**Tourism**

**Travel Package Prediction**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Anoop S V**

**Baby Shamin**

**Divya Sukesh**

**Shanija Ragesh**

**Taniya Mary Jaimon**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Nov 2022**

**List of Figures**

**Figures Page No**

figure 1 25

figure 2 27

figure 3 27

**List of Abbreviations**

1. **EDA**- Exploratory Data Analysis
2. **IQR**-Interquartile Range
3. **SMOTE**- Synthetic Minority Oversampling Technique
4. **KNN-** K-Nearest Neighbor
5. **SVM-** Support Vector Machine
6. **XGBoost-** Extreme Gradient Boosting.

**Table of Contents**

1. **Abstract**
2. **Problem Definition**
   1. **Overview**
   2. **Problem Statement**
3. **Introduction**
4. **Literature Survey**
5. **Data Overview**
6. **EDA** 
   1. **Insights from EDA**
7. **Data Preprocessing**
   1. **Outlier Handling**
   2. **Encoding**
   3. **Normalization**
   4. **Over Sampling**
8. **Model Building**
   1. **Classification Models**
9. **Python Flask**
10. **Abstract**

As tourism is one of the fastest growing industries today, thus within the tourism industry events are getting more and more important. It drives economic growth, creates jobs, improves social development, and promotes peace.

A tourism company wants to build a feasible business model that helps you to understand the existing ways of doing the business and how to change the ways for the benefit of the tourism sector. To harness the available data of existing and potential customers to make the marketing expenditure more efficient and to predict which customer is more likely to purchase the newly introduced travel package. The main findings of the study indicate the ways to expand the customer base is to introduce a new offering of packages.

At the end of the study, we are going to predict whether the customer purchased the package or not and the product pitched by the salesperson depending on different features.

**2. Problem Definition**

**2.1 Overview**

To predict which customer is more likely to purchase the newly introduced travel package, which variables are most significant and which segment of customers should be targeted more.

**2.2 Problem Statement**

To utilize the available data of existing and potential customers to make the marketing expenditure more efficient and to predict which customer is more likely to purchase the newly introduced travel package.

One of the ways to expand the customer base is to introduce a new offering of packages. Currently, there are 5 types of packages the company is offering- Basic, Standard, Deluxe, Super Deluxe and King. Looking at the data, we observed that 18% of the customers purchased the packages.

There is a need to analyze the customer’s data and information to provide recommendations to the Policy Maker and the Marketing Team as well as build a model to predict the potential customer who is going to purchase the newly introduced travel package.

**3. Introduction**

In this project, data from a company named “EasyTravel.Com” is used with the goal to make marketing expenditure more efficient. The company is planning to launch a new product, a Travel Package Prediction. This package provides the travelers a warm welcome, in a relaxed and friendly atmosphere to maintain, enhance or kick-start a healthy lifestyle, and support or increase one’s sense of well-being. For this purpose, available data shall be used to predict the potential customer who is going to purchase the newly introduced travel package. Different classification models will be used for this approach.

**4. Literature Survey**

1. **E. Rahmadian, D. Feitosa & A. Zwitter (2021):** Information is essential to the tourism sector. Because of this, research on tourism needs a lot of active, current, timely, and relevant information to assist and aid decision-making processes. Big data can provide current and incredibly informed judgements about behavior and human activity that benefit the travel and tourism sector (Xu & Whitmarsh, 2020). Numerous data sources can be accessed through the tourist to help with decision-making at various stages, including before, during, and after a trip. Social media is one example of a common data source that is used by tourists and tourism firms to communicate, locate or offer pertinent information or any recommendations, and learn about significant news and disasters (Park et al., 2019).
2. **Mathew and Sreejesh(2017)**: In their study, the authors examined the impact of sustainable, responsible tourism on destination sustainability and community quality of life. Thekkady, Kovalam, and Kumarakom were the three main tourist locations in Kerala where data for the study were gathered. The study was carried out in Indian settings in the tourism state of Kerala. The study found that perceived responsible sustainable tourism in terms of economic, social, cultural, and environmental responsibilities strongly influences perceived destination sustainability. This finding was supported by confirmatory factor analysis. The quality of life of locals was found to be impacted by perceived destination sustainability in terms of their material well-being, communal well-being, emotional well-being, health and safety well-being, and overall quality of life.
3. **Hardy and Beeton (2002):** The growth of the environment and the economy have long been linked to sustainable tourism. It must, however, also be connected to community development. Sustainable tourist growth necessitates community involvement. In 2005, Choi and Sirakaya underlined the importance of the host community for sustainable tourism and created a sustainable tourism attitude scale to gauge locals' attitudes toward sustainable travel. Seven components were found using exploratory factor analysis. Environmental sustainability, social costs, economic advantages, community participation, long-term planning, visitor happiness, and a community-centered economy are the contributing aspects.
4. **Hosani, N.A., & Salam, A. (2018):**  Sustainable City Transport Strategies: Dubai from the Viewpoints of Tourists Tourism Research & Hospitality Journal, The study's main goals were to find out how tourists used various forms of transportation, how far they traveled inside the city, and what they thought were the best sustainable transportation policies for Dubai to follow. Techniques - Surveys were carried out between September 15, 2015, and September 20, 2015, for this. Where the following concerns were addressed via questionnaires: 1) To comprehend the mode of transportation chosen by tourists. 2) To determine, if any, whether the tourist's residence and mode of transportation are related. 3) Determine what tourists think about environmentally friendly forms of transportation 4) To gather views from visitors regarding the most sustainable.
5. **Data Overview**

* **CustomerID:** Unique customer ID
* **ProdTaken:** Whether the customer has purchased a package or not (0: No, 1: Yes)
* **Age:** Age of customer
* **TypeofContact:** How customer was contacted (Company Invited or Self Inquiry)
* **CityTier**: City tier depends on the development of a city, population, facilities, and living standards. The categories are ordered i.e., Tier 1 > Tier 2 > Tier 3
* **Occupation:** Occupation of customer
* **Gender:** Gender of customer
* **NumberOfPersonVisiting:** Total number of persons planning to take the trip with the customer
* **PreferredPropertyStar**: Preferred hotel property rating by customer
* **MaritalStatus:** Marital status of customer
* **NumberOfTrips:** Average number of trips in a year by customer
* **Passport:** The customer has a passport or not (0: No, 1: Yes)
* **OwnCar:** Whether the customers own a car or not (0: No, 1: Yes)
* **NumberOfChildrenVisiting:** Total number of children with age less than 5 planning to take the trip with the customer
* **Designation:** Designation of the customer in the current organization
* **MonthlyIncome:** Gross monthly income of the customer
* **PitchSatisfactionScore:** Sales pitch satisfaction score
* **ProductPitched:** Product pitched by the salesperson
* **NumberOfFollowups:** Total number of follow-ups has been done by the salesperson after the sales pitch
* **DurationOfPitch:** Duration of the pitch by a salesperson to the customer

1. **Exploratory Data Analysis(EDA)**

EDA is the most important part of any analysis. We will get to know many things about the data.

EDA steps performed:

* Statistical Summary
* Information about dataset
* Checking duplicates
* Find the data types
* Finding null values
* Filling null values
* Finding the unique values
* Replaced the value
* Univariate analysis
* Correlation Matrix
* Bivariate analysis
* Multivariate analysis

**6.1 Insights from EDA**

**Univariate Analysis**

**Age**:- Normally distributed without any skewness. We see that most customers are in the range of 30- 40 years of age.

**ProdTaken**:- We see that only 18.82% of the total customers purchased any of the travel packages.

**TypeOfContact**:- It is clear from the graph that self-enquiry is more as it is 70.97%.

**DurationOfPitch**:- Duration of pitch is slightly right skewed. Most customer’s pitch duration is below 20 minutes.

**Occupation**:- 48.45% of customers are salaried and very few are freelancers.

**Gender**:- Number of Male customers(59.7%) are higher than Female customers (40.3%).

**NumberOfPersonVisiting**:- 49.1% of customers plan to take at least 3 persons with them during the trip.

**NumberOfFollowups**:- The highest number of follow-ups is 4 followed by 3.

**ProductPitched**:- Basic(37.7%) and Deluxe(35.4%) are the most popular travel packages.

**PreferredPropertyStar**:- 61.8% customers prefer a three star hotel rating compared to four (18.7%) and five (19.6%) star rating hotels.

**MaritalStatus**:- Married customers(47.87%) are more when compared to other marital status.

**NumberOfTrips**:- Number of Trips is right-skewed a little and majority of the customers seem to take at least 2 trips per year.

**PitchSatisfactionScore**:- 30.24% of customers rated 3 star, followed by 19.84% with 5 star.

**NumberOfChildrenVisiting**:- Around 43.9% of customers have at least one child under age Five.

**Bivariate Analysis**

**Age Vs Gender**:- The mean of both the genders are almost of same age.

**TypeofContact Vs ProdTaken**:- 80% of the customers who were contacted through self-enquiry did not purchase a travel package. 80% of the customers who were contacted through company invited did not purchase a travel package.

**CityTier Vs ProdTaken**:- A little above 80% of the customers in Tier 1 city did not purchase a travel package. A little below 80% of the customers in Tier 2 and 3 cities did not purchase a travel package.

**Occupation Vs ProdTaken**:- Though customers who are Freelancers by Occupation have bought travel packages, the sample size is only two. Of the 434 Large Business owning customers, almost 30% bought travel packages. Among Salaried and Small Business owning customers, close to 20% have bought travel packages.

**NumberOfPersonVisiting Vs ProdTaken**:- Those traveling with 1 and 5 other people did not purchase a travel package. 19% of those traveling with 2, 3 and 4 other people did purchase a travel package.

**NumberOfFollowups Vs ProdTaken**:- The more follow-ups a customer pitched, the more likely the customer purchased a travel package. 40% of the customers who had 6 follow ups purchased a travel package.

**ProductPitched Vs ProdTaken**:- Customers who had the basic package pitched had around 30% of them purchase a travel package. While customers who had the super deluxe package pitched only had around 5% of them purchase a travel package.

**PreferredPropertyStar Vs ProdTaken**:- Those who preferred 5 stars had the highest purchase rate at 25%. Those who preferred 3 stars had the lowest purchase rate at 15%.

**MaritalStatus Vs ProdTaken**:- Single customers had the highest purchase rate at 35%. Divorced customers had the lowest purchase rate at 15%.

**PitchSatisfactionScore Vs ProdTaken**:- Majority of customers have given a score of 3.0 to the Sale pitch for the products. But we observe that the number of customers who purchased any product is almost equal across all pitch scores.

**NumberOfChildrenVisiting Vs ProdTaken**:- We see that the percentage of customers who purchased a product is fairly same across all categories of variable NumberOfChildrenVisiting.

**Designation Vs ProdTaken**:- Around 30% Customers with Executive Designation have purchased a product. Sr. Manager - 16% and Manager - 11% Designation customers have purchased a product. Very few customers of VP and AVP Designation have purchased a product.

**Multivariate Analysis**

**MaritalStatus Vs MonthlyIncome and ‘hue’ as ProdTaken**:- Here unmarried people with higher monthly income preferred to take the product.

**Occupation Vs MonthlyIncome and ‘hue’ as ProdTaken**:- In small business and salaried, people with an average monthly income of 20,000 have taken the product.

**DurationOfPitch Vs Occupation and ‘hue’ as ProdTaken**:- The salesperson took more time to convince small business people to take the product.

**7. Data Preprocessing**

Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for building and training Machine Learning models. In simple words, data preprocessing in Machine Learning is a data mining technique that transforms raw data into an understandable and readable format.

As a preprocessing step we did:

* Outlier Handling
* Encoding
* Normalization
* Over Sampling

**7.1 Outlier Handling**

In data analytics, outliers are values within a dataset that vary greatly from the others—they're either much larger, or significantly smaller. Outliers may indicate variabilities in a measurement, experimental errors, or a novelty.

We simply take the difference between the third and first quartiles (Q3 minus Q1). Once we have the IQR, we can use it to detect outliers in our data columns. Using IQR to detect outliers is called the 1.5 x IQR rule. Using this rule, we calculate the upper and lower bounds, which we can use to detect outliers.

Below are some of the methods of treating the outliers

* Trimming/removing the outlier
* Quantile based flooring and capping
* Mean/Median imputation

We found outliers for continuous numerical columns using for loop in a box plot.

As outliers of ‘DurationOfPitch’ are not genuine, we handled the outliers to the 25th and 75th quantile using the clipping method.

**7.2 Encoding**

Machine learning models can only work with numerical values. For this reason, it is necessary to transform the categorical values of the relevant features into numerical ones. This process is called feature encoding.

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important preprocessing step for the structured dataset in supervised learning.

We separated the categorical columns and encoded the same using the Label Encoding method. We used a for loop as we have more than one categorical column.

**7.3 Normalization**

The process of transforming the columns in a dataset to the same scale is referred to as normalization. Every dataset does not need to be normalized for machine learning. It is only required when the ranges of characteristics are different. Normalization avoids raw data and various problems of datasets by creating new values and maintaining general distribution as well as a ratio in data. Further, it also improves the performance and accuracy of machine learning models using various techniques and algorithms.

We dropped all other numerical columns and took continuous numerical variables and did normalization. As the result is in array format, we converted it back to dataframe.

**7.4 Over Sampling**

SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them.

We did sampling in ProdTaken as the data is imbalanced. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model.

**8. Model Building**

Building a model in machine learning is creating a mathematical representation by generalizing and learning from training data. Then, the built machine learning model is applied to new data to make predictions and obtain results.

We did a train test split and took ProdTaken as our target column. Then we did different classification model building methods for finding the score, as our target column is discrete numeric.

**8.1 Classification models**

**Logistic Regression:** Logistic regression is commonly used for prediction and classification problems. From the sklearn module we will use the LogisticRegression() method to create a logistic regression object. Logistic Regression is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.)

**KNN:** K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification of predictive problems in industry.

The following two properties would define KNN well -

Lazy learning algorithm − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.

Non-parametric learning algorithm − KNN is also a non-parametric learning algorithm because it doesn’t assume anything about the underlying data.

**Random Forest:** The random forest classifier can be used to solve regression or classification problems. The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble consists of a data sample drawn from a training set with replacement, called the bootstrap sample.

**SVM:** Support Vector Machine (SVM) is a relatively simple Supervised Machine Learning Algorithm used for classification and/or regression. It is more preferred for classification but is sometimes very useful for regression as well. Basically, SVM finds a hyper-plane that creates a boundary between the types of data. In 2-dimensional space, this hyper-plane is nothing but a line. In SVM, we plot each data item in the dataset in an N-dimensional space, where N is the number of features/attributes in the data. Next, find the optimal hyperplane to separate the data. So by this, you must have understood that inherently, SVM can only perform binary classification (i.e., choose between two classes). However, there are various techniques to use for multi-class problems. Support Vector Machine for Multi-CLass Problems To perform SVM on multi-class problems, we can create a binary classifier for each class of the data.

**Bagging Classifier:** Bootstrap Aggregation, or Bagging for short, is an ensemble machine learning algorithm. Specifically, it is an ensemble of decision tree models, although the bagging technique can also be used to combine the predictions of other types of models

Bagging aims to improve the accuracy and performance of machine learning algorithms. It does this by taking random subsets of an original dataset, with replacement, and fits either a classifier (for classification) or regressor (for regression) to each subset.

**AdaBoost Classifier:** AdaBoost was the first really successful boosting algorithm developed for the purpose of binary classification. AdaBoost is short for Adaptive Boosting and is a very popular boosting technique that combines multiple “weak classifiers” into a single “strong classifier”. It was formulated by Yoav Freund and Robert Schapire. They also won the 2003 Gödel Prize for their work.

**Decision Tree Classifier:** A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical tree structure, which consists of a root node, branches, internal nodes and leaf nodes. Decision Tree is one of the most powerful and popular algorithms. Decision-tree algorithms fall under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

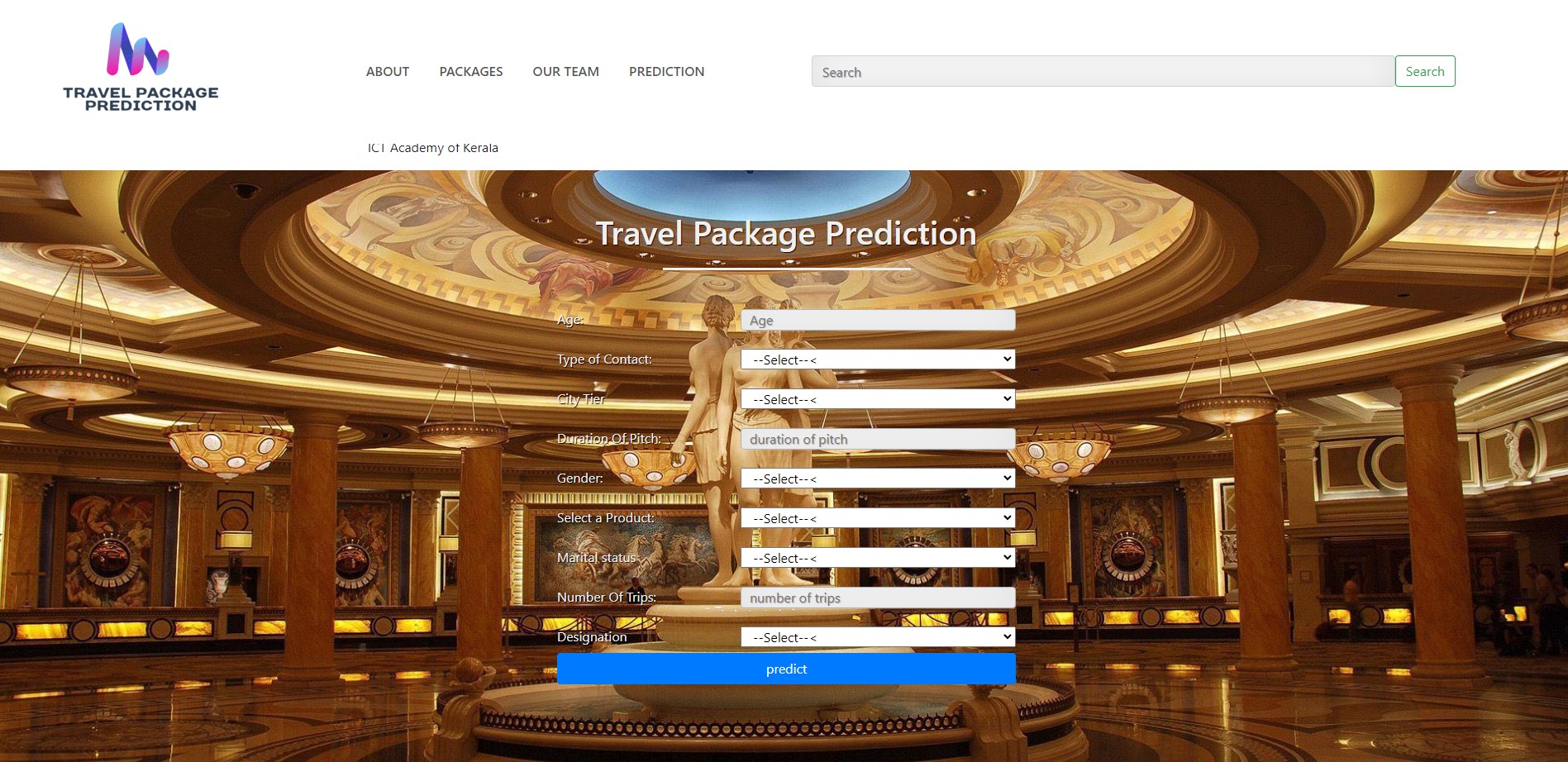
**Gradient Boosting Classifier:** Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient Boosting is a tree-based algorithm, which sits under the supervised branch of Machine Learning.It can be used for both classification and regression problems.

**XGBoost Classifier:** XGBoost is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

**9. Python Flask**

Web application development is made simple with the help of the Python module Flask. Being a microframework, it lacks capabilities like an ORM (Object Relational Manager) and has a compact and simple-to-extend core. It does offer a lot of awesome features, including URL routing and a template engine. This web app framework uses WSGI.

First we created a pickle(pkl) file for the highest accuracy model (Random Forest). Then with the help of ‘html’ and ‘CSS’ we created a web page for predicting whether the customer will accept the package or not.



**Figure 1: Prediction**

**10. Result**

The following table lists the classification models' accuracy scores:

|  |  |
| --- | --- |
| Classification models | Accuracy Score (%) |
| Logistic Regression | 63.3 |
| KNN | 77.1 |
| Random Forest Classifier | 88.7 |
| SVM | 62.5 |
| Bagging Classifier | 87.1 |
| AdaBoost Classifier | 68.7 |
| Decision Tree Classifier | 87 |
| Gradient Boosting Classifier | 72.9 |
| XGBoost Classifier | 72.8 |

### Feature and Model

* The features we use for modeling after selecting there are 9: Age, TypeofContact, CityTier, DurationOfPitch, Gender, ProductPitched, MaritalStatus, NumberOfTrips, Designation.
* The Best model that we choose is Random Forest.

The predicted result is shown as below:

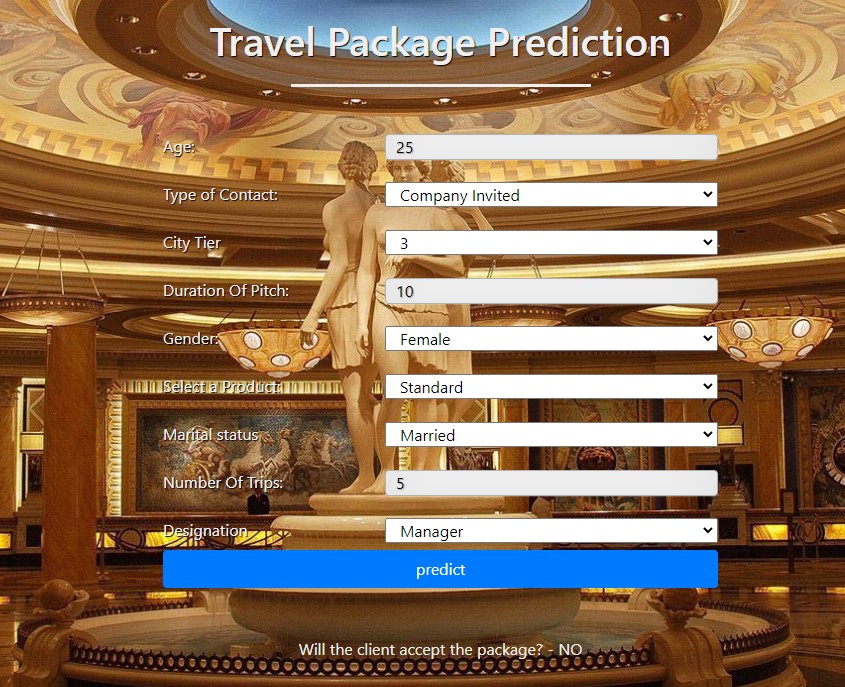


Figure 2: when predicted ‘NO’

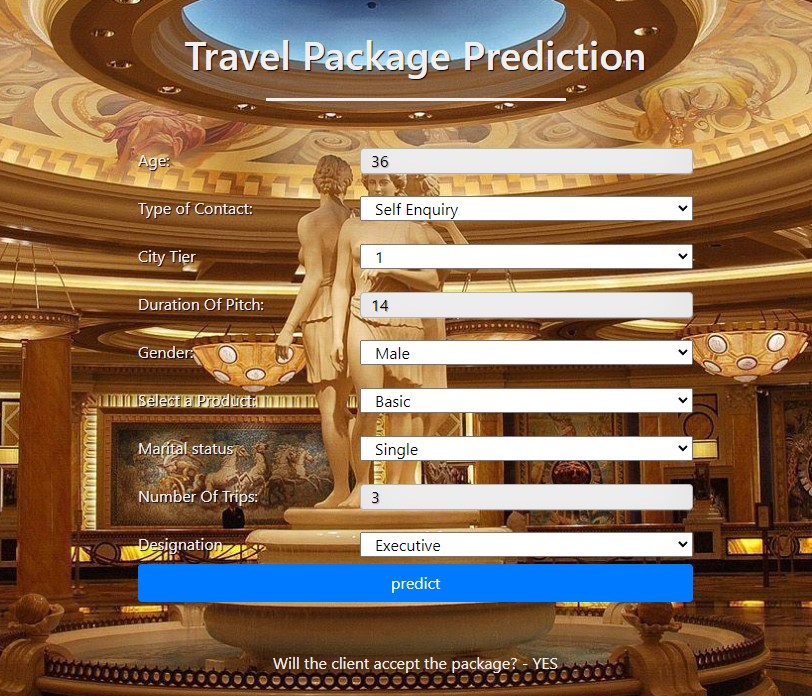


Figure 3: When predicted ‘YES’

**11. Conclusion**

After the modeling we took the model with highest accuracy (Random Forest- 87.9%). With this model we created a pickle file for web page implementation using python flask. After entering the customer details, the model will predict whether the customer will accept the package or not.

**Hosted Webpage URL-** <http://dsaproject5.pythonanywhere.com>

**References**

1. [**https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/**](https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/)
2. [**https://www.geeksforgeeks.org/numpy-clip-in-python/**](https://www.geeksforgeeks.org/numpy-clip-in-python/)
3. [**https://www.kaggle.com/datasets/riteshvsharma/tourismxlsx**](https://www.kaggle.com/datasets/riteshvsharma/tourismxlsx)
4. [**https://www.geeksforgeeks.org/create-a-stacked-bar-plot-in-matplotlib/**](https://www.geeksforgeeks.org/create-a-stacked-bar-plot-in-matplotlib/)
5. [**https://stackoverflow.com/questions/25447700/annotate-bars-with-values-on-pandas-bar-plots**](https://stackoverflow.com/questions/25447700/annotate-bars-with-values-on-pandas-bar-plots)
6. [**https://www.geeksforgeeks.org/python-pearsons-chi-square-test/**](https://www.geeksforgeeks.org/python-pearsons-chi-square-test/)
7. [**https://www.researchgate.net/publication/327117594\_India\_-\_Sustainability\_and\_the\_Tourism\_Rankings**](https://www.researchgate.net/publication/327117594_India_-_Sustainability_and_the_Tourism_Rankings)
8. **https://www.kaggle.com/code/abhimanyukspillai/demand-forecasting-simple**