**Capstone Report**

**Canon EMEA - Matching Projects to Employees**

**Saad Joiya**

**Aim**: Match employees and projects.

**Method**: Need to create mock data based on client requirements for both employees and also projects. Moreover, need to figure out the data templates, data types and how different elements come together from both sources to produce final scoring/ ranking. Using text matching and other mathematical formulas for matching criteria.

**Limitations**: No data available. We had to create ourselves based on feature set. Also we do not have any true labels so we cannot train a ML model so we need to use other approaches.

**Disadvantages**: A lot of research and manual work would be needed in data creation and also algorithm implementation. This model cannot learn with new data as this is not a ML model.

**Advantages**: High interpretability. We can use weights as needed to control how the employees are ranked and which element contributes more than others. This method is not black box so we can tweak based on client expectations. WE can also apply post filtering based on employee availability etc that is a plus.

**Output Format**: A jupyter notebook that contains end to end pipeline from Data Creation to Weighted Scoring.

**Suggestions**: Canon Learn and Development team should try to gather project and employee profiles in the structure describe for using this matching algorithm. As data starts to accumulate, we can test the efficacy of these matchings. With time, as we gather performance labels, we can use the features from projects and employees as predictors to train classification model which automates the matching process.

**Evaluation Criterion**: Arena approach is implemented for now, the results are periodically shared with the client. The feedback received on matches is then used to optimize the algorithm.

**Project Table Details:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Features** | **Description** | **Data Type** | **Details** |
| Project Summary | Give a brief summary | Text | Free text, ex. "This project will support the european patent attorney team by building internal systems and support processes aligned with their strategic objectives." |
| Scope and Deliverables | Outline the scope and list milestones | Text | Free text, ex. "Design support processes, create documentation for best practices, collaborate with business units, and track implementation progress." |
| Customer Industry | Specify the customer’s industry (e.g., finance, healthcare) | List | Finance, Healthcare, Retail etc |
| Customer Preferences or Standards | Any specific standards, compliance, or customer communication preferences | Text | Free text or non finite list, ex "Needs to be compliant with ABC100 standard" or "XYZ project methodology is a must" |
| Products Involved | The specific solutions the project centers on, which is crucial if certain employees have experience in those | List | List. For e.g. MVP AI SCAN, WORKFLOW2000, PRINT2.0 |
| Integration Requirements | Any required integrations with existing systems, third-party tools, or APIs | Text | Integration with ERP System, Master data exchange with rest api etc (non finite list can put everything else in others) |
| Required Skills and Expertise | Skill Requirments for the project e.g., JavaScript, project management, Printing | List | JavaScript, Python, Project Management, Graphic Designing, and expertise level (1 to 10 expertise level) |
| Complexity Rating | A subjective rating classify the project’s complexity | Integer | 1 to 10 |
| Work Location | Location of Work Office | Category | Canon Offices Cities/European City |
| Work Flexibility | Indicate whether the project can be remote, hybrid, or if it requires on-site work in specific locations | Category | Value from either of On site, Remote, Hybrid |
| Language Requirements | Languages required to communicate in this project | Dictionary (key) | Langauges like English, German, French, Portuguese etc (non finite list can put everything else in others) |
| Language Level | CEFR Level of languages required to communicate in this project | Dictionary (value) | A1, A2, B1, B2, C1, C2 |
| Effort | Estimated workload in hours required for the project | Integer | Hours Required |
| Requested Timeline | Desired End Date | Date | Delivery Date |

**Employee Table Details:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Features** | **Description** | **Data Type** | **Details** |
| Role | Job Description | Text | For e.g "Provides high-quality patent services including drafting, filing, and prosecuting patents" |
| Industry Experience | Specify industries experienced in | List | Finance, Healthcare, Retail etc |
| Internal /External Certifications | Relevant certifications (e.g., PMP, Six Sigma). | List | For e.g. PMP, Six Sigma certifications etc. |
| Product Experience | Canon Product Names, redacted | List | List. For e.g. MVP AI SCAN, WORKFLOW2000, PRINT2.0 |
| Expertise | Define expertise of employee | List | For e.g. i.e. Scripting, Integration, Color Reproduction, Cloud & Infrastructure |
| Core Competencies | Specific skills or technologies, e.g., programming languages, project management methodologies. | Dictionary (key) | For e.g. JavaScript, Python, Project Management, Graphic Designing |
| Core Competencies (Expertise) | Rating Specific skills or technologies, e.g., programming languages, project management methodologies. | Dictionary (value) | Core Competencies expertise level (1 to 10 expertise level) |
| Work Location | Location of Work Office | Category | Canon Offices Cities/European City |
| Work Flexibility | Indicate whether the project can be remote, hybrid, or if it requires on-site work in specific locations | Category | Value from either of Onsite, Remote, Hybrid |
| Languages Known | Languages required to communicate in this project | Dictionary (key) | Languages like English, German, French, Portuguese etc (nonfinite list can put everything else in others) |
| Communication Skills | CEFR Level of languages required to communicate in this project | Dictionary (value) | A1, A2, B1, B2, C1, C2 |
| Cultural Awareness | Openness to diversity and experience in international settings, valuable for global teams. | Integer | Openness Rating 1 to 5 |
| Problem Solving | Critical thinking and adaptability in projects | Integer | Previous project count |
| Leadership | Leadership or mentoring/coaching experience. | Integer | Overall Years of experience |
| Collaboration | Experience in team environments, cross functional collaboration etc. | Integer | Ratings 1 to 5 based on experience, project count & teamwork |

**Matching and Scoring Criterions:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Feature** | **Employee Feature** | **Matching Method** | **Notes** |
| Project Summary | Role | Text Embedding / Cosine Similarity | General fit & thematic similarity based on text embeddings and cosine similarity |
| Scope and Deliverables | Role | Text Embedding / Cosine Similarity | Task alignment based on text embeddings and cosine similarity |
| Customer Industry | Industry Experience | Category Similarity (Fuzzy match with threshold) | Match industries for coverage |
| Customer Preferences or Standards | External Certifications / Internal Certifications | Keyword Matching / Set Overlap | Standards/compliance mapping. Cleaning and tokenizing strings into sets, then measuring overlapping of keywords. |
| Products Involved | Product Experience | Category Similarity Coverage (Fuzzy match with threshold) | Matching Products for coverage |
| Integration Requirements | Expertise | Keyword Matching / Set Overlap | Technical integration alignment. Cleaning and tokenizing strings into sets, then measuring overlapping of keywords. |
| Required Skills and Expertise | Core Competencies | Category Similarity Coverage (Fuzzy match with threshold) | Primary skill matching coverage |
| Complexity Rating | Core Competencies (expertise level) | Coverage x expertise fit (employee capability score) then compare complexity | Compare Employee Capability based on skills to Complexity |
| Work Location | Work Location | Category Similarity (Fuzzy match with threshold). Irrelevant if remote | Location coverage |
| Work flexibility | Work flexibility | Category Similarity (Fuzzy match with threshold) | Remote/Hybrid/On-site scoring factoring for location coverage |
| Language Requirements | Language Proficiency | Category Similarity Coverage (Fuzzy match with threshold) | Check if required language is present for coverage |
| Language Level | Communication skills | Coverage x language fit (language capability) | Scoring using match or near-match on required fluency factoring for language coverage |
| Effort | - | Filter on availability after scoring | Filter on availability after scoring |
| Requested Timeline | - | Filter on availability after scoring | Filter on availability after scoring |
| - | Cultural Awareness | Optional Bonus (Rating Match) | Use as soft bonus if needed |
| - | Collaboration | Optional Bonus (Rating Match) | Soft skill scoring if needed |
| - | Problem Solving | Optional Bonus (Normalized Score) | Tiebreaker or bonus |
| - | Leadership | Optional Bonus (Experience-Based) | Soft bonus for senior roles |

**Matching Different Features:**

**Product Score is calculated using:**

**Tables and Columns Used**

| **Table 1** | **Variable** | **Table 2** | **Variable** |
| --- | --- | --- | --- |
| Projects | Products Involved | Employees | Products Experience |

Here we will have a list of products In the projects table and similarly a list in the employees table. As the products involved field in projects data is human filled, we first use fuzzy matching to get rid of any mismatching with products experience column in employee table caused by spelling mistakes and then we match the columns. The coverage is calculated based on how many required products are possessed by the employee. For example, if a project requires products AIScan, Print2.0 and Workflow2000 and employee knows only AIScan, the coverage would be 33%. If an employee knows any two of the above then 66% and in case of knowing all three or more, it would be 100%.

**Location Score is calculated using:**

**Table and Columns Used**

| **Table 1** | **Variable** | **Table 2** | **Variable** |
| --- | --- | --- | --- |
| Projects | Work Location | Employees | Work Location |
| Projects | Work Flexibility | Employees | Work Flexibility |

Scoring Mechanism:

This calculation assigns a **location match score** based on both:

1. **Work flexibility compatibility**
2. **Location similarity (using fuzzy matching)**

Note: As Locations are filled by Humans, to match them we would use fuzzy matching.

| **Project Flexibility** | **Employee Flexibility** | **Location Match** | **Score** | **Explanation** |
| --- | --- | --- | --- | --- |
| remote | Any | — | 1.0 | Location irrelevant for remote work |
| Any except remote | Any | ❌ | 0.0 | Location mismatch or wrong flexibility |
| onsite | onsite | ✅ | 1.0 | Perfect location and presence match |
| onsite | hybrid | ✅ | 0.5 | Partial match; available some days onsite |
| onsite | remote | ✅ | 0.0 | No physical presence at required location |
| hybrid | onsite | ✅ | 1.0 | Can fully accommodate onsite days |
| hybrid | hybrid | ✅ | 1.0 | Flex on both sides; good match |
| hybrid | remote | ✅ | 0.5 | Remote match possible but less ideal |

**Language Score is calculated using:**

**Language Match with Proficiency Logic**

We match project language requirements with employee language fluency, handling:

* Typos/misspellings in language names (fuzzy matching)
* Fluency level comparison using the CEFR scale
* Finally, scoring based on coverage and fluency fit

**Table and Columns Used**

| **Table 1** | **Variable** | **Table 2** | **Variable** |
| --- | --- | --- | --- |
| Projects | Language Requirements | Employees | Language Proficiency |
| Projects | Language level | Employees | Communication skills |

**Mapping used for CEFR Scale:**

**CEFR Level Numeric Value**

**A1 1**

**A2 2**

**B1 3**

**B2 4**

**C1 5**

**C2 6**

**Steps for calculating Language Score:**

1️. For each project language, find the best fuzzy match in an employee’s known languages

2️. If a match is found:

Compare CEFR levels using the cefr\_scale

3️. Score per language:

• If employee level ≥ required → score = 1.0

• Else → 1 - (diff / 6)

4.Calculate Average Fit:

The mean of all individual language scores per language (fluency comparison), **only** for matched languages.

5. Coverage = matched languages / total required

6. Final Score = coverage × average fit