NGEE ANN POLYTECHNIC

School of InfoComm Technology

Predictive Modeling

Diploma in Financial Informatics (FI)
Diploma in Information Technology (IT)
YR 2/3 (2020/21), Semester 4/6

INDIVIDUAL ASSIGNMENT 1

(30% of Predictive Analytics Module)

Deadline for Submission:

Presentation: Week 9, 14th – 18th Dec 2020
Report: 27th Dec 2020 (Sunday), 2359 Hours

Student Name :	
Student Number :	

Penalty for late submission:

10% of the marks will be deducted every day after the deadline.
NO submission will be accepted after 03rd Jan 2021, 23:59.

PREDICTIVE ANALYTICS ASSIGNMENT 1

1. OBJECTIVES

In this assignment we will explore and analyze two datasets (one classification problem and one regression problem) to understand the data and prepare the data ready for the predictive modelling in Assignment 2.

- To conduct data preparation, exploration and analysis through visualization and statistical approaches
- To prepare the data ready for predictive modeling
- To document the analysis and findings

2. DATASETS

2.1. HR ANALYTICS (CLASSIFICATION PROBLEM)

HR analytics is revolutionizing the way human resources departments operate, leading to higher efficiency and better results overall. Human resources have been using analytics for years. However, the collection, processing and analysis of data has been largely manual, and given the nature of human resources dynamics and HR KPIs, the approach has been constraining HR. Here is an opportunity to try predictive analytics in identifying the employees most likely to get promoted.

This dataset (**hr_data.csv**) contains employee personal information, education background, past performance and etc. Detailed information can be found in the below table. You can utilize all these variables to make prediction on whether the employee will be promoted or not.

hr_data.csv

Variable	Definition
employee_id	Unique ID for employee
department	Department of employee
region	Region of employment (unordered)
education	Education Level
gender	Gender of Employee
recruitment_channel	Channel of recruitment for employee
no_of_trainings	no of other trainings completed in previous year on soft skills, technical skills etc.
age	Age of Employee
previous_year_rating	Employee Rating for the previous year
length_of_service	Length of service in years
KPIs_met >80%	if Percent of KPIs(Key performance Indicators) >80% then 1 else 0
awards_won?	if awards won during previous year then 1 else 0
avg_training_score	Average score in current training evaluations
is_promoted	(Target) Recommended for promotion

2.2. AIRBNB SINGAPORE (REGRESSION PROBLEM)

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world.

This dataset (**listings.csv**) describes the listing activity and metrics in Singapore from year 2013 to 2019. The data file includes the hosts information, the condition of listed properties, the reviews and etc. Detailed information can be found in the below table. You can utilize all these variables to make predictions on the rental price of the listed properties.

listings.csv

Variable	Definition
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	region
neighbourhood	sub region
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	(Target) daily rental price in dollars
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

3. SUGGESTED TASKS

You are suggested to tackle each dataset in the below FOUR steps.

Step 1: **Obtain the datasets** from MEL and **Explore the Data**

Download the datasets (hr_data.csv and listings.csv) from MeL. You are encouraged to utilize both statistical and visualization approaches to familiarize yourself with the datasets.

Step 2: **Cleanse and Transform the Data**

Are there any missing values? How did you handle them? Are there any outliers? How did you identify them and how to deal with them? Do you need to transform the Categorical Data into numbers? Do you need to scale the data or not?

Step 3: **Correlation Analysis**

Investigate the relationships between different features/variables. Which features are likely helpful for making predications? Did you create any new features/variables? Did you drop any features/variables and why?

Step 4: **Export the Data**

After you finish the above preparation tasks, export the newly created data into csv files (**hr_data_new.csv and listings_new.csv**) accordingly. We will be using these newly created datasets to build Predicative Models in Assignment 2.

4. SUGGESTED REPORT FORMAT & CONTENT GUIDELINES

Write an **INDIVIDUAL** report with the following sections (see Table below). Sample content description is provided for each section. You are free to include other relevant information you deem necessary in the sections. You are strongly encouraged to include screen shots in your explanation, description and/or analysis.

*(Note: For a page with 1 inch margins, 11 point Calibri font, and minimal spacing elements, a good rule of thumb is **500 words** for a single spaced page)*

	Suggested Report Sections & Content Guidelines	**Word Count**
1.	Table of Contents	NA
2.	Summary/Overview	500 words
3.	HR Analytics • Problem Understanding • Data Exploration • Data Cleansing and Transformation • Correlation Analysis • Others **Hint:** for this binary classification problem, do you have equal sized samples for the two classes? If not, how did you handle this? Stratified Sampling?	Min: 1000 words Max: 3000 words
4.	Airbnb Singapore • Problem Understanding • Data Exploration • Data Cleansing and Transformation • Correlation Analysis • Others: **Hint:** Do you plan to utilize all the samples to build a generic price prediction model? It is likely more effective to focus on a subset of data (e.g. certain region, certain price range and etc.) to build a customized price prediction model.	Min: 1000 words Max: 3000 words
5.	Summary and Further Improvements • Summarize your findings on the two datasets • Explain the possible further improvements	Min: 500 words Max: 1000 words

5. DELIVERABLES

Presentation and demonstration

- Each student will be required to submit a video recorded presentation to showcase and demo the work. The video recorded presentation should be not exceed 10 minutes. Video recorded presentations which exceed the allotted time will be penalized.
- Students to submit the presentation slides in MeL. Deadline for slides submission is **Sunday 13th Dec 2020, 2359 hours.**
- The video presentation must be recorded using Microsoft Teams. After completion of the video recorded presentation, submit the link to the video (from Microsoft Stream). Submit the link to your video recorded presentation using the link below:
  - Assignment 1 Video Presentation Submission Link
    (Login using only your NP student account)
- Deadline for the video submission is **Sunday 20th Dec 2020, 2359 hours.**

Assignment report

- Submit the **softcopy** of the report via **SafeAssign** in MeL. Deadline for softcopy submission is **Sunday 27th Dec 2020, 2359 hours**
- Submit the Jupyter Notebook file (PA_Assignment_1.ipynb) in MeL. Deadline for softcopy submission is **Sunday 27th Dec 2020, 2359 hours**

Note: DO NOT PLAGIARIZE (please refer to Ngee Ann Polytechnic Plagiarism Policy webpage for more information)

6. GRADING CRITERIA

	Grading Criteria	Component Weightage
Presentation	a) Quality of work b) Flow of presentation based on content guidelines (see section 4) c) Quality of presentation slides d) Presentation and articulation skills	**50%**
Final Report	a) Quality of work b) Completeness of report based on suggested report sections and content guidelines (see section 4) c) Clarity of report, Quality of discussions, Use of proper visual aids and Use of proper grammar d) Quality of recommendations for further improvements	**50%**