Cahier des charges

Cadrage

Auteure de ce cahier des charges : Carole Lemort
Client : Amine Bibane (INSY2S)

Présentation générale du besoin

Besoin : augmenter le nombre de collaborateurs d'une entreprise qui a besoin de recruter rapidement -> augmenter la vitesse de recrutement par rapport au processus classique en facilitant le travail des ressources humaines

objectif du besoin = gain de temps

Problématique = Relier des profils de personnes en adéquation avec des appels d'offres

Périmètre = Ensemble de mots clés pour sélectionner les profils et leur attribuer une note

Définition Scoring : attribuer une note à un profil en fonction des différents critères

Il existe deux catégories de compétences à évaluer :

- Compétences techniques (Ce qui est écrit) (Ce qui est lié aux compétences et à l'exercice du métier)
- Savoir être (Lire entre les lignes) (Explicite ou entre les lignes -Soft skills-)

Type de projet = Analyse de données : Sélection des variables pertinentes et attribution d'un poids à chaque critère

Source des données - > Réseaux sociaux, LinkedIn, APEC, CADREMPLOI…

Solution proposée

Règles de gestion

- Récupération des données = Ciblage Réseaux Sociaux, Rapatrier des données
- Intelligence Artificielle = Scoring, TALN (Traitement Automatique du Langage Naturel)
- Base De Données (stockage)
- Entrepôt de données, Reporting
- Réutilisation de la base de données pour une utilisation ultérieure

Lors de la récupération des données, le profil donne plus d'informations qu'un simple CV. Il est possible d'obtenir des choses tels que les hobbies, les moyens de locomotion, …

Comment attribuer une note à un profil ?
- pourcentage de correspondance
- parsing et matching

Etapes

1. Parsing : Lecture des CV/profils à l'aide d'une API pour créer une Base de données (quoi mettre, liens entre les tables, ….)
2. Matching : identifier les éléments qui correspondent et ceux qui ne "correspondent pas" = mots clés, compétences, expérience, niveau d'études, …
3. Occurrences et fréquences des expressions clés recherchées : une compétence demandée dans la fiche de poste et qui est présente dans peu de CV aura plus de poids qu'une compétence que tous les profils possèdent

Avantages

- Pour la présélection, l'Intelligence Artificielle est beaucoup plus rapide qu'un être humain -> économie de temps pour le département des ressources humaines

- Le tri et le classement des candidats sont automatiques selon la fiche de poste ce qui offre 2 intérêts. Premièrement, l'analyse est toujours identique ce qui offre une sélection plus équitable car il n'y a pas de biais cognitif. Deuxièmement, cela offre la possibilité d'ajouter un filtre selon un ou plusieurs critères pour un tri rapide : le lieu, le secteur d'activités, le niveau d'études …..

- Cette méthode promeut une meilleure image de marque de l'entreprise.

Inconvénients

- L'analyse est uniquement factuelle et ne prends pas en compte les softs skills tel que le savoir être (Pour l'instant) -> voir si on a le temps de faire cette partie durant le mois imparti

- L'analyse n'indique rien sur la personnalité du candidat : adéquation avec l'équipe, la vision de l'entreprise …

- L'intelligence artificielle ne doit pas être confondue avec l'omniscience : certains profils peuvent être exclus à cause du parsing / matching car ils ne présentent que peu ou pas de mots clés alors que la personne possède des qualités et des compétences intéressantes pour le poste ciblé.

Evaluation des algorithmes : courbe de ROC et de LIFT, sur-apprentissage, courbe précision-rappel, de la probabilité à la décision, performance et tuning (pistes d'amélioration)

Architecture

Equipes

Jessy, Elias = API, Data Integration

Adrien, Bilal, Saïd, Amin = Data Science : Scoring, TALN

Balkihisse, Sandra, Fatiha, Moussa, Mouloud, Mejdi = Data Analyse : Entrepôt de données, Reporting

Contraintes de planning
1 mois du 28 février au 25 mars 2022

Contraintes de budget
Pas de budget alloué

Méthode de projet = méthode agile
Les sprints ont une durée d'une semaine

Planification et organisation

Data Integration

Objectif final = Mise en place de l'extraction des données, mise en place des flux de données vers la base de données, intégration et mise en ligne

Semaine 1 :

- Création d'une application LinkedIn puis essai (en attente de validation pour récupération profil : 90 jours)
- Récupération de données sur des banques de données : lecture et nettoyage
- Récupération de template CV puis tentative de transformation en JSON. Impossibilité de le faire en algorithme pur donc problème d'extraction des données : possibilité d'extractions de données à l'aide du machine learning avec l'équipe data science.
- Extraction et parsage de données à partir de fichiers textes

Problème rencontré : Les informations venant de plusieurs sources de données différentes, il est difficile d'obtenir des colonnes de données homogènes.

Semaine 2 :

- Trouver un format de données qui puisse contenir les différentes informations extraites de sources hétéroclites et qui puisse convenir à la fois aux data scientist et aux data analyst.
- Extraction impeccable des données et conversion en un format adéquat pour mettre sur une base de données en suivant le Modèle Conceptuel de Données.

Semaine 3 :
Mise en place Talend pour automatisation et communication avec la base de données

Semaine 4 :
Installation des données récupérées aux étapes précédentes sur le serveur
Le serveur sera mis en ligne, et ajout de données supplémentaires à partir d'API ou de Talend, en ligne.

Data Science

Semaine 1 : Bibliographie

- Adrien + Bilal = algorithmes de scoring (clustering et K-means)
- Saïd + Amin = soft skills et NLP

Recherche de l’algorithme adéquat mais on a eu des problèmes et on a finit par trouver k-means

Problème : pour commencer le preprocessing, il faut les vraies données, pas des fausses données pour s'entraîner sur les algorithmes sinon le nettoyage ne sera pas correct

Semaine 2 : Preprocessing : Nettoyer et transformer les données

- Adapter les données à l’algorithme utilisé
- Extraction des soft skills à partir du traitement du langage naturel et analyse sémantique

Semaine 3-4 : Calcul du Score

- Création de l’interface utilisateur
- Détermination du seuil à partir duquel les profils sont affichés
- Amélioration du score à partir du calcul de base

Détails des sprints

Adrien et Bilal :

Après une semaine de recherche, on a recherché des modèles de prédiction, nous avons les prédictions supervisées et non supervisées. Dans les modèles supervisés, le knn (les plus proches voisins) était un modèle envisagé, mais ces modèles ont besoin de données étiquetées (de profil type).
Dans les modèles non supervisés nous avons le k-means (on recherche le centre des clusters). Les modèles non supervisés permettent de rechercher des profils atypiques.

Dans les modèles non supervisés, nous avons choisi le modèle de clustering, les différents métiers ont des compétences spécifiques, un préparateur de commande n’aura pas les mêmes compétences qu’un commercial, on a donc des clusters. Le k-means est un modèle de clustering de base, il nous permettra de faire le nettoyage, la transformation et la sélection des données (phase 2). Par la suite, nous testerons d'autres modèles de clustering pour avoir un modèle plus fiable.

Pour le score (phase 3) nous avons pensé calculer le score grâce aux “distances” du profil recherché, plus les profils des candidats sont proches du profil recherché meilleur sera le score, nous n’avons pas encore choisi l’échelle.