Certification « Développement et conduite de projets d'intelligence artificielle »

Notice

Le but de la certification est la confirmation de l'apprentissage du contenu dispensé lors des différentes formations. La certification repose sur un cas d'usage, que vous devrez préparer pour la soutenance, ainsi qu'un QCM.

Le cas d'usage a été imaginé pour vous mettre en situation au travers des sujets couverts au cours de votre parcours de formation. Il constituera le support principal lors de la soutenance et permettra ensuite d'ouvrir la discussion.

Ce cas d'usage aborde les sujets suivants :

- Cadrage d'un projet de data science
- Stockage et accès à la donnée
- Mise en place d'un modèle de data science
- Industrialisation d'un modèle

La cas d'usage se trouve dans ce document. Il est intitulé : OCTO FLY. Les attendus du cas d'usage sont détaillés à la fin de l'énoncé. Vous trouverez dans l'énoncé un lien vers les données. Un second jeu de données viendra compléter le premier au bout de 2 semaines. Le lien vous sera fourni par email.

La performance du code ou du modèle ne rentre pas en compte dans l'évaluation. Nous nous attendons à voir du code dans une librairie ou module englobant la phase de feature engineering et d'apprentissage du modèle.

Si vous avez des questions ou des remarques, nous nous tenons à votre entière disposition et répondrons dans les délais les plus brefs.

Déroulement de la certification :

- Cas d'usage à travailler qui servira de point départ pour la soutenance
- Une soutenance de 3h

Déroulement du passage de la soutenance :

- QCM autours des sujets abordés dans le parcours ~ 1h
- Restitution du cas d'usage 2h
    - Présentation de 20m

    - Analyse du code
    - Discussion libre autour du cas d'usage

Vos contacts :

- Baptiste O'Jeanson : bapo@octo.com
- Baptiste Saintot : basa@octo.com

Planning

Le planning de préparation de la certification est le suivant :

Date	**Commentaire**
05/05/2021	Envoi de l'énoncé et des données pour préparer la soutenance
20/05/2021	Envoi d'un complément de données
TBD	Passage de la soutenance

L'évaluation prendra en compte ce planning, notamment les délais courts.

Cas d’usage : OCTO FLY

Contexte

Les retards de vols coûtent des millions d’euros aux compagnies aériennes (correspondances ratées, trajet suivant aussi en retard, coût d’attente à l’aéroport, …) et génèrent beaucoup de frustration pour les usagers qui sont sévères envers les compagnies.

ACCENTO est une compagnie spécialisée dans l’analyse des données. Elle a pour principaux clients de grandes compagnies aériennes telles que BlaBlaJet, Creuse Airline ou LeBonAvion. Elle offre un large panel de services autour de la data : visualisation, et prédiction de retards et d’annulations de vols.

Les performances du modèle utilisé pour les prédictions ne sont plus alignées avec la réalité. Elle vous a donc engagé en tant que data scientist afin de construire un nouvel algorithme permettant la prédiction des retards.

Données disponibles

Le directeur de la DSI d’ACCENTO vous a mis à disposition une base de données contenant les informations de nombreux vols. Cette base de données a été reconstruite à partir d’un backup de notre base principale. La base de données est un fichier de base Sqlite (Python inclut une librairie pour lire ce type de fichier).

Vous pouvez contacter les personnes du métier via les contacts définis dans la notice de la soutenance.

Documentation des base de données

Vous trouverez les deux fichiers contenant des bases de données au format sqlite3. Une troisième base (au même format) vous sera fournie pendant la préparation.

Base de données	**Liens**

batch_1.db	https://drive.google.com/drive/folders/1Bejsu2cFhXamwy7CkJH0xJhKj2fK8JvD
test.db	dans le fichier zip disponible dans le lien ci-dessus
batch_2.db	Cette base sera disponible le 20/05/2021 Elle contient le même schéma que batch_1.db avec de nouvelles lignes.

Fichier batch_1.db :

Le fichier batch_1.db contient 4 tables :

- aeroports
- compagnies
- vols
- fuel

VOLS

Description

La table contient toutes les informations relatives aux vols.

Colonnes

Nom de la colonne	type
IDENTIFIANT	INTEGER
VOL	INTEGER
CODE AVION	TEXT
AEROPORT DEPART (code IATA)	TEXT
AEROPORT ARRIVEE (code IATA)	TEXT
DEPART PROGRAMME	INTEGER
HEURE DE DEPART	REAL
RETART DE DEPART	REAL
TEMPS DE DEPLACEMENT A TERRE AU DECOLLAGE	REAL
DECOLLAGE	REAL
TEMPS PROGRAMME	REAL

TEMPS PASSE	REAL
TEMPS DE VOL	REAL
DISTANCE	INTEGER
ATTERRISSAGE	REAL
TEMPS DE DÉPLACEMENT À TERRE A L'ATTERRISSAGE	REAL
ARRIVEE PROGRAMMEE	INTEGER
HEURE D'ARRIVEE	REAL
RETARD A L'ARRIVEE	REAL
DETOURNEMENT	INTEGER
ANNULATION	INTEGER
RAISON D'ANNULATION	TEXT
RETARD SYSTEM	REAL
RETARD SECURITE	REAL
RETARD COMPAGNIE	REAL
RETARD AVION	REAL
RETARD METEO	REAL
DATE	TEXT
NIVEAU DE SECURITE	INTEGER
COMPAGNIE AÉRIENNE	TEXT
NOMBRE DE PASSAGERS	INTEGER

COMPAGNIES

Description

La table contient toutes les informations relatives aux compagnies aériennes.

Colonnes

Nom de la colonne	type
COMPAGNIE	TEXT
CODE	TEXT
NOMBRE D EMPLOYES	INTEGER

CHIFFRE D AFFAIRE	INTEGER

AEROPORTS

Description

La table contient toutes les informations relatives aux aéroports.

Colonnes

Nom de la colonne	type
CODE IATA (code aéroport)	TEXT
NOM	TEXT
LIEU	TEXT
PAYS	TEXT
LONGITUDE	TEXT
LATITUDE	TEXT
HAUTEUR	REAL
PRIX RETARD PREMIERE 20 MINUTES	INTEGER
PRIS RETARD POUR CHAQUE MINUTE APRES 10 MINUTES	INTEGER

FUEL

Description

La table contient les prix du baril de kérosène d'avion sur la plage de période de la table `vol`

Colonnes

Nom de la colonne	type
DATE	TEXT
PRIX DU BARIL	REAL

Fichier `test.db` :

Ce jeu de données est un exemple de fichier fourni par l'un de nos clients. Elle contient en partie les mêmes informations que la table vol de la base Batch à l'exception de plusieurs colonnes. Cela vous laisse le loisir de choisir la colonne à prédire. Le fichier test.db contient une unique table :

- `vols`

VOLS

Description

La table contient toutes les informations relatives aux vols.

Colonnes

Nom de la colonne	type
IDENTIFIANT	INTEGER
VOL	INTEGER
CODE AVION	TEXT
AEROPORT DEPART	TEXT
AEROPORT ARRIVEE	TEXT
DEPART PROGRAMME	INTEGER
TEMPS DE DEPLACEMENT A TERRE AU DECOLLAGE	REAL
TEMPS PROGRAMME	REAL
DISTANCE	INTEGER
TEMPS DE DEPLACEMENT A TERRE A L'ATTERRISSAGE	REAL
ARRIVEE PROGRAMMEE	INTEGER
DATE	TEXT
NIVEAU DE SECURITE	INTEGER
COMPAGNIE AERIENNE	TEXT
NOMBRE DE PASSAGERS	INTEGER

Organisation du travail sur le projet

Avant la date d'entretien fixée, vous pouvez poser, par email, des questions au Directeur du Système d'Information.

Le Directeur de la DSI sera certainement amené à communiquer avec vous lors de la phase de conception.

Vous réaliserez, pour le jour de l'entretien avec OCTO, une présentation permettant de présenter en 20 minutes votre analyse du contexte et la solution technique que vous avez développée ainsi qu'une présentation des résultats. Le code fera l'objet d'une code review pendant la soutenance.

En l'absence éventuelle d'informations, vous êtes invité à prendre des hypothèses réalistes qui seront détaillées dans votre présentation.

Pour ce projet, il n'y a pas de bonne ou de mauvaise réponse. Il a pour objectif de mettre en lumière vos capacités à analyser une situation floue et complexe, à formuler des solutions adaptées au contexte et à restituer vos conclusions à l'écrit. Ce projet a pour but de mettre en application les savoirs acquis lors des formations. Une revue de code du projet est prévue lors de la soutenance. La performance du modèle ne rentre pas en compte dans la notation.

La suite de la soutenance portera sur des sujets de fonds liés aux sujets des différentes formations suivies lors du cursus.

But

Vous devez construire le nouveau projet de prévision de retard des vols. Votre objectif est de fournir un maximum d'informations que vous estimez utiles à notre clientèle.

Le livrable de ce projet sera composé des éléments suivants :

- Une présentation détaillant :
    - La démarche de cadrage expliquant :
        - Comment vous avez choisi la colonne à prédire
        - Le choix de la métrique d'optimisation
    - Une analyse du dataset (un rapide commentaire sur chacune des colonnes)

- Un module python respectant au maximum les recommandations présentées lors de la formation sur l'industrialisation d’un projet de machine learning et un modèle prédictif

- Les prédictions réalisées à partir de la base de tests dans un fichier csv contenant deux colonnes (ID, prédiction).

Bon courage !