda-modelling-holiday-package-89-acc>
* Jupyter Notebook 6.3.0 for coding
  + Data Pre-processing
  + Modelling
  + A/B Testing
* Libraries:
  + Pandas, Numpy, Dataprep
  + Scipy
  + Matplotlib, Seaborn
  + SKlearn, MLXtend