**HBA511 Data for Decision Analysis**

**Graded Set of Exercises**

Background and aims of the data analysis

You work as a data analyst for the hotel chain Amazing Hotels Ltd. Throughout the years, the chain has amassed a large amount of data about its bookings, collected through the chain’s Property Management System database. You have been asked to analyse this data, with the aim of understanding what factors may influence whether a booking is cancelled or not, understand characteristics of cancelled and uncancelled bookings and predict whether a booking will be cancelled. For your analysis, Amazing Hotels requires you to complete a number of tasks.

Data details

The data have been extracted from the database into two csv files (*Hotel1.csv*, and *Hotel2.csv*). The variables in these files and their description can be seen in the table below:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| IsCancelled | Categorical, with the categories: 1 = booking was cancelled  0 = not cancelled |
| ADR | Numeric, Average daily rate for the stay |
| Adults | Integer, Number of adults |
| Children | Integer, Number of children |
| DepositType | Categorical, with the categories:  No Deposit – no deposit was made  Non Refund – a deposit was made in the value of the total stay cost |
| IsRepeatedGuest | Categorical, with the categories: 1 = is repeat guest  0 = is not repeat guest |
| LeadTime | Integer, Number of days between booking time and arrival time |
| Meal | Categorical, with the categories: No meal - No meal  BB – Bed & Breakfast  HB – Half board  FB – Full board |
| PreviousBookingsNotCancelled | Integer, Number of previous bookings not cancelled by the customer prior to the current booking |
| PreviousCancellations | Integer, Number of previous bookings that were cancelled by the customer prior to the current booking |
| TotalSpecialRequests | Integer, Number of special requests made by the customer (e.g., twin bed or high floor) |
| BookingChanges | Integer, Number of changes made to the booking until check-in or cancellation |
| StaysWeekendNights | Integer, Number of days booked during the weekend |
| StaysWeekNights | Integer, Number of days booked during the week |

Additional information

The datasets are based on real data from Antonio et al., (2019), collected from real Portuguese hotels. The original data have been modified to avoid plagiarism, and to reflect the characteristics of real-life data which is often imperfect and requires cleaning and pre-processing.

**Reference:** Antonio, N., de Almeida, A. and Nunes, L., (2019). Hotel booking demand datasets. *Data in brief*, 22, pp.41-49.

The task – Data Analysis for Amazing Hotels

You have been asked to complete a set of tasks as part of your data analysis for Amazing Hotels. First, you need to load the two data files into Python and combine them. After you have done this, please complete the following tasks, ensuring that the analysis you produce is reproducible, and can be verified by another person running your code:

1. Clean and pre-process the data to prepare it for analysis (weight 20%). As part of this task, you should identify problems within the data, decide how/whether to address the problems identified, act based on your decision and justify your decision.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rubric for Task 1** | | | | | |
| **A/A+ (Excellent)** | **B/B+ (Very good)** | **C/C+ (Good)** | **D/D+ (Adequate)** | **MF (Marginal fail)** | **F (Fail)** |
| Identifies all | Identifies most | Identifies some | Identifies a limited | Identifies a very | Identifies no |
| problems within the | problems within the | problems within the | range of problems | limited range of | problems within the |
| data. Where | data. Where | data. Where | within the data. | problems within the | data. Fails to address |
| necessary, addresses identified problems | necessary, addresses identified problems | necessary, addresses identified problems. | Where necessary, addresses identified | data. Fails to address most problems. | all the problems. Provides no |
| effectively. | mostly effectively. | Provides good | problems, although | Provides inadequate | justifications for |
| Provides excellent | Provides very good | justifications for | methods used may | or very limited | how/whether to |
| justifications for | justifications for | how/whether to | not be optimal. | justifications for | address problems. |
| how/whether to | how/whether to | address problems. | Provides adequate | how/whether to | Unprofessional |
| address problems. | address problems. | Writing and | justifications for | address problems. | writing and |
| Employs professional writing and | Employs mostly professional writing | programming style, are unprofessional in | how/whether to address problems. | Unprofessional writing and | programming style, with numerous |
| programming style, with no style errors. | and programming style, with very rare style errors. | places, with rare style errors. | Writing and programming style are frequently unprofessional, with several style errors. | programming style with numerous  errors. | errors. |

1. Conduct exploratory data analysis relevant to the aims outlined in the *Background and aims of the data analysis* section. As part of your exploratory analysis, you should provide one correlation heatmap, two bar charts, and suitable charts to visualise the distribution of 4 variables. You should identify key takeaways for each of the graphs presented (weight 20%).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rubric for Task 2** | | | | | |
| **A/A+ (Excellent)** | **B/B+ (Very good)** | **C/C+ (Good)** | **D/D+ (Adequate)** | **MF (Marginal fail)** | **F (Fail)** |
| Provides all required | Provides most of the | Provides some of the | Provides a limited | Provides a very | Provides none of the |
| charts. Where | required charts. | required charts. | amount of the | limited amount of | required charts. |
| necessary, provides rationale for choice of chart. Provides excellent,  comprehensive | Where necessary, provides rationale for choice of chart. Provides very good interpretation of the | Where necessary, provides rationale for choice of charts, which may be incomplete in places. | required charts. Where necessary, provides limited rationale for choice of charts. Some | the required charts. Very limited  rationale is provided for choice of charts. Very limited | Provides no rationale for choice of charts. No interpretation is provided.  Unprofessional |
| interpretation of the | charts, which | Good interpretation | incomplete | interpretation is | writing and |
| charts. Employs | identifies most key | is provided, which | interpretation is | provided. | programming style |
| professional writing | insights. Employs | may be missing | provided. Writing | Unprofessional | with numerous |
| and programming | mostly professional | some key insights. | and programming | writing and | errors. |
| style, with no style | writing and | Writing and | style is frequently | programming style |  |
| errors. | programming style, with very rare style errors. | programming style, are unprofessional in places, with rare style errors. | unprofessional, with several style errors. | with numerous errors. |  |

1. Test a hypothesis about a correlation between two variables of your choice, which you believe to be relevant to the aims outlined in the *Background and aims of the data analysis* section. Justify your analytical and, if required, data manipulation choices, and report and interpret the results (weighting 10%).
2. Test a hypothesis about a difference between two groups of your choice, which you believe to be relevant to the aims outlined in the *Background and aims of the data analysis* section. Justify your analytical and, if required, data manipulation choices, and report and interpret the results (weighting 10%).
3. Build a linear regression model including the variables Adults, Children, IsRepeatedGuest, LeadTime and TotalSpecialRequests as predictors of ADR. You should generate standardised regression coefficients to understand which predictors have the strongest relationship to ADR. Justify your analytical and, if required, data manipulation choices, and report and interpret the statistical results (weight 20%).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rubric for Tasks 3, 4 & 5** | | | | | |
| **A/A+ (Excellent)** | **B/B+ (Very good)** | **C/C+ (Good)** | **D/D+ (Adequate)** | **MF (Marginal fail)** | **F (Fail)** |
| Analysis is | Analysis is | Analysis is | Analysis is | Analysis is mostly | No analysis or |
| performed correctly. | performed correctly. | performed correctly. | performed mostly | incorrect. Almost | analysis is |
| All decisions are | Most decisions are | Some decisions are | correctly, with minor | none of the | completely |
| justified effectively | justified effectively. | justified effectively. | errors that do not | decisions are | incorrect. No |
| and in-depth. | Mostly complete, | Partially complete | substantially affect | justified, and where | justification provided |
| Insightful and | correct | interpretation is | the findings. Most | justification is | for decisions. |
| Comprehensive | interpretation of the | provided, which may | decisions are not | attempted, this is | No interpretation is |
| interpretation of the | results is provided. | contain rare errors. | justified, but where | limited. Attempt at | provided. |
| results is provided. | Employs mostly | Writing and | justifications are | interpretation is | Unprofessional |
| Employs professional writing and | professional writing and programming | programming style, are unprofessional in | provided, these are reasonable. Limited | made, but this is very limited and/or | writing and programming style |
| programming style, with no style errors. | style, with very rare style errors. | places, with rare style errors. | interpretation is provided, which may contain several  errors. Writing and programming style are frequently unprofessional, with several style errors. | incorrect. Unprofessional writing and programming style with numerous errors. | with numerous errors. |

1. Build and compare two machine learning models that predict whether a booking will be cancelled or not. Your target feature should be IsCancelled, and all the other variables in the dataset should be included as predictive features. Evaluate the machine learning models and based on your evaluation, provide a recommendation as to whether either of the models should be implemented or not by Amazing Hotels to predict cancellations (weight 20%).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rubric for Task 6** | | | | | |
| **A/A+ (Excellent)** | **B/B+ (Very good)** | **C/C+ (Good)** | **D/D+ (Adequate)** | **MF (Marginal fail)** | **F (Fail)** |
| Model is built | Model is built | Model is built | Model is built mostly | Model is mostly | No model built, or |
| correctly. Model is | correctly. Model is | correctly. Model is | correctly, with some | incorrectly built. | model is built |
| Evaluated | evaluated mostly | evaluated, although | minor errors that do | Inadequate or | completely |
| comprehensively, using appropriate | comprehensively, using appropriate | several aspects of evaluation have | not substantially affect the results. | incorrect evaluation is provided. | incorrectly. Model is not evaluated. |
| procedures. | procedures. | been missed. | Model evaluation is | Inadequate | No |
| Well-articulated, reasoned | Mostly well- articulated and | Recommendations for action are | limited with numerous aspects | recommendations for action are | recommendations for action are |
| recommendations | reasoned | provided, although | missing. | provided. | provided. |
| for action are | recommendations | they lack in clarity | Recommendations | Unprofessional | Unprofessional |
| provided. Employs | for action are | and rationale in | for action are | writing and | writing and |
| professional writing | provided. Employs | places. Writing and | provided, although | programming style | programming style |
| and programming style, with no style | mostly professional writing and | programming style, are unprofessional in | they are limited in clarity and rationale. | with numerous errors. | with numerous errors. |
| errors. | programming style, with very rare style errors. | places, with rare style errors. | Writing and programming style are frequently unprofessional, with several style errors. |  |  |

Marking

You will receive a grade for each of the tasks (based on the criteria associated with each task). The grade you receive for each task will be weighted based on the weighting associated with each task, to calculate the overall grade for this assessment.

Submission details

Your work must be submitted in a Jupyter notebook (.ipynb file) with code cells and Markdown text cells. Your code must be accompanied by comments to explain what the code is doing. You must use Markdown text cells to provide justifications for your approach and interpretations of the outputs generated. Your use of the text cells must be concise and to the point, focusing on the key information needed to understand the work you have done.

In addition to the .ipynb file, please also export the notebook as a .pdf file (see instructions here: <https://jupyterlab.readthedocs.io/en/stable/user/export.html)> and upload the .pdf file together with your .ipynb file.

Any sources used, either in your code, or in your text cells, must be cited using the Harvard style of referencing. You should include a Markdown cell with the reference list at the end of your Notebook. Importantly, if you use code taken from another source, this needs to be referenced appropriately (as you would any other source for an academic assessment).

The marker must be able to run the notebook to verify your results, without modifying the code, so for this reason, you will need to use pre-specified file paths for loading the datasets. Please ensure you use the relative file paths below, enclosed within double or single quotation marks:

Datasets/Hotel1.csv  
Datasets/Hotel2.csv

This means that, when you are working on your analysis, the datasets should be stored in a folder called “Datasets”, located in the same folder as your Jupyter notebook.

Your submission should include:

* The .ipynb notebook with your code (make sure to run all the cells and save the notebook before uploading it)
* The .pdf version of the notebook (obtained by exporting your notebook to a .pdf file, after you have run all the cells in the notebook).