**Detailed Project Report**

**TRAVEL PACKAGE PURCHASE PREDICTION**

|  |  |
| --- | --- |
| Written By | Aadil Hussain Teeli |
| Version | 1.0 |
| Date | 24-02-2024 |

**Document Change Control Record**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Reviews

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Reviewer** | **Comments** |
|  |  |  |  |

### Approval Status:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Version** | **Review date** | **Reviewed By** | **Approved By** | **Comments** |
|  |  |  |  |  |

# **Introduction**

# 

# **1.1 Abstract**

The "Travel Purchase Package Prediction" project revolves around a binary classification problem within the realms of travel and data science. Specifically designed to forecast travel package purchases, the system harnesses historical data encompassing customer interactions and purchase behaviors. The primary goal is to categorize customers into two distinct groups: those likely to purchase the travel package and those not likely to do so.

Drawing valuable insights from past purchase records and various customer attributes, the system employs a diverse range of machine learning algorithms. These algorithms meticulously identify intricate patterns and relationships within the data, contributing to the development of a robust binary classification model. The model's efficacy lies in its ability to accurately classify customers, providing businesses with insights to tailor marketing strategies and optimize sales efforts.

In essence, this project revolutionizes customer engagement and revenue generation in the travel industry by empowering businesses to proactively address the diverse needs and preferences of their customer base. The outcome is a sophisticated binary classification model that enhances the precision of predictions, serving as a strategic tool for businesses seeking to elevate their marketing initiatives and maximize sales in the dynamic landscape of the travel sector.

**1.2 Machine Learning**

The data available is increasing day by day and such a huge amount of unprocessed data is needed to be analyzed precisely, as it can give very informative and finely pure gradient results as per current standard requirements. It is not wrong to say as with the evolution of Artificial Intelligence (AI) over the past two decades, Machine Learning (ML) is also on a fast pace for its evolution. ML is an important mainstay of IT sector and with that, a rather central, albeit usually hidden, part of our life. As the technology progresses, the analysis and understanding of data to give good results will also increase as the data is very useful in current aspects.

In machine learning, one deals with both supervised and unsupervised types of tasks and generally a classification type problem accounts as a resource for knowledge discovery. It generates resources and employs regression to make precise predictions about future, the main emphasis being laid on making a system self-efficient, to be able to do computations and analysis to generate much accurate and precise results. By using statistic and probabilistic tools, data can be converted into knowledge. The statistical inferencing uses sampling distributions as a conceptual key.

ML can appear in many guises. In this paper, firstly, various applications of ML and the types of data they deal with are discussed. Next, the problem statement addressed through this work is stated in a formalized way.

## Problem Statement

In the fast-growing world of tourism, predicting whether people will buy travel packages has become really important. The tourism industry can have unpredictable changes in demand, so being able to accurately predict if customers will purchase travel packages helps organizations plan better. The main goal here is to figure out if a customer is likely to buy a travel package or not. This information is valuable for tourism businesses to make smart decisions and adapt to the changing needs of travelers.

**2. Architecture:**

Following architecture was followed during project development:

Start

Data gathering

Data Cleaning

Handling Missing Data

Parameter tuning

Model building

Model saving

End

Feature Generation

Deployment

Export into csv

Push to GitHub

Flask setup

Encoding Categorical Data

New feature creation

**2.1 Data gathering:**

Data source: <https://question.transtutors.com/6129343_1_tourism-data.xlsx>

Train and Test data are stored in .csv format.

**2.2 Raw Data Validation:**

Following the loading of data, it is essential to conduct various validations before proceeding with any further operations. These validations include checking for zero standard deviation across all columns and identifying columns with complete missing values. These checks are imperative because attributes exhibiting these characteristics are deemed useless and do not contribute to the sales of items from respective outlets.

For instance, if an attribute displays zero standard deviation, it implies that all values are the same, with a mean of zero. This suggests that regardless of whether the sales are increasing or decreasing, that attribute will remain constant. Similarly, if an attribute contains entirely missing values, including it in operations serves no purpose and unnecessarily increases the risk of the curse of dimensionality.

**2.3 Data Transformation**

Prior to sending the data into the database, it is necessary to undergo data transformation to convert it into a format suitable for easy insertion into the database. Notably, the 'Age,' 'Duration of Pitch,' and 'Monthly Income' attributes exhibit missing values. Therefore, in both the train set and the test set, these missing values are filled with appropriate data types to ensure completeness and compatibility with the database structure.

**2.4 Data preprocessing**

Before constructing the model, extensive pre-processing was conducted on the customer data. The handling of missing values took into account data type and distribution; for instance, numerical features were imputed with the mean, categorical features were filled with the mode, and potentially removed if deemed significant. Correction or removal of invalid values depended on their severity and impact. Outliers were identified, and their influence on the analysis led to their addressing and removal. Additionally, feature scaling and normalization were implemented to guarantee uniform scales for all features, thereby enhancing the efficiency of the model building process.

**2.5 Feature Engineering:**

Feature Engineering was conducted after the pre-processing stage, revealing that certain attributes were deemed unimportant for the specific outlet. Consequently, these irrelevant attributes were removed from the dataset. Additionally, one-hot encoding was implemented to transform categorical features into numerical features, enhancing their compatibility for subsequent analysis and modeling.

**2.6 PipeLining:**

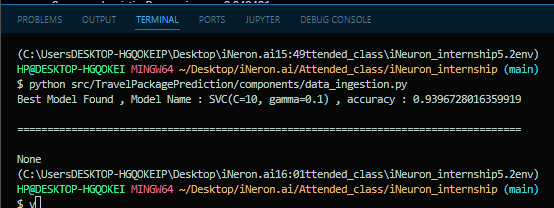
During the pre-processing phase of my project, distinct pipelines were set up for handling numerical and categorical features separately. The numerical pipeline is designed to address tasks such as imputation and scaling, while the categorical pipeline employs methods such as one-hot encoding. This customized approach ensures that each type of feature undergoes suitable processing, thereby optimizing both the performance and interpretability of the model. The implementation of these pipelines streamlines the pre-processing procedures, facilitating the efficient transformation of the dataset and ultimately improving the accuracy of our predictive models.

**2.7 Parameter tunning:**

The tuning of parameters was carried out using GridSearchCV, where various algorithms including Logistic Regression, Decision Tree SVM, Random Forest, Naive\_Bayes\_Classifier, KNeighborsClassifier and many more algorithms were employed for solving the problem. The parameters of these algorithms were fine-tuned and incorporated into the models. Remarkably, the SVC emerged as the most effective, yielding a training accuracy of 99% and a testing accuracy of 93.9%.

**2.8 Model building:**

After completing the comprehensive preprocessing operations outlined earlier, including scaling and hyperparameter tuning, the dataset underwent evaluation using different models. Among them, the Support Vector Classifier (SVC) exhibited the most favorable performance, achieving a remarkable training accuracy of 99% and a testing accuracy of 93.9%. These results indicate that the SVC model excelled in addressing the complexities of the given problem and demonstrated robust performance, particularly in accurately classifying instances in both the training and testing datasets.



**2.10 Model saving:**

Model is then saved using pickle library in .pkl format.

**2.11 Git Hub**

Whole project directory will be pushed into GitHub repository.

**2.12 Deployment:**

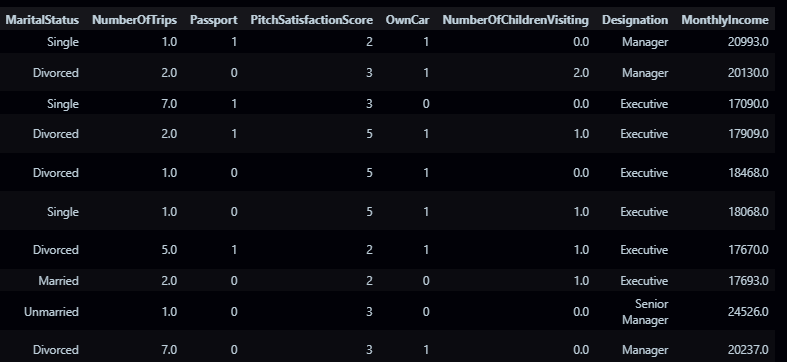
Cloud environment was set up and project was deployed form GitHub into AWS cloud platform.

App link https://vxh7bvmiaw.eu-west-3.awsapprunner.com/predictdata

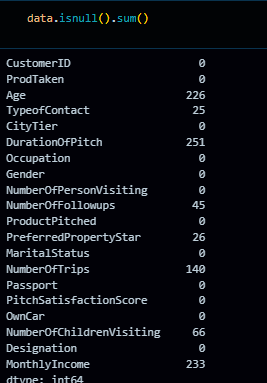
**3. Data set description**

The dataset for Travel Package Purchase Prediction contains 4888 observations and 20 features. Each observation represents a potential customer, while the features provide information about various aspects related to travel package purchases. The columns in the dataset are described as follows:



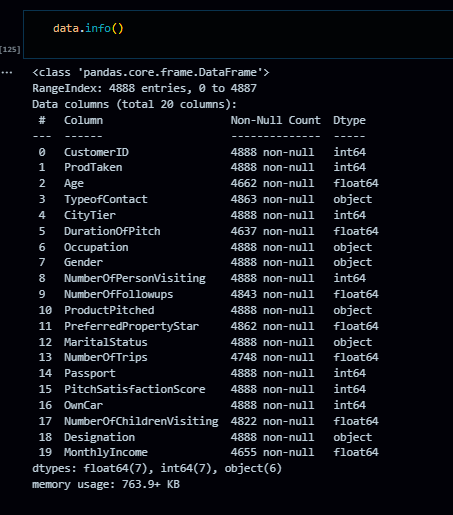


The data Set has various missing values:

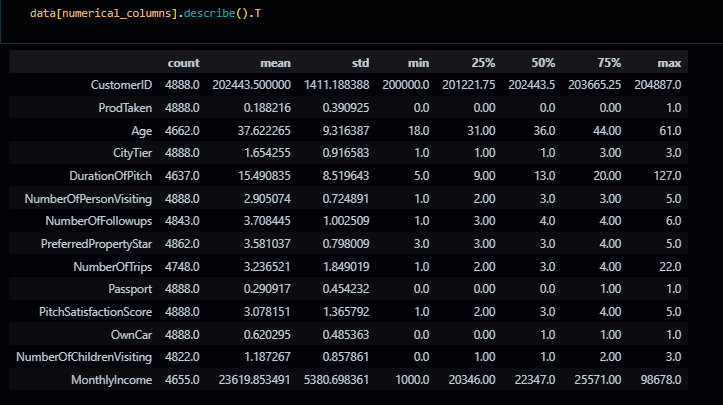


The data set consists of various data types from integer to floating to object as shown in Fig.In the raw data, there can be various types of underlying patterns which also gives an in-depth knowledge about the subject of interest and provides insights into the problem. But caution should be observed

with respect to data as it may contain null values, or redundant values, or various types of ambiguity, which also demands pre-processing of data. The dataset should therefore be explored as much as possible.



Various factors important by statistical means like mean, standard deviation, median, count of values and maximum value, etc. are shown below for numerical attributes.



The preprocessing steps for this dataset involve an in-depth analysis of the independent variables. This includes identifying and addressing null values in each column, replacing them with suitable data types to ensure a seamless flow in analysis and model fitting without compromising accuracy. The visual representations provided above utilize Pandas tools to showcase variable counts for numerical columns and mode values for categorical columns.

Examining numerical columns involves assessing maximum and minimum values, as well as percentile values for the median. This information is pivotal in determining the priority for selecting values for subsequent exploration tasks and analysis. Additionally, understanding the data types of various columns becomes crucial for label processing and implementing a one-hot encoding scheme during the model-building phase.

# **4. Implementation and Results**

In this section we will be discusing, the programming language, libraries, implementation platform along with the data modeling and the observations and results obtained.

## 4.1 Implementation Platform and Language

## Python stands out as a versatile, interpreted high-level language widely employed for addressing domain-specific problems, allowing users to focus on problem-solving rather than grappling with system intricacies. Often referred to as the 'batteries included language' for programming, Python boasts an array of libraries tailored for scientific computations and research, complemented by numerous third-party libraries that enhance problem-solving efficiency.

## This project leverages prominent Python libraries, including NumPy for scientific computation, Seaborn and Matplotlib for 2D plotting, and the Pandas tool for comprehensive data analysis. The choice of these libraries reflects a commitment to leveraging robust tools that streamline scientific tasks and data exploration

## For the development environment, Visual Studio Code (VS Code) has been selected. Recognized for its excellence in 'literate programming,' VS Code seamlessly integrates human-friendly code into code blocks, providing an exceptional platform for development. This choice underscores a dedication to a development environment that prioritizes readability and efficiency in code creation and analysis.

## 4.3 Metrics for Data Modelling

* For the Travel Package Purchase Prediction project, which involves classification tasks, the following metrics are more relevant:
* **Accuracy Score**: Measures the proportion of correctly predicted labels out of the total number of instances. It is a fundamental metric for evaluating classification models.
* **F1 Score**: Represents the harmonic mean of precision and recall. It provides a balanced measure of the model's accuracy, especially in scenarios with imbalanced class distributions.
* **Classification Report**: Provides a comprehensive overview of the model's performance, including precision, recall, and F1 score for each class. It helps in understanding the model's ability to classify instances correctly across different classes.

## 4.4 Prediction results

## The analysis uncovered that specific customer demographics and characteristics played a crucial role in predicting travel package purchases. Notably, features such as 'Age', 'TypeofContact', 'CityTier', 'Occupation', 'Gender', 'PreferredPropertyStar', 'MaritalStatus', 'Passport', 'PitchSatisfactionScore', 'OwnCar', 'NumberOfChildrenVisiting', 'Designation', and 'MonthlyIncome' exhibited varying degrees of significance in determining whether a customer would choose a travel package.

## Further investigation into the model's performance highlighted the influential role of certain features in prediction outcomes. For instance, 'PreferredPropertyStar' and 'NumberOfTrips' emerged as particularly impactful, suggesting that customer preferences for property type and past travel experiences significantly influenced their decision-making process.

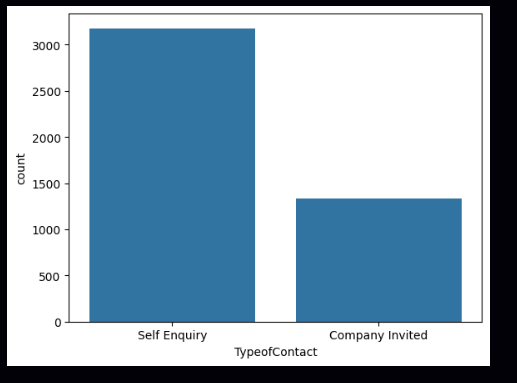
## Additionally, the classification model demonstrated strong performance metrics, with high accuracy and F1-score values. These results signify the model's effectiveness in precisely categorizing customers into purchase and non-purchase segments. The robust performance suggests that the model successfully captured intricate patterns and relationships within the data, enabling accurate predictions of travel package purchases.

## In summary, these findings underscore the significance of utilizing customer demographic and behavioral data for predicting travel package purchases. Through a comprehensive understanding and analysis of these factors, businesses can tailor their marketing strategies and product offerings to effectively target and engage potential customers. This, in turn, can lead to increased sales and revenue growth within the travel industry.

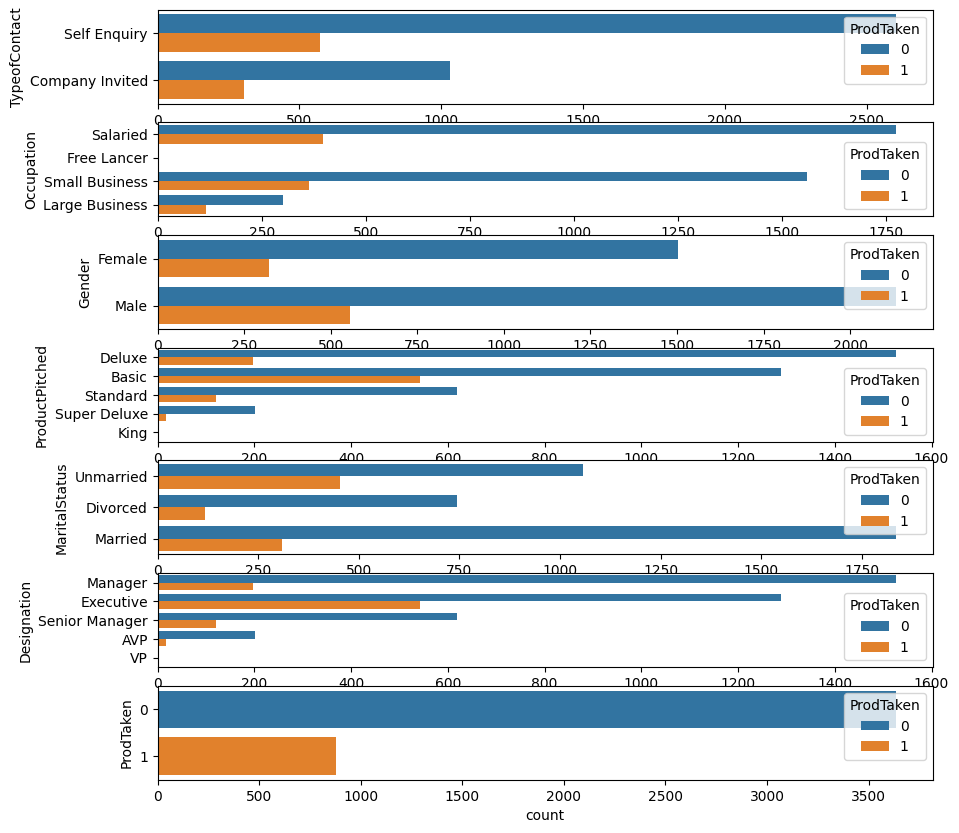
## 5. Conclusion

The objective of the Travel Package Purchase Prediction project is to offer valuable insights to travel companies, aiding them in refining marketing strategies, improving customer targeting, and optimizing revenue generation. Through the analysis of past customer behavior and travel package characteristics, our goal is to build a strong machine learning model capable of precisely forecasting future demand for travel packages. This predictive system is designed to enable businesses to make well-informed decisions regarding product offerings and resource allocation, while also streamlining supply chain management processes. By leveraging historical data and advanced machine learning techniques, we aim to foster growth and efficiency in the travel industry, delivering tangible value to stakeholders.

Our findings underscore the importance of understanding customer preferences and behaviors. Notably, we've observed that the "Self Enquiry" option is favored by the majority of customers.



I want to high light the important obseravtions here, that we have drawn from the data:



* TypeofContact: This variable shows that customers with a certain type of contact are more likely to have taken the product.
* Occupation: some occupations have a higher number of customers who have taken the product compared to others, like salaried and small Bussiness.
* Gender: The number of men and women buying the product is different.
* Product Pitched: How the product is presented or talked about affects if people will buy it.
* Marital Status: Being married or single affects how likely someone is to buy the product.
* Designation: Some titles or roles of people have make them more likely to buy the product.
* Car Ownership: People who own cars are more likely to buy the product compared to those who don't.

## Future Scope

The Travel Package Purchase Prediction project offers a solid foundation for growth and improvement, presenting various opportunities for innovation. Here's how we can enhance the project's success:

1. **Continuous Feature Evolution:**Implement ongoing updates and enhancements to feature engineering techniques, adapting to shifting customer preferences and market dynamics. This approach ensures the model remains agile and accurate over time.
2. **Integration of External Data Streams:**Explore the integration of diverse external data sources, such as social media feeds, economic indices, or emerging travel trends. By incorporating this data, we can enrich insights and bolster prediction accuracy.
3. **Visualization of Feature Importance:**Develop interactive visualization tools or dashboards to showcase feature importance rankings and model predictions. This visualization aids stakeholders in comprehending the model's insights and facilitates decision-making processes.
4. **Customer Segmentation Analysis:**Conduct thorough customer segmentation analyses based on classification predictions. This analysis uncovers nuanced customer patterns, empowering targeted marketing strategies and personalized customer experiences.
5. **Integration of User Feedback:**Incorporate mechanisms for gathering and integrating user feedback into model refinement processes. This feedback loop enables adjustments based on real-world user experiences and preferences.

**7. Q & A:**

**Q1) What’s the source of data?**

Ans. The data for training is provided by the client from:

<https://question.transtutors.com/6129343_1_tourism-data.xlsx>

**Q 2) What was the type of data?**

Ans. The data was the combination of numerical and Categorical values.

**Q 3) What’s the complete flow you followed in this Project?**

Ans. Refer the Architecture section for this.

**Q 5) What techniques were you using for data pre-processing?**

* + Removing unwanted attributes
  + Visualizing relation of independent variables with each other and output variables
  + Checking and changing Distribution of continuous values
  + Removing outliers
  + Cleaning data and imputing if null values are present.
  + Converting categorical data into numeric values.
  + Scaling the data

**Q 6) How training was done or what models were used?**

* Before diving the data in training and validation set we performed clustering over fit to divide the data into clusters.
* As per cluster the training and validation data were divided.
* The scaling was performed over training and validation data
* Algorithms like logistic Regression ,Decision tree, XGBoost, SVM, AdaBoost,Random Forest etc.

**Q 7) How Prediction was done?**

Ans. The testing files are shared by the client. We pass its data to the best model which we have saved in pickle format and get the prediction.

**Q 8) Where the model was deployed?**

Ans. When the model is ready, we deploy it in AWS platform. This model is an web application where user can enter the data and these data gets extracted in the backend and user gets the prediction result.