Exercise Readme

Dataset Background

PredictiveHire's FirstInterview™ is a text-chat based candidate screening system where job applicants answer 5-7 situational judgement and past behaviour questions. Responses to these structured interview questions are then analysed by PredictiveHire AI (PHAI™), a Natural Language Processing (NLP) and machine learning based evaluation algorithm, to generate a candidate "suitability recommendation" (i.e. in the form of "Yes", "No", or "Maybe") used in screening candidates. In addition to the "suitability recommendation", PHAI also generates personality and behaviour-based traits of the candidates to further assist the recruiters and hiring managers to make decisions.

You will be provided with a mock dataset that is similar to the actual datasets from FirstInterview systems. Teams at PredictiveHire are keen to understand (1) how the product performs against different KPIs or key factors, (2) insights that will help make business decisions.

Key Fields of the datasets

The three datasets represent three different data sources. Typically there are multiple data sources where different types of data are stored and part of your work is to identify how these data sources can be combined to gain better insights.

Dataset "1_product_F_CSAT"

At the end of each FirstInterview section, candidates are provided with a chance to give feedback on the experience about FirstInterview, which contains two parts: (1) a score rating in the range of 1 to 10, and (2) text feedback. This dataset is a processed version of the feedback, in which two sentiment classification methods are used to detect the sentiment of the text feedback.

Field	Description	Comments
prediction_id	Data Entry ID	Candidate Identifier
comment_sentiment	Sentiment of the candidates' feedbacks for product FirstInterview	Sentiment derived from candidate's feedback text based on sentiment classification method 1
polarity	Sentiment polarity of the candidates' feedbacks for product FirstInterview	Sentiment derived from candidate's feedback text based on sentiment classification method 2, the value typically ranges from -1 to 1. The larger the value, the more positive the sentiment is.

Dataset “2_main_data_source_A”

Field	Description	Comments
prediction_id	Data Entry ID	Candidate Identifier
submittedAt	The timestamp when the candidate submits the responses in FirstInterview.	
cohortName	Job listing name	
timeTaken	The number of minutes a candidate spent on completing PredictiveHire’s FirstInterview.	
Recommend	Suitability Recommendation given by PredictiveHire AI	Values can be “YES”, “MAYBE”, “NO”.
ftq_count	Number of questions answers by the candidates during the interview	
total_words	Total number of words candidates write in FirstInterview	
plagiarismType	Plagiarism Testing Result	“TRUE_POSITIVE”: positive plagiarism test result on candidates answer in FirstInterview. “NONE”: plagiarism test result for the response is not available or unknown. “NEGATIVE”: negative plagiarism test result on candidate’s answer in FirstInterview, i.e., the candidate does not copy from others. “SELF_PLAGIARISM”: candidates copy from themselves when answering FirstInterview. For example, some candidates may apply to multiple jobs and provide very similar answers.
gender	Gender of the candidate	
hexaco_extraversion_general	Extraversion value based on the HEXACO Personality Model. A larger value indicates a person is more extravert.	More information about the HEXACO personality model can be found here. Columns with names such as hexaco_xxx_general are all HEXACO personality model values.
easl	English as a second language	Showing whether a candidate speaks a language other than English, which is considered their native tongue.
trait_team_player_percentile	Behaviour trait value for team_player. The larger the value, the more likely a candidate engages in teamwork as predicted by PHAI.	Columns with names “trait_xxx_percentile” are all behaviour traits values.

Dataset “3_main_data_source_B”

Field	Description	Comments
prediction_id	Data Entry ID	Candidate Identifier, a dropped-out candidate does not have this ID.
cohortId	Id of the job posting.	For example, a job posting can be about a carer role in a hospital.
submittedAt	The time stamp when the candidate submits the responses in FirstInterview.	
customerSatisfaction_score	The Net Promoter Score (NPS) given by the candidate for the FirstInterview experience.	Higher scores indicate that candidates are happier about the FirstInterview experience.
status	Whether the candidates finish the FirstInterview	“dropped out” indicates that a candidate did not finish answering all the questions in the FirstInterview. “completed” indicates that a candidate finishes answering all the questions in the FirstInterview
recruiter_id	Id of the recruiter that publish the job advertisement	
job_publish_date	The timestamp when the job advertisement is published	
job_close_date	The timestamp when the job advertisement is closed	
customerSatisfaction_comment_length	The number of words candidates write as comments to the FirstInterview experience.	Alongside the NPS scores, candidates are asked to write a few comments about their experiences in FirstInterview. This field counts the length of their responses.

Task

We are keen to understand how you would approach a reporting and analytics task with complicated business background information.

We do not expect you to build a comprehensive solution or report, but keen to understand:

(1) How you preprocess the dataset, including cleaning, linking, and merging the dataset to facilitate any further analysis.
(2) What are the key insights in the dataset you can find that may help a business with improving the product, promoting the service in the market, making the right business decision, etc.
(3) How you can present your findings in a clear and precise way.

You are free to make fair assumptions about the data. Note that the data in the dataset are deliberately de-identified and distorted so please do not be surprised if some of the findings are counterintuitive.

Effort

No more than 3hrs

Outcome

Write a short description of the methodology and key findings. For example: What was your approach? What are the most important pre-processing steps? What are the interesting insights? What are the suggestions to the business given the findings? Are there any assumptions you have made and questions you might ask etc

You might want to keep the advanced steps, methods, or analysis you would like to consider in a comprehensive solution as future steps in the description.

This is an open-ended analytics exercise. You may choose from the following key factors and analytics angles. Also, feel free to choose a tool you are familiar with for analysing data and presenting results.

Key factors:

(1) Dropout
(2) CSAT (customer satisfaction, or NPS)
(3) Plagiarism
(4) Suitability Recommendations
(5) Candidates engagements
(6) Other key factors that can be supported by the data in the dataset.

Analytics angles to evaluate the key factors:

(1) Descriptive Statistics.
(2) How does a key factor change over time?
(3) How does a key factor look differently from one type of role to another, etc?
(4) Is there any interesting discovery when looking at more than one key factor at the same time?
(5) Other angles that might help the business.

Presentation tools:
You might choose the tools you like to present the findings, for example:
(1) Microsoft Word or PowerPoint,
(2) Python, Jupyter notebooks,
(3) BI tools such as ProwerBI, Tableau or AWS QuickSight.
(4) Combination of the above tools or other tools you are familiar with.

Example:

An example outcome can be a word report showing that candidates' dropout ratio in customer A is significantly higher than that in customer B so that a suggestion is made to the business that an update needs to happen to the frontend interface for customer A to make it less intimidating in the hope that it will reduce the number of dropouts for customer A.