Note Technique

Prediction des Tarifs de Taxi a New York

Projet Machine Learning

Auteurs

Ayoub REZALA
Yassine FRIKICH

Module

Concepts et Technologies IA

Janvier 2026

Table des matières

1 Analyse des Besoins Metiers

1.1 Contexte et Secteur d'Activite

Le secteur du transport de personnes par taxi constitue un pilier essentiel de la mobilite urbaine, particulierement dans les grandes metropoles mondiales. New York City, avec ses cinq boroughs et ses 8,3 millions d'habitants, represente l'un des marches de taxi les plus importants au monde.

La Taxi and Limousine Commission (TLC) de New York regule environ 13 000 taxis jaunes (Yellow Cabs) et supervise plus de 80 000 vehicules de location. En 2018, annee de reference de notre etude, le secteur a enregistre plus de 100 millions de courses, generant un chiffre d'affaires de plusieurs milliards de dollars.

1.2 Problematique Identifiee

1.2.1 Description du Probleme

La tarification des courses de taxi a New York repose sur un systeme complexe combinant plusieurs facteurs :

— Tarif de base initial
— Tarification au kilometre parcouru
— Tarification au temps d'attente
— Supplements (heures de pointe, nuit, peages)
— Taxes diverses (MTA tax, improvement surcharge)

Cette complexite pose plusieurs defis :

1. **Pour les clients** : Difficulte a estimer le cout d'une course avant le depart
2. **Pour les chauffeurs** : Optimisation difficile des revenus journaliers
3. **Pour les gestionnaires de flotte** : Allocation inefficace des ressources

1.2.2 Preuve de l'Existence du Probleme

L'analyse de notre dataset de 8 millions de courses revele :

— Une variance importante des tarifs pour des distances similaires
— Un ecart-type eleve sur la variable `fare_amount`
— Des correlations significatives entre plusieurs variables et le tarif final

1.3 Impact sur l'Activite

1.3.1 Impact Financier

Indicateur	**Valeur**
Nombre de courses analysees	8 000 000+
Tarif moyen par course	12.50 $
Ecart-type du tarif	10.20 $
Pourboire moyen	1.85 $

TABLE 1 – Statistiques cles du dataset

Une meilleure prediction des tarifs permettrait :
— Reduction des litiges client-chauffeur (estimes a 2-3% des courses)
— Amelioration de la satisfaction client
— Optimisation de la planification des courses

1.3.2 Impact Operationnel

La capacite a predire le tarif d'une course permet :
— Estimation en temps reel pour les applications mobiles
— Aide a la decision pour les chauffeurs (choix des zones)
— Planification budgetaire pour les entreprises clientes

1.4 Solution Proposee

Nous proposons un modele de Machine Learning base sur la regression lineaire pour predire le tarif d'une course de taxi en fonction de plusieurs variables explicatives. Cette approche presente les avantages suivants :
— **Simplicite** : Modele interpretable et explicable
— **Rapidite** : Prediction en temps reel possible
— **Scalabilite** : Applicable a grande echelle
— **Cout** : Faible cout computationnel

2 Conformite Legale et Reglementaire

2.1 Cadre Juridique

2.1.1 Non-Application du RGPD

Le Reglement General sur la Protection des Donnees (RGPD) est une reglementation europeenne qui s'applique :
— Aux entreprises etablies dans l'Union Europeenne
— Au traitement de donnees de residents europeens

Notre dataset n'est pas soumis au RGPD pour les raisons suivantes :

1. **Origine geographique** : Les donnees proviennent de la Taxi and Limousine Commission de New York, une agence gouvernementale americaine.
2. **Anonymisation** : Le dataset public ne contient aucune donnee permettant d'identifier directement ou indirectement les passagers ou les chauffeurs :
   — Pas de noms ou identifiants personnels
   — Pas de numeros de telephone
   — Pas d'adresses email
   — Pas de numeros de carte bancaire
3. **Donnees agregees** : Les coordonnees GPS de depart et d'arrivee sont arrondies et ne permettent pas de tracer des individus specifiques.
4. **Dataset public** : Ces donnees sont mises a disposition du public par la ville de New York dans le cadre de sa politique Open Data.

2.1.2 Cadre Juridique Applicable

Les donnees sont soumises a la reglementation americaine, notamment :
— **Freedom of Information Law (FOIL)** de l'Etat de New York
— **NYC Open Data Law** (Local Law 11 of 2012)
Ces reglementations encouragent la publication de donnees gouvernementales pour favoriser la transparence et l'innovation.

2.2 Utilisation Ethique des Donnees

Bien que non soumis au RGPD, nous nous engageons a :
— Utiliser les donnees uniquement a des fins academiques et de recherche
— Ne pas tenter de re-identifier des individus
— Citer la source des donnees (NYC TLC)
— Respecter les conditions d'utilisation du dataset

3 Analyse Exploratoire des Donnees

3.1 Description du Dataset

3.1.1 Source et Origine

Attribut	Valeur
Source	NYC Taxi and Limousine Commission
Annee	2018
Format	CSV
Taille	8 000 000+ enregistrements
Colonnes	17 variables
URL	https://www.kaggle.com/datasets/neilclack/nyc-taxi-trip-data-google-public

TABLE 2 – Informations sur le dataset

3.1.2 Variables Disponibles

Variable	Type	Description
fare_amount	Float	Tarif de la course (variable cible)
trip_distance	Float	Distance parcourue en miles
trip_duration	Integer	Duree du trajet en secondes
rate_code	Integer	Type de tarification (1-6)
payment_type	Integer	Mode de paiement (1=CB, 2=Cash)
tip_amount	Float	Montant du pourboire
tolls_amount	Float	Montant des peages
mta_tax	Float	Taxe MTA
extra	Float	Supplements divers
pickup_longitude	Float	Longitude de depart
pickup_latitude	Float	Latitude de depart
dropoff_longitude	Float	Longitude d'arrivee

Variable	Type	Description
dropoff_latitude	Float	Latitude d'arrivee
passenger_count	Integer	Nombre de passagers

TABLE 3: Description des variables du dataset

3.2 Statistiques Descriptives

3.2.1 Variables Numeriques Principales

Variable	Moyenne	Mediane	Ecart-type	Min	Max
fare_amount	12.50	9.50	10.20	0.01	500+
trip_distance	2.90	1.71	3.70	0.00	100+
trip_duration	850	660	700	0	7200+
tip_amount	1.85	1.45	2.50	0.00	100+

TABLE 4 – Statistiques descriptives des variables principales

3.3 Analyse des Valeurs Manquantes

L'analyse du dataset revele une excellente qualite des donnees :
— Aucune valeur manquante sur les variables principales
— Dataset pre-nettoye par la source

3.4 Analyse des Correlations

3.4.1 Correlations avec la Variable Cible

L'analyse des correlations avec `fare_amount` revele :

Variable	Correlation	Force
trip_distance	+0.94	Tres forte
trip_duration	+0.82	Tres forte
tolls_amount	+0.32	Moderee
tip_amount	+0.28	Faible
rate_code	+0.15	Faible
payment_type	-0.06	Tres faible

TABLE 5 – Correlations avec fare_amount

3.4.2 Interpretation

La forte correlation entre `trip_distance` et `fare_amount` (r = 0.94) confirme que la distance est le principal determinant du tarif, conformement a la structure tarifaire officielle des taxis new-yorkais.

4 Experimentation Machine Learning

4.1 Methodologie

4.1.1 Approche Adoptee

Nous avons adopte une approche supervisee de regression pour predire la variable continue `fare_amount`. Le choix de la regression lineaire se justifie par :

— La relation lineaire observee entre distance et tarif
— L'interpretabilite du modele (coefficients explicables)
— La rapidite d'entrainement et de prediction
— La facilite de mise en production

4.1.2 Pipeline de Traitement

1. **Echantillonnage** : Selection de 100 000 lignes (contrainte computationnelle)
2. **Nettoyage** : Suppression des valeurs aberrantes
3. **Selection des features** : Choix des variables explicatives
4. **Division** : Split train/test (80%/20%)
5. **Entrainement** : Regression lineaire
6. **Evaluation** : Calcul des metriques

4.2 Preparation des Donnees

4.2.1 Nettoyage des Donnees

Criteres de filtrage appliques pour supprimer les outliers :

```
df_clean = df_sample[
    (df_sample['fare_amount'] > 0) &
    (df_sample['fare_amount'] < 200) &
    (df_sample['trip_distance'] > 0) &
    (df_sample['trip_distance'] < 100) &
    (df_sample['trip_duration'] > 60) &
    (df_sample['trip_duration'] < 7200)
].copy()
```

Justification des seuils :

— `fare_amount` $<$ 200$: Exclusion des tarifs extremes (erreurs de saisie)
— `trip_distance` $<$ 100 miles : Courses realistes dans NYC
— `trip_duration` entre 1 min et 2h : Durees coherentes

4.2.2 Selection des Features

Variables retenues pour le modele :

```
features = ['trip_distance', 'rate_code', 'payment_type',
            'extra', 'mta_tax', 'tip_amount']
```

4.3 Modelisation

4.3.1 Modele de Regression Lineaire

Le modele de regression lineaire suppose une relation de la forme :

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{1}$$

Ou :
— $\hat{y}$: tarif predit (fare_amount)
— β_0 : intercept (tarif de base)
— β_i : coefficients des features
— x_i : valeurs des features

4.3.2 Division Train/Test

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
```

— Ensemble d'entrainement : 80% (environ 80 000 lignes)
— Ensemble de test : 20% (environ 20 000 lignes)
— Random state fixe pour reproductibilite

4.4 Evaluation du Modele

4.4.1 Metriques Utilisees

Nous utilisons trois metriques complementaires pour evaluer les performances :

R-Squared (R2) - Coefficient de Determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

— **Ce que mesure R2** : Proportion de la variance de la variable cible expliquee par le modele
— **Interpretation** : Valeur entre 0 et 1 ; R2 = 0.85 signifie que 85% de la variance est expliquee
— **Pertinence** : Permet de comparer le pouvoir explicatif global du modele
— **Limites** : Peut etre artificiellement eleve avec beaucoup de features ; ne mesure pas la precision en unites interpretables

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

— **Ce que mesure RMSE** : Erreur moyenne de prediction en unites de la variable cible

— **Interpretation** : RMSE = 3.50$ signifie une erreur moyenne de 3.50$ par prediction
— **Pertinence** : Donne une mesure directement interpretable en dollars
— **Limites** : Sensible aux outliers (erreurs au carre) ; dependant de l'echelle

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

— **Ce que mesure MAE** : Erreur absolue moyenne de prediction
— **Interpretation** : MAE = 2.50$ signifie une erreur moyenne de 2.50$
— **Pertinence** : Plus robuste aux outliers que RMSE
— **Limites** : Ne penalise pas les grandes erreurs autant que RMSE

Complementarite des Metriques L'utilisation conjointe de ces trois metriques permet :
— R2 : Vision globale du pouvoir explicatif
— RMSE : Detection des grandes erreurs (si RMSE » MAE, presence d'outliers mal predits)
— MAE : Erreur typique attendue pour une prediction

4.4.2 Resultats Obtenus

Metrique	Valeur	Interpretation
R2 Score	0.92	92% de variance expliquee
RMSE	3.20 $	Erreur quadratique moyenne
MAE	2.10 $	Erreur absolue moyenne

TABLE 6 – Resultats de l'evaluation du modele

4.4.3 Interpretation des Resultats

— **R2 = 0.92** : Excellent pouvoir explicatif ; le modele capture tres bien la variance des tarifs
— **RMSE = 3.20$** : Erreur acceptable pour des courses moyennes de 12$
— **RMSE proche de MAE** : Peu de predictions avec des erreurs extremes

4.5 Coefficients du Modele

4.5.1 Formule de Regression

$$fare\_amount = 4.12 + 2.65 \times trip\_distance + 0.85 \times rate\_code + ... \tag{5}$$

4.5.2 Interpretation des Coefficients

Variable	Coefficient	Interpretation
Intercept	4.12	Tarif de base (prise en charge)
trip_distance	2.65	+2.65$ par mile parcouru
rate_code	0.85	Impact du type de tarification
tip_amount	0.42	Correlation positive avec le tarif
extra	1.05	Impact des supplements
mta_tax	2.10	Impact de la taxe MTA

TABLE 7 – Coefficients du modele de regression

4.6 Reproductibilite

4.6.1 Environnement Technique

```
# Fichier requirements.txt
pandas==2.0.0
numpy==1.24.0
scikit-learn==1.3.0
matplotlib==3.7.0
seaborn==0.12.0
jupyter==1.0.0
```

4.6.2 Instructions de Reproduction

1. Cloner le repository :

```
git clone https://github.com/yassinefri/Campaign-Analytics-
    Platform.git
cd Campaign-Analytics-Platform
```

2. Creer l'environnement virtuel :

```
python -m venv venv
source venv/bin/activate  # Linux/Mac
venv\Scripts\activate     # Windows
```

3. Installer les dependances :

```
pip install -r requirements.txt
```

4. Executer les notebooks :

```
jupyter notebook
```

4.6.3 Structure du Projet

```
Campaign-Analytics-Platform/
    README.md
    requirements.txt
    nyc_taxi_analysis.ipynb       # Analyse exploratoire
    fare_prediction_model.ipynb   # Modelisation ML
    fare_prediction_model.pkl     # Modele sauvegarde
    datasets/
        original_cleaned_nyc_taxi_data_2018.csv
```

5 Metriques et KPIs a Surveiller

5.1 KPIs de Performance du Modele

KPI	Seuil Acceptable	Seuil Optimal	Frequence
R2 Score	> 0.80	> 0.90	Mensuel
RMSE	< 5.00\$	< 3.00\$	Hebdomadaire
MAE	< 4.00\$	< 2.50\$	Hebdomadaire

TABLE 8 – KPIs de performance a surveiller

5.2 KPIs de Stabilite

Pour assurer la stabilite du modele en production :
— **Data Drift** : Surveiller les changements dans la distribution des features
— **Concept Drift** : Detecter les changements dans la relation feature-target
— **Prediction Drift** : Monitorer la distribution des predictions

5.3 KPIs Metier

— Taux d'erreur de prediction $> 20\%$ du tarif reel
— Nombre de predictions aberrantes par jour
— Satisfaction client (si feedback disponible)

6 Estimation des Couts de Deploiement Azure

6.1 Architecture Proposee

Pour le deploiement de notre solution, nous proposons une architecture Azure comprenant :

1. **Azure Machine Learning** : Hebergement et serving du modele
2. **Azure Container Instances** : Deploiement du conteneur d'inference
3. **Azure Blob Storage** : Stockage des donnees et du modele
4. **Azure Application Insights** : Monitoring et logs

6.2 Estimation des Couts Mensuels

Service	Configuration	Cout/Mois
Azure ML Workspace	Basic	0 $ (gratuit)
Container Instances	1 vCPU, 1.5 GB RAM	35 $
Blob Storage	10 GB, LRS	2 $
Application Insights	5 GB logs/mois	12 $
Total		**49 $/mois**

TABLE 9 – Estimation des couts Azure (usage leger)

6.3 Scenarios de Charge

6.3.1 Scenario 1 : Usage Leger (< 10 000 predictions/jour)

— Container Instances : 1 instance
— Cout estime : 49 $/mois

6.3.2 Scenario 2 : Usage Moyen (10 000 - 100 000 predictions/jour)

— Azure Kubernetes Service : 2-3 pods
— Cout estime : 150-250 $/mois

6.3.3 Scenario 3 : Usage Intensif (> 100 000 predictions/jour)

— Azure Kubernetes Service avec auto-scaling
— Azure Cache for Redis
— Cout estime : 500-1000 $/mois

6.4 Optimisation des Couts

Recommandations pour reduire les couts :
— Utiliser les instances Spot pour les entrainements
— Activer l'auto-scaling avec scale-to-zero
— Choisir la region Azure la moins chere
— Beneficier des credits Azure for Students si applicable

7 Pistes d'Amelioration

7.1 Ameliorations du Modele

1. **Feature Engineering** :
   — Ajouter des features temporelles (heure, jour de la semaine)
   — Calculer la distance euclidienne a partir des coordonnees GPS
   — Creer des features de zone (Manhattan, Brooklyn, etc.)
2. **Algorithmes Alternatifs** :
   — Random Forest pour capturer les non-linearites
   — Gradient Boosting (XGBoost, LightGBM)

— Reseaux de neurones pour les patterns complexes

3. **Validation** :
   — Implementer la validation croisee k-fold
   — Tester sur des donnees plus recentes (2019-2024)

7.2 Ameliorations de la Pipeline

— Automatiser le re-entrainement periodique
— Implementer des tests unitaires et d'integration
— Ajouter un systeme d'alerte en cas de degradation des performances

8 Conclusion

Ce projet a permis de developper une solution de prediction des tarifs de taxi a New York basee sur le Machine Learning. Les principaux resultats sont :

— Un modele de regression lineaire avec un R2 de 0.92
— Une erreur moyenne de prediction (MAE) de 2.10$
— Une solution reproductible et documentee
— Un cout de deploiement estime a 49$/mois pour un usage leger

La simplicite du modele de regression lineaire permet une mise en production rapide tout en offrant des performances satisfaisantes. Les pistes d'amelioration identifiees permettront d'augmenter la precision si necessaire.

A Annexe : Resume des Metriques d'Evaluation

A.1 Recapitulatif

Metrique	Description
R2	Proportion de variance expliquee (0 a 1)
RMSE	Racine de l'erreur quadratique moyenne (en $)
MAE	Erreur absolue moyenne (en $)

A.2 Limites de l'Evaluation

— Le modele assume une relation lineaire entre les features et la cible
— Les metriques sont calculees sur un echantillon de 100 000 lignes
— Pas de validation croisee implementee
— Les donnees datent de 2018 (potentiel decalage temporel)
— L'heteroscedasticite des residus n'a pas ete formellement testee