Project description

ENSAE - Data Science & Machine Learning
2021 - 2022

Mously Diaw

Segmentation of the customers of an E-commerce site (1/2)

Vous êtes consultant pour Olist, une entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne.

Olist souhaite que vous fournissiez à ses équipes d'e-commerce une **segmentation des clients** qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.

Votre objectif est de **comprendre les différents types d'utilisateurs** grâce à leur comportement et à leurs données personnelles.

Vous devrez **fournir à l'équipe marketing une description actionable** de votre segmentation et de sa logique sous-jacente pour une utilisation optimale, ainsi qu'une **proposition de contrat de maintenance** basée sur une analyse de la stabilité des segments au cours du temps.

Votre mission

Votre mission est d'aider les équipes d'Olist à comprendre les différents types d'utilisateurs. Vous utiliserez donc des méthodes non supervisées pour regrouper des clients de profils similaires. Ces catégories pourront être utilisées par l'équipe Marketing pour mieux communiquer.

Segmentation of the customers of an E-commerce site (2/2)

Livrables

- Un **notebook** (ou code commenté au choix) **de l'analyse exploratoire & d'essais** des différentes approches de modélisation.
- Un **notebook de simulation pour déterminer la fréquence nécessaire de mise à jour** du modèle de segmentation.
- Un **support de présentation** pour présenter votre travail à un collègue.

Pour faciliter votre passage devant le jury, déposez sur la plateforme, dans un dossier nommé "<Class>_segmentation<NumeroGroupe>" (exemple: its4_segmentation02, cf fichier excel sur les groupes), tous les livrables du projet. Chaque livrable doit être nommé avec le numéro du projet et selon l'ordre dans lequel il apparaît, par exemple "segmentation_01_notebook*analyse*", "segmentation_02_notebook*essais*", et ainsi de suite. Ce dossier compressé doit être envoyé par mail avec l'objet: "**DS & ML: Projet final**"

Modalités de la soutenance

- 25 min de Présentation
  - 5 min - Présentation de la problématique, du cleaning effectué, du feature engineering et de l'exploration
  - 10 min - Présentation des différentes pistes de modélisation effectuées
  - 5 min - Présentation du modèle final sélectionné ainsi que des améliorations effectuées.
  - 5 min - Présentation de la simulation pour définir le délai de maintenance du modèle (contrat de maintenance)
- 5 à 10 min de questions-réponses.
- 5 min de debriefing à la fin de la soutenance

Prediction of Building energy (1/2)

La ville de Seattle s'intéresse de près aux émissions des bâtiments non destinés à l'habitation: **Prédiction des émissions de CO2**

Votre prédiction se basera sur les données déclaratives du permis d'exploitation commerciale (taille et usage des bâtiments, mention de travaux récents, date de construction..)

Vous cherchez également à **évaluer l'intérêt de l'ENERGY STAR Score pour la prédiction d'émissions**, qui est fastidieux à calculer avec l'approche utilisée actuellement par votre équipe.

Votre mission

Voici un récapitulatif de votre mission :

1. Réaliser une analyse exploratoire.
2. Tester différents modèles de prédiction afin de répondre au mieux à la problématique.

Faites bien attention au traitement des différentes variables, à la fois pour trouver de nouvelles informations (peut-on déduire des choses intéressantes d'une simple adresse ?) et optimiser les performances en appliquant des transformations simples aux variables (normalisation, passage au log, etc.).

Prediction of Building energy (2/2)

Seattle

Livrables attendus

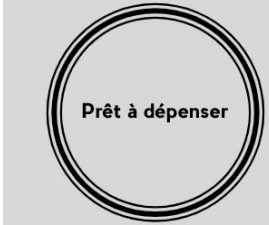
- Un **notebook** de l'analyse exploratoire mis au propre et annoté.
- Le **code** (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifierez clairement le modèle final choisi.
- Un support de **présentation** pour la soutenance.

Pour faciliter votre passage au jury, déposez dans un dossier nommé "<Class>_buildingenergy<NumeroGroupe>" (exemple: its4_buildingenergy01, cf fichier excel sur les groupes), tous les livrables du projet. Chaque livrable doit être nommé avec le numéro du projet et selon l'ordre dans lequel il apparaît, par exemple "buildingenergy_01_notebook", "buildingenergy_02_code", et ainsi de suite. Ce dossier compressé doit être envoyé par mail avec l'objet: "**DS & ML: Projet final**"

Modalités de la soutenance

- 25 min de Présentation
  - 5 min - Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.
  - 5 min - Présentation du cleaning effectué, du feature engineering et de l'exploration.
  - 10 min - Présentation des différentes pistes de modélisation effectuées.
  - 5 min - Présentation du modèle final sélectionné ainsi que des améliorations effectuées.
- 5 à 10 min de questions-réponses.
- 5 min de debrief à la fin de la soutenance

Home credit default risk (1/2)



Vous êtes Data Scientist au sein d'une société financière, nommée **"Prêt à dépenser"**, qui propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt.

L'entreprise souhaite **mettre en œuvre un outil de "scoring crédit" pour calculer la probabilité** qu'un client rembourse son crédit, puis classifie la demande en crédit accordé ou refusé. Elle souhaite donc développer un **algorithme de classification** en s'appuyant sur des sources de données variées (données comportementales, données provenant d'autres institutions financières, etc.).

Vous aurez sûrement besoin de joindre les différentes tables entre elles.

De plus, les chargés de relation client ont fait remonter le fait que les clients sont de plus en plus demandeurs de **transparence** vis-à-vis des décisions d'octroi de crédit.

Votre mission

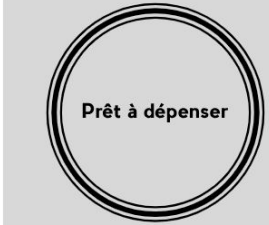
1. Construire un modèle de scoring qui donnera une prédiction sur la probabilité de faillite d'un client de façon automatique.

2. Visualiser le score et l'interprétation de ce score pour chaque client de façon intelligible pour une personne non experte en data science.

3. Construire un dashboard interactif à destination des gestionnaires de la relation client permettant d'interpréter les prédictions faites par le modèle, et d'améliorer la connaissance client des chargés de relation client.

Vous pourrez ainsi vous focaliser sur l'élaboration du modèle, son optimisation et sa compréhension. A cet effet, je vous incite à sélectionner un kernel Kaggle pour vous faciliter la préparation des données nécessaires à l'élaboration du modèle de scoring. Vous analyserez ce kernel et l'adapterez pour vous assurer qu'il répond aux besoins de votre mission.

Home credit default risk (2/2)



Livrables

- Un **notebook** de l'analyse exploratoire mis au propre et annoté.
- Le **code** (ou un notebook) des différents tests de modèles mis au propre (fonction cout métier, optimisation, métrique d'évaluation, interprétabilité globale/locale du modèle final), dans lequel vous identifierez clairement le modèle final choisi.
- Un support de **présentation** pour la soutenance, détaillant le travail réalisé.

Pour faciliter votre passage au jury, déposez sur la plateforme, dans un dossier nommé "<*Classe*>_*creditdefault*<*NumeroGroupe*>" (exemple: ise2_creditdefault02, cf fichier excel sur les groupes), tous les livrables du projet. Chaque livrable doit être nommé avec le numéro du projet et selon l'ordre dans lequel il apparaît, par exemple "*creditdefaul*t_01_notebookanalyse", "*creditdefaul*t_02_notebookessai", et ainsi de suite. Ce dossier compressé doit être envoyé par mail avec l'objet: "**DS & ML: Projet final**"

Modalités de la soutenance

- 25 min de Présentation
  - 5 min - Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.
  - 5 min - Présentation du cleaning effectué, du feature engineering et de l'exploration.
  - 10 min - Présentation des différentes pistes de modélisation effectuées.
  - 5 min - Présentation du modèle final sélectionné ainsi que des améliorations effectuées.
- 5 à 10 min de questions-réponses.
- 5 min de debrief à la fin de la soutenance

Natural Language Processing with Disaster Tweets (1/2)

Vous êtes consultant pour Twitter, réseau social. Twitter est devenu un important canal de communication en cas d'urgence.

L'omniprésence des smartphones permet aux gens d'annoncer une urgence qu'ils observent en temps réel. C'est pourquoi de plus en plus d'organismes s'intéressent à la surveillance programmatique de Twitter (par exemple, les organisations de secours en cas de catastrophe et les agences de presse).

Mais il n'est pas toujours évident de savoir si les mots d'une personne annoncent réellement une catastrophe. A cet effet, Twitter souhaite avoir un modèle de détection des tweets qui annoncent des catastrophes.

Votre mission

- Vous devez construire un modèle d'apprentissage automatique qui prédit quels tweets sont liés à des catastrophes réelles et lesquels ne le sont pas.

Natural Language Processing with Disaster Tweets (2/2)

Livrables

- Un **notebook** de l'analyse exploratoire mis au propre et annoté.
- Le **code** (ou un notebook) des différents tests de modèles mis au propre (optimisation, métrique d'évaluation, interprétabilité du modèle final), dans lequel vous identifierez clairement le modèle final choisi.
- Un support de **présentation** pour la soutenance, détaillant le travail réalisé.

Pour faciliter votre passage au jury, déposez sur la plateforme, dans un dossier nommé "<Classe>_disastertweets<NumeroGroupe>" (exemple: ise2_disastertweets11, cf fichier excel sur les groupes), tous les livrables du projet. Chaque livrable doit être nommé avec le numéro du projet et selon l'ordre dans lequel il apparaît, par exemple "disastertweets_01_notebookanalyse", "disastertweets_02_notebookessai", et ainsi de suite. Ce dossier compressé doit être envoyé par mail avec l'objet: "**DS & ML: Projet final**"

Modalités de la soutenance

- 25 min de Présentation
  - 5 min - Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.
  - 5 min - Présentation du cleaning effectué, du feature engineering et de l'exploration.
  - 10 min - Présentation des différentes pistes de modélisation effectuées.
  - 5 min - Présentation du modèle final sélectionné ainsi que des améliorations effectuées.
- 5 à 10 min de questions-réponses.
- 5 min de debrief à la fin de la soutenance

Bonus (facultatif)

- **Dashboard** (dash, streamlit, ...): construire un dashboard interactif permettant d'interpréter les prédictions faites par le modèle, et d'améliorer la connaissance client.
- Déployer sur le cloud (heroku, ...) le modèle sous forme d'**API** (flask, fast api, etc...)