

You are part of the business analytics team at "RentBrisk", an online platform that allows customers to book vacation rentals. Your team is tasked with the development of a geographical preference model for new customers.

## You have two options in this assignment:

1. The goal is to predict whether a customer is likely to make their first booking in the European Union (EU) or the United States (US). This makes the case a binary classification task.
2. The goal is to predict whether a customer is likely to make their first booking in Australia (AU), Canada (CA), Germany (DE), Spain (ES), France (FR), Great Britain (GB), Italy (IT), Netherlands (NL), Portugal (PT) or the United States (US). This makes the case a multi-class classification task.

The purpose of this model is to tailor RentBrisk's marketing campaigns to the predicted preferences of these customers, providing content that leans towards either an EU or US cultural charm *or if applicable a country specific culture* *for option two above*. By customizing promotional materials according to geographical preference, RentBrisk aims to drive higher engagement and convert potential customers into making their first bookings.

The marketing and customer satisfaction department will be conducting an email and mobile notification campaign focused on households that your model predicts are most likely to make their first booking. Upon completion and evaluation of your model, you will present your findings, including a list of top 500 customers as identified by your model, to RentBrisk's CMO (Chief Marketing Officer).

## Tasks

**1. Data Examination**

You are provided with historical data from more than 17,000 new customers, along with their initial booking locations. Study the datasets carefully, understand the variables, clean and preprocess the data as needed prior to modelling. *The data has been partitioned into train, validation and test sets for you.*

**2. Develop Geographical Preference Model**

Based on the preprocessed data, devise a model that predicts which key geographical areas (EU/US for option 1 or AU/CA/DE/ES/FR/GB/IT/NL/PT/US for option 2) new customers are most likely to make their first bookings. Ensure the model is trained, tested, and refined to achieve optimal results.

*Keep in mind this is a small dataset originally based on publicly released data but results in this modified version may not extremely accurate. The case goal is for you to demonstrate machine learning mechanics, and business acumen with messy, realistic data. The grade earned will NOT be impacted by accuracy alone.*

**3. Evaluate the Model**

Evaluate the performance of your model using suitable metrics. Mention the methods used and the reason behind their selection.

**4. Identify Top Prospective Customers in the Test Set**

Using the best model(s) predict the geographic preferences for the test set of 5000 users.

**5. Customer Profile & Model Behavior Analysis**

Using the information learned in EDA present an understanding of the data characteristics. Once the model(s) have been applied to the test set, reexamination the characteristics of each classified group. The goal is to provide additional exploratory data analysis (EDA) to the RentBrisk executives to offer more insights into customer and modeled behavior.

## Data

The dataset you will be using is a modified version of the publicly available Airbnb dataset, <https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data>. Additional fictiioutus demographic data has been appended. The test set destination (y) column has NDF which is “no destination found”.

## Example Data

|  |  |  |
| --- | --- | --- |
| Variable | Example Value | Notes |
| id | h39zx0jq8i |  |
| date\_account\_created | 2/15/11 |  |
| timestamp\_first\_active | 2/15/11 21:29 |  |
| date\_first\_booking | 2/17/11 |  |
| gender | OTHER |  |
| age | 25 |  |
| signup\_method | basic |  |
| signup\_flow | 0 |  |
| language | en |  |
| affiliate\_channel | content |  |
| affiliate\_provider | google |  |
| first\_affiliate\_tracked | untracked |  |
| signup\_app | Web |  |
| first\_device\_type | Other |  |
| first\_browser | IE |  |
| income | 141226.42 |  |
| job\_sector | Hospitality |  |
| education\_level | Master |  |
| age\_isBad | 0 |  |
| country\_destination | US | This is the dependent variable in either choice (binary or multi-class) |

## Deliverables

Your deliverables should include: (1) your R code, documenting your data processing, model creation and evaluation steps, (2) a presentation deck illustrating your process and findings, designed with the CMO's perspective in mind, (3) a written report, 2-5 pages, discussing your process, findings and business implications, (4) a CSV file of predictions for the test set and (5) a narration of your presentation. The narration can be embedded within the slide deck, recorded and submitted, or recorded and the URL shared in a simple file which is submitted.

## Delivery and Narration Guidance

You are not allowed to use an ai avatar, or speech creation for narration. While this is certainly useful technology in many instances, the purpose of the business case presentation is to improve *your* presentation skills. In a business setting you will still be expected to articulate your findings and not send an avatar for this type of business meeting. As a result services like <https://elevenlabs.io/> or <https://www.heygen.com/> are not permitted. Your “boss” or “audience” in the case will not accept these technologies.

## Criteria for Success

|  |
| --- |
| Organization – Was the presentation well organized? |
| Delivery – Was the content delivered clearly and persuasively with the audience in mind? |
| Code Documentation – Was the data mined to support the conclusion? |
| Written Supplemental – Is it grammatically acceptable, organized and error free.  -Is the data supported clearly and coincides with the data, and narration while being contextualized with external information? |
| Data Mining Process – Overall, as a complete portfolio of work, is the topic interesting, organized, researched, supported and delivered effectively? Was CRISP-DM, SEMMA, or a similar workflow followed to organize the work if appropriate? |

## Additional Guidance

While AI avatars or speech creation tools for narration are strictly off the table, your ability to communicate your findings clearly and effectively will play a major role in your success. Your written supplemental should avoid the use of Markdown or code screenshots, and should instead contain structured text. When turning in your assignments, ensure all links are accessible to reviewers. Private, unshared links will not be considered for review.

* Be sure to review any submitted files, your teaching staff will not track you down for missing components or wrong versions. We will grade components of the case that are submitted and assign 0 for any rubric sections missing. For example, if you neglect to submit your narration the Delivery section of the rubric will be assigned a 0.