**PROJET TINDER**

**Préambule.**

Bonjour à tous et bienvenu à tous dans les locaux de Jedha. Je suis Olivier Evrard, data analyste. J’ai été missionné par l’entreprise Tinder pour essayer de comprendre pourquoi les matchs proposés ne se soldent pas tous par un rendez-vous afin d’optimiser ceux-ci.

Tinder est une application de rencontres en ligne et de réseautage géo-social. Dans Tinder, les utilisateurs « glissent vers la droite » pour aimer ou « glissent vers la gauche » pour ne pas aimer les profils des autres utilisateurs, qui incluent leurs photos, une courte biographie et une liste de leurs intérêts.

En 2021, Tinder a enregistré plus de 65 milliards de matchs dans le monde.

**Projet :**

L'équipe marketing a besoin d'aide pour un nouveau projet. Elle constate une diminution du nombre de correspondances et essaie de trouver un moyen de comprendre ce qui motive l'intérêt des gens pour les autres.

Ils ont décidé de mener une expérience de speed dating avec des personnes qui devaient donner à Tinder de nombreuses informations sur elles-mêmes qui pourraient finalement se refléter sur leur profil de rencontre sur l'application.

Tinder a ensuite rassemblé les données de cette expérience. Chaque ligne de l'ensemble de données représente un speed dating entre deux personnes et indique si chacune d'elles a secrètement accepté d'aller à un deuxième rendez-vous avec l'autre personne.

**Objectif :**

Utilisez l'ensemble de données pour comprendre ce qui incite les gens à s'intéresser l'un à l'autre pour aller à un deuxième rendez-vous ensemble

**Portée du projet :**

Les données ont été recueillies auprès de participants à des rencontres rapides expérimentales de 2002 à 2004. Au cours de ces rencontres, les participants avaient un « premier rendez-vous » de quatre minutes avec chaque autre participant du sexe opposé. À la fin de ces quatre minutes, on leur demandait s'ils souhaitaient revoir leur partenaire. On leur demandait également d'évaluer leur partenaire sur six critères : l'attrait, la sincérité, l'intelligence, le plaisir, l'ambition et les intérêts communs.

L'ensemble de données comprend également des données de questionnaire recueillies auprès des participants à différents moments du processus. Ces champs comprennent : les données démographiques, les habitudes de rencontre, la perception de soi à travers des attributs clés, les croyances sur ce que les autres trouvent précieux chez un partenaire et les informations sur le mode de vie.

**Données fournies.**

Les informations fournies sont un fichier csv qui contient les données d’une étude faite suivant un panel d’utilisateurs et un document explicatif sur cette base de données.

**Analyses de ces données.**

Le fichier comporte 195 colonnes sur 8 378 lignes. Une ligne correspond à rendez-vous, nous avons donc 552 personnes qui ont participé à cette expérience. Le nom des colonnes n'étant pas très explicite, la documentation fournie nous sera d'une grande aide. Le nombre de 552 participants à cette étude n'est pas très élevé aux vues des 65 milliards de match dans le monde. Beaucoup de données sont manquantes et donc demanderont une attention particulière. Les évaluations demandées ne sont pas notées dans la même base en fonction de la wave ce qui posera un problème lors de calculs. Beaucoup de données catégorielles ont été converties numériquement.

**Profil des participants.**

**Physique** :

La répartition homme/Femme est de 50/50. On remarque aussi que nous avons à peine une cinquantaine de personnes en moyenne par tranche d'âge. La race majoritairement représentée est de type European/Caucasian-American suivie par le type Asian/Pacific Islander/Asian-American. L’opération de speed dating a eu lieu un peu partout dans le monde.

**Social** : à revoir pour comparer les professions manuelles et intellectuelles qui ont l’air le mieux représentées.

**Habitudes de vie des participants.**

Les participants sont en grandes majorité des personnes qui sortent assez souvent. Malgré tout, elles n'ont pas forcément un nombre de rendez-vous conséquents. Globalement ces personnes n'attachent pas une grande importance au fait d'être de la même race ou de la même religion dans une relation.

**Attentes des participants dans le sexe opposé.**

Une grande majorité des participants sont là pour passer une agréable soirée et rencontrer des gens sans forcément chercher le grand Amour. Les qualités recherchées rentrent tous dans la catégorie « personne idéale », avec une attractivité (caractère physique) plus prononcée pour les hommes, alors que les femmes recherchent plus l’intelligence.

**Speed Dating.**

Lors de ce speed dating, il a été redemandé les qualités recherchées, avec une évolution vers l’attractivité pour les femmes. Ensuite après le speed dating, il a été demandé les qualités sur les personnes rencontrés avec une séparation pour celle où une demande de rendez-vous a été demandé. On constate qu’il n’y a pas de réelles raisons qui caractérisent une demande de rendez-vous, on peut noter quand même qu’un manque d’attractivité a certainement freiné une demande d’un second rendez-vous.

Les inquiétudes de Tinder sont bien fondées, à peine un peu plus de la moitié des participants sont satisfaits des rencontres proposées.

Le temps imparti de 4 mn a été jugé trop court, ce temps a certainement été volontairement court pour suscité l’envie de connaitre plus l’autre personne avec un second rendez-vous.

Le nombre de speed dating a lui aussi été jugé peu important.

**Bilan suite aux speed Dating.**

À la suite des 8378 rencontres organisées par le speed dating, seulement 680 matchs obtenus, sachant qu’un match est comptabilisé lorsque 2 personnes ont décidé de se revoir, soit 8.12 %.

On a aussi la confirmation que le critère des races n’influe pas sur la décision de se revoir.

Le critère de l’attractivité semble maintenant très prononcé. La différence d’âge n’est pas très probante.

**Deuxième rendez-vous.**

Sur les 680 matchs, suivront 531 rendez-vous soit au final moins d’un rendez-vous par personnes.

**Conclusion.**

**1 - Considérations des éléments fournis**

L'étude s'appuie sur un échantillon de 552 personnes, réparties équitablement entre hommes et femmes, provenant de plusieurs continents. Cet échantillon représente une fraction très limitée des utilisateurs de l'application, estimés à plusieurs millions dans le monde. Comme observé, nous disposons de seulement une cinquantaine de personnes par tranche d'âge.

De plus, de nombreuses données sont manquantes, probablement en raison de la redondance des informations demandées aux participants. Des incohérences ont également été relevées dans les réponses fournies.

**2 - Conclusion**

Les inquiétudes de Tinder sont bien fondées, Les résultats montrent un taux moyen de satisfaction dans les propositions de rencontre lors des speed datings. Bien que l'attrait physique soit le premier critère de sélection, la poursuite des rendez-vous semble principalement motivée par le désir de mieux connaître l'autre. Cela est probablement dû à la durée de 4 minutes du speed dating, jugée trop courte par les participants des deux sexes. La différence d’âge entre les femmes et les hommes reste perceptible, avec des femmes plus jeunes que les hommes en général mais n'est pas fondamental.

**3 - Propositions d'améliorations**

Les matchs proposés devront prendre en considération d’avantage le coté attractif des personnes avec un âge plus jeune pour les femmes. Il faut donc développer les arguments concernant les goûts et les couleurs des personnes. Ces paramètres étant très subjectifs, il est à noter que les participants sont ouverts à la mixité sociale et garder une part de mystère forçant le désir de découvrir l'autre.

**Prédiction.**

Nous pouvons maintenant prédire suivant votre profil les chances d’obtenir un rendez-vous.

F1 score : Moyenne harmonique de la précision et du rappel. Un bon score signifie que son rappel et sa précision sont élevés. Nous avons une valeur de 0.69 ce qui est pas mal.

MSE : l’erreur quadratique moyenne a aussi un bon score.

R2 score : les valeurs d’entrainement et de test sont assez proche de 1. le modèle est assez performant

**Le score F1**

Le score F1 (ou F-measure en anglais) est une **moyenne harmonique de la précision et du rappel**. Il équivaut au double du produit de ces deux paramètres sur leur somme. Sa valeur est maximale lorsque le rappel et la précision sont équivalents.

Dans certains modèles, on cherche à connaître la distribution des faux positifs et des faux négatifs. La métrique score F1 est alors utilisée pour évaluer la performance de l'algorithme. De même, il est particulièrement difficile de comparer deux modèles avec une faible précision et un rappel élevé. Le contraire est également vérifié. Dans ces conditions, le score F1 permet de mesurer ces deux paramètres simultanément.

**La précision**

La précision indique le **rapport entre les prévisions positives correctes et le nombre total de prévisions positives**. Ce paramètre répond donc à la question suivante : sur tous les enregistrements positifs prédits, combien sont réellement positifs ? En reprenant l'exemple des courriers, cela désigne le rapport entre le nombre de courriers pertinents reconnus comme tels (TP) par le modèle sur la somme combinée de ce même paramètre avec le nombre de spams considérés comme des courriers pertinents (FP).

**Le rappel**

Le rappel (ou recall en anglais) est un **paramètre qui permet de mesurer le nombre de prévisions positives correctes sur le nombre total de données positives**. Il permet de répondre à la question suivante : sur tous les enregistrements positifs, combien ont été correctement prédits ? Dans l'exemple des courriers, cela représente le rapport entre le nombre de courriers pertinents reconnus comme tels sur le nombre réel de courriers pertinents. Celui-ci équivaut à la somme des vrais positifs (TP) et des faux négatifs (FN).

Pour ce système, le taux de rappel devrait dans l'idéal être de 100 %. Autrement dit, le nombre de faux négatifs devrait être égal à zéro. Il n'est en effet pas souhaitable que des courriers authentiques soient reconnus comme spams. Cela peut être très risqué. La métrique de rappel a donc une plus grande importance dans l'évaluation de la performance de ce système.

**L’accuracy**

L’accuracy est une métrique de performance qui évalue la capacité d’un modèle de classification à **bien prédire à la fois les individus positifs et les individus négatifs**.

L’accuracy permet de décrire la performance du modèle sur les individus positifs et négatifs de façon symétrique. Elle mesure le **taux de prédictions correctes** sur l’ensemble des individus

En tant qu'analyste scientifique, le choix de la métrique de performance pour une régression logistique dépend grandement du contexte de votre problème et des objectifs que vous souhaitez atteindre.

**Comprendre les métriques de base:**

* **Matrice de confusion:** C'est le point de départ pour calculer la plupart des autres métriques. Elle vous donne une vue d'ensemble des prédictions correctes et incorrectes, en les catégorisant en vrais positifs, faux positifs, vrais négatifs et faux négatifs.
* **Accuracy:** C'est la proportion globale de prédictions correctes. Cependant, elle peut être trompeuse si les classes sont déséquilibrées.
* **Precision:** Elle mesure la proportion de prédictions positives qui sont effectivement correctes.
* **Recall (sensibilité):** Elle mesure la proportion d'exemples positifs qui ont été correctement identifiés.
* **F1-score:** C'est la moyenne harmonique de la précision et du rappel, offrant un bon compromis entre les deux.

**Autres métriques:**

* **Courbe ROC:** Elle permet de visualiser le compromis entre le taux de vrais positifs et le taux de faux positifs pour différents seuils de classification.
* **AUC (Area Under the Curve):** C'est une mesure unique de la performance globale du modèle, basée sur la courbe ROC.

Puisque le coût d'une fausse prédiction positive ou négative est négligeable dans votre cas, l'**accuracy** (exactitude) semble être une métrique tout à fait appropriée pour évaluer les performances de votre modèle.

**Pourquoi l'accuracy est-elle un bon choix dans ce cas ?**

* **Simplicité:** Elle vous donne une mesure globale de la performance de votre modèle, en indiquant simplement le pourcentage de prédictions correctes.
* **Facilité d'interprétation:** L'accuracy est une métrique intuitive, facile à comprendre et à communiquer.

**Cependant, il est important de noter que:**

* **Déséquilibre des classes:** Si vos classes ne sont pas équilibrées (par exemple, si vous avez beaucoup plus de 0 que de 1), l'accuracy peut être trompeuse. Dans ce cas, il pourrait être plus pertinent d'examiner d'autres métriques comme la précision, le rappel ou le F1-score.
* **Autres aspects:** Même si l'accuracy est élevée, il peut être intéressant d'analyser plus en détail les performances de votre modèle en utilisant une matrice de confusion ou une courbe ROC. Cela vous permettra d'identifier les types d'erreurs que votre modèle commet le plus fréquemment.

**En résumé:**

Dans votre situation particulière où le coût des erreurs est nul, l'**accuracy** est une métrique simple et efficace pour évaluer votre modèle de régression logistique.