Travel Package Purchase Prediction - Problem Statement

**Submission type**

:

File Upload

**Due Date**

:

Aug 28, 8:30 AM

**Total Score**

:

60

**Available from**

:

Aug 11, 9:30 PM

**Description**

**Background and Context**

You are a Data Scientist for a tourism company named "Visit with us". The Policy Maker of the company wants to enable and establish a viable business model to expand the customer base.

A viable business model is a central concept 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.

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, King. Looking at the data of the last year, we observed that 18% of the customers purchased the packages.

However, the marketing cost was quite high because customers were contacted at random without looking at the available information.

The company is now planning to launch a new product i.e. Wellness Tourism Package. Wellness Tourism is defined as Travel that allows the traveler to maintain, enhance or kick-start a healthy lifestyle, and support or increase one's sense of well-being.

However, this time company wants to harness the available data of existing and potential customers to make the marketing expenditure more efficient.

You as a Data Scientist at "Visit with us" travel company have to analyze the customers' data and information to provide recommendations to the Policy Maker and Marketing Team and also build a model to predict the potential customer who is going to purchase the newly introduced travel package.

**Objective**

To predict which customer is more likely to purchase the newly introduced travel package.

**Data Dictionary**

**Customer details:**

1. CustomerID: Unique customer ID
2. ProdTaken: Whether the customer has purchased a package or not (0: No, 1: Yes)
3. Age: Age of customer
4. TypeofContact: How customer was contacted (Company Invited or Self Inquiry)
5. 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
6. Occupation: Occupation of customer
7. Gender: Gender of customer
8. NumberOfPersonVisiting: Total number of persons planning to take the trip with the customer
9. PreferredPropertyStar: Preferred hotel property rating by customer
10. MaritalStatus: Marital status of customer
11. NumberOfTrips: Average number of trips in a year by customer
12. Passport: The customer has a passport or not (0: No, 1: Yes)
13. OwnCar: Whether the customers own a car or not (0: No, 1: Yes)
14. NumberOfChildrenVisiting: Total number of children with age less than 5 planning to take the trip with the customer
15. Designation: Designation of the customer in the current organization
16. MonthlyIncome: Gross monthly income of the customer

**Customer interaction data:**

1. PitchSatisfactionScore: Sales pitch satisfaction score
2. ProductPitched: Product pitched by the salesperson
3. NumberOfFollowups: Total number of follow-ups has been done by the salesperson after the sales pitch
4. DurationOfPitch: Duration of the pitch by a salesperson to the customer

**Note:**

Please note XGBoost can take a significantly longer time to run, so if you have time complexity issues then you can avoid tuning XGBoost. No marks will be deducted if XGBoost tuning is not attempted.

**Best Practices for Notebook :**

* The notebook should be well-documented, with inline comments explaining the functionality of code and markdown cells containing comments on the observations and insights.
* The notebook should be run from start to finish in a sequential manner before submission.
* It is preferable to remove all warnings and errors before submission.
* The notebook should be submitted as an HTML file (.html) and as a notebook file (.ipynb)

**Submission Guidelines :**

1. There are two parts to the submission:
   1. A well commented Jupyter notebook [format - .ipynb]
   2. File converted to HTML format
2. Any assignment found copied/ plagiarized with other groups will not be graded and awarded zero marks
3. Please ensure timely submission as any submission post-deadline will not be accepted for evaluation
4. Submission will not be evaluated if,
   1. it is submitted post-deadline, or,
   2. more than 2 files are submitted

Happy Learning!!

**Scoring guide (Rubric) - Travel Package Purchase Prediction**

| **Criteria** | **Points** |
| --- | --- |
| **Perform an Exploratory Data Analysis on the data**  - Univariate analysis - Bivariate analysis - Use appropriate visualizations to identify the patterns and insights - Come up with a customer profile (characteristics of a customer) of the different packages - Any other exploratory deep dive | 8 |
| **Illustrate the insights based on EDA**  Key meaningful observations on individual variables and the relationship between variables | 5 |
| **Data Pre-processing**  - Prepare the data for analysis - Missing value Treatment, - Outlier Detection(treat, if needed- why or why not ), - Feature Engineering, - Prepare data for modeling | 8 |
| **Model building - Bagging**  - Build Bagging classifier, Random Forest, and Decision Tree. - Comment on model performance | 6 |
| **Model performance improvement - Bagging**  - Comment on which metric is right for model performance evaluation and why? - Comment on the model performance after tuning the Decision Tree, Bagging, and Random Forest classifier to improve the model performance. | 9 |
| **Model building - Boosting**  - Build Adaboost, GradientBoost, XGBoost, and Stacking classifiers - Comment on model performance | 6 |
| **Model performance improvement - Boosting**  - Comment on which metric is right for model performance evaluation and why? - Comment on the model performance after tuning the AdaBoost, and Gradient Boosting classifier on the appropriate metric to improve the model performance. \* Please note XGBoost can take a significantly longer time to run, so if you have time complexity issues then you can avoid tuning XGBoost. | 7 |
| **Actionable Insights & Recommendations**  - Compare model performance on various metrics. - Conclude with the key takeaways - What would your advice be to grow the business? | 6 |
| **Notebook - Overall**  - Structure and flow - Well commented code | 5 |
| Points | 60 |

#### **1. How should one approach the Travel Package Purchase Prediction project?**

* Before starting the project, please read the problem statement carefully and go through the criteria and descriptions mentioned in the rubric.
* Once you understand the task, download and import the dataset into a Jupyter notebook to get started with the project.
* To work on the project, you should start with an exploratory analysis of the data. Try to use descriptive statistics and visualization to understand the data.
* Once the EDA is completed, you should start building possible models (tuned and untuned) and test for corresponding performance metrics to find the best-performing model.
* It is important to close the analysis with key findings and recommendations to the business.

#### **2. Is the “Fe male” entry the same as “Female” in the Gender column? If yes, how to treat them?**

Yes, the entries “Fe male” and “Female” are the same. Considering it as a data entry error, it is recommended to convert “Fe male” into “Female”.

#### **3. I am getting the below error with the Decision Tree classifier:**

ValueError: could not convert string to float: 'Self Enquiry'

#### **How to solve it?**

As decision trees cannot handle non-numeric data, it is preferable to either encode or create dummy variables for the categorical features.

#### **4. How to install XGBoost on Windows?**

To install XGBoost on Windows, use “!pip install xgboost” in Jupyter notebook or “conda install -c anaconda py-xgboost” in the Anaconda prompt in Anaconda.

Before this, it must be ensured that previous versions of XGBoost are uninstalled by using the command "!pip uninstall xgboost".

#### **5. How to install XGBoost on Mac?**

To install XGBoost on Mac, run the following command in terminal “conda install -c conda-forge xgboost”.

In some cases, trying to install XGBoost may return the following error:

XGBoostError: XGBoost Library (libxgboost.dylib) could not be loaded.

Likely causes:

\* OpenMP runtime is not installed (vcomp140.dll or libgomp-1.dll for Windows, libomp.dylib for Mac OSX, libgomp.so for Linux and other UNIX-like OSes). Mac OSX users: Run `brew install libomp` to install OpenMP runtime.

\* You are running 32-bit Python on a 64-bit OS

In this scenario, run the following commands on the Terminal on macOS:

**Note: Open a new Terminal window before using the below commands in the macOS Terminal.**

**To open the Terminal, open the Launchpad and then click on the Terminal icon.**

1. Install the Homebrew package manager by running the following command on the Terminal.

/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"

2. After this, use Homebrew to install the OpenMP library by running the following command on the Terminal.

brew install libomp

This should allow you to install XGBoost on your Mac.

#### **6. Are “unmarried” and “single” in “Marital status” the same? If not how?**

Unmarried and single are not the same in the marital status feature, since an “unmarried” person could potentially have a partner but not be married to them, whereas a "single" person is merely one without a partner.

#### **7. What is the currency of 'MonthlyIncome'? Is it too much to be in US dollars?**

As there is no currency information mentioned in the data, you are free to use a currency of your choice. This does not make a difference to the algorithm's predictive capabilities.

#### **8. Error in reading CSV file Tourism because the file extension is .xlsx. Any solution?**

The file is in .xlsx format (Microsoft Excel file) and not .csv. You need to write the extension of the file as .xlsx and you can then use the function **pd.read\_excel()** to import the file.

#### **9. Tuning XGBoost is taking too much time. What to do?**

Please find below the options that can be tried to reduce the run time for XGBoost tuning:

1. Pick individual features at a time and try adjusting them. Get the optimal value beyond which no improvement is seen, note that optimal value, and move on to the next feature.
2. Only use parameters that are necessary. Do not use all the parameters just because they are there.
3. Do not use more than 2-3 options in each list of hyperparameters given to the tuning function.
4. Try a small dataset randomly chosen from the original dataset, tune hyperparameters on them, and then use that set of hyperparameters over the entire dataset.

#### **10.** **I am getting the below error with Bagging Classifier with Logistic Regression as base\_estimator:**

AttributeError: 'str' object has no attribute 'decode'

#### **How to solve it?**

The LogisticRegression function has a parameter named solver, which was earlier by default set to 'liblinear', but after an update in sklearn, the default solver was changed to 'lbfgs'. Kindly try setting the solver as liblinear, and try the code again, as shown below:

bagging\_lr=BaggingClassifier(base\_estimator=LogisticRegression(solver='liblinear', random\_state=1),random\_state=1)   
bagging\_lr.fit(X\_train,y\_train)

#### **11. I am getting the below warning while fitting XGBoost Classifier**

WARNING: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval\_metric if you'd like to restore the old behavior.

#### **How to solve it?**

To remove the warning kindly try setting the eval\_metric hyperparameter as 'logloss', as shown below:

xgb = XGBClassifier(eval\_metric='logloss')