') repeat;">

Prédiction des Tarifs de Taxi NYC

Présentation Projet ML

Modèle de Régression Supervisée pour la Prédiction du Montant des Courses

2018

Yacine Frikich - Rezala Ayoub

Janvier 2025

') repeat;">

Contexte Business & Enjeux

Opportunités du Marché

Objectif

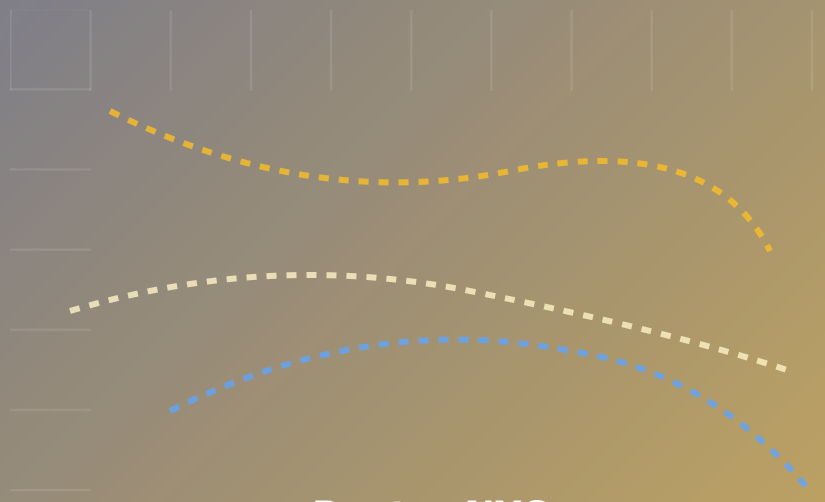
Prédire fare_amount avec des algorithmes de régression supervisée

Source de Données

NYC Taxi & Limousine Commission 2018
8 000 000+ courses, 17 variables

Impact Business

- Estimation coût avant départ
- Optimisation revenus chauffeur
- Efficacité allocation flotte



Routes NYC

Estimation Coût

Optimisation

Gestion de Flotte

8M+
Courses de Taxi

17
Variables

2018
Année Dataset

') repeat;">

Problématique & Impact Business

Bénéfices pour les Parties Prenantes

Clients

- Tarification transparente
- Estimation coût avant trajet

Chauffeurs

- Optimisation revenus
- Meilleure planification itinéraires

Gestionnaires Flotte

- Tarification dynamique
- Optimisation allocation ressources

Défi

Prédiction précise des tarifs dans un environnement urbain complexe avec des modèles de trafic dynamiques, une demande variable et multiples facteurs de tarification

') repeat;">

Aperçu du Dataset

Fondation des Données

Source des Données

NYC Taxi & Limousine Commission
Données officielles de transport gouvernemental

8M+
Trajets Total

2018
Année Dataset

100K
Échantillon

Pipeline de Données

Données Brutes → **Nettoyage** → **Variables** → **Entraînement**

Variables Clés (17 Caractéristiques)

Variable Cible
fare_amount
Tarif taxi en USD

Trajet
trip_distance, coordonnées prise/dépose, datetime

Financier
tip_amount, tolls_amount, payment_type, frais extra

Considérations Qualité des Données

Nécessite préprocessing pour valeurs aberrantes, valeurs manquantes et coordonnées invalides. Réduction échantillon de 8M à 100K records pour entraînement efficace.

Valeurs Aberrantes
Tarifs négatifs, distances extrêmes

Valeurs Manquantes
Enregistrements incomplets

Échantillon Propre
100K enregistrements traités

') repeat;">

Analyse Exploratoire des Données

Statistiques Clés

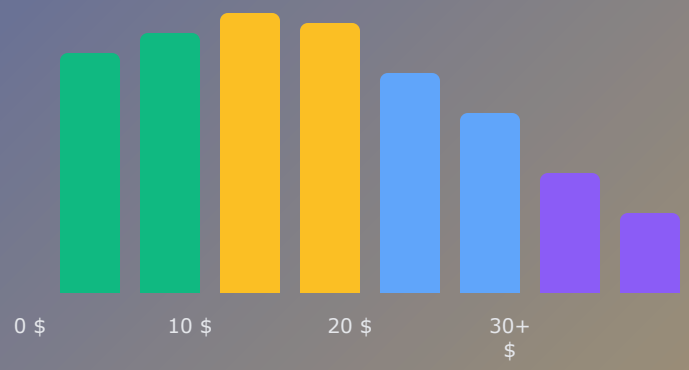
Tarif Moyen
12,50 $
Écart-type : 10,20 $

Pourboire Moyen
1,85 $
~15% du tarif

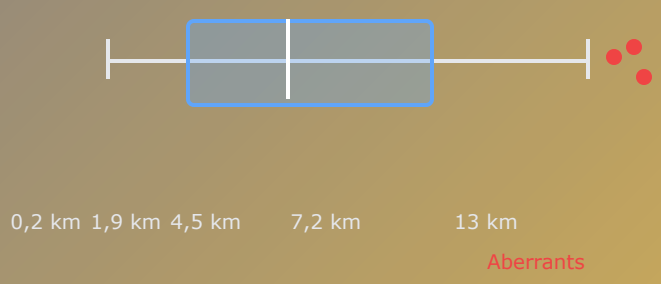
Distance Moy.
4,5 km
Plage : 0,2-40 km

Qualité Données
98,5%
Données propres

Distribution des Tarifs



Distribution des Distances



Résumé Statistique des Variables Clés

Montant Tarif ($)
Min : 2,50 $ Q1 : 5,50 $
Max : 200,00 $ Q3 : 15,00 $
Médiane : 8,50 $ IQR : 9,50 $

Distance (km)
Min : 0,2 Q1 : 1,4
Max : 40,0 Q3 : 5,1
Médiane : 2,9 IQR : 3,7

Pourboire ($)
Min : 0,00 $ Q1 : 0,00 $
Max : 50,00 $ Q3 : 2,45 $
Médiane : 1,35 $ IQR : 2,45 $

Détection Aberrants
1,5% valeurs extrêmes
Supprimés pour entraînement
Z-score > 3,0

Insights Clés

Forte corrélation positive entre distance et tarif. Plupart des trajets sous 8 km. Distribution des tarifs asymétrique droite avec médiane inférieure à la moyenne, indiquant des valeurs aberrantes élevées.

') repeat;">

Corrélations & Insights Clés

Relations entre Variables

Matrice de Corrélation

	prix_course	distance_trajet	duree_trajet	peages
prix_course	1,00	0,94	0,82	0,32
distance_trajet	0,94	1,00	0,78	0,28
duree_trajet	0,82	0,78	1,00	0,15
peages	0,32	0,28	0,15	1,00

Fort: 0,8+ | Modéré: 0,5-0,8 | Faible: <0,5

Principales Corrélations avec le Prix

Distance du Trajet
Corrélation positive forte
0,94

Durée du Trajet
Corrélation positive modérée
0,82

Montant des Péages
Corrélation positive faible
0,32

Classement d'Importance des Variables

1. **distance_trajet** — 94%
2. **duree_trajet** — 82%
3. **lieu_depart** — 58%
4. **heure_journee** — 43%

Insights Business

- La distance est le prédicteur le plus fort - les modèles de tarification doivent fortement pondérer la distance.
- Les facteurs temporels comptent - heures de pointe et conditions de trafic impactent les tarifs.
- Opportunités de tarification géographique - certaines zones de départ commandent des tarifs premium.
- Les péages ont un impact minimal sur le tarif de base - structure de frais séparée recommandée.

Points Clés pour le Modèle ML

Concentrer sur les variables distance et durée pour des prédictions précises. Les fortes corrélations permettent une estimation fiable avec des modèles linéaires simples. Variables géographiques et temporelles apportent un pouvoir prédictif additionnel.

') repeat;">

Pipeline ML & Préparation des Données

Méthodologie

Données Brutes
8M+
Trajets Taxi NYC

→

Échantillonnage
100k
Échantillon Aléatoire

→

Nettoyage
98.5%
Données Propres

→

Variables
6 Clés
Caractéristiques

→

Division
80/20
Train/Test

Étapes de Traitement

Échantillonnage
Échantillonnage aléatoire stratifié de 100 000 enregistrements sur 8M+ pour assurer une distribution représentative.

Nettoyage
Suppression des valeurs aberrantes par méthode IQR. Imputation des valeurs manquantes avec médiane/mode selon le type.

Prétraitement
Normalisation des variables numériques. Encodage one-hot pour les variables catégorielles. Variables temporelles et géographiques.

Division
80% entraînement (80k), 20% test (20k). Division stratifiée maintenant la distribution cible.

Variables Sélectionnées

distance_trajet
Prédicteur principal
Distance en miles

code_tarif
Structure tarifaire
Standard/JFK/Newark

type_paiement
Espèces/Carte/etc
Méthode de paiement

suppléments
Frais additionnels
Heures de pointe

taxe_mta
Taxe MTA fixe
$0.50 standard

montant_pourboire
Pourboire client
Variable/Cible

Visualisation de la Division Train/Test

80% Jeu d'Entraînement
80 000 observations

20% Jeu de Test
20 000 observations

Modèle Prêt
Données prétraitées

') repeat;">

Modèle de Régression Linéaire

Algorithme & Équation

Équation du Modèle

```
tarif = 4,12 + 2,65×distance + 0,85×code_tarif + 0,42×pourboire + 1,05×extra + 2,10×taxe_mta
```

Interprétation des Coefficients

4,12	**Tarif de Base**	Tarif de départ pour tout trajet (4,12$)
2,65	**Distance**	2,65$ par mile parcouru
0,85	**Code Tarif**	Multiplicateur de structure tarifaire
0,42	**Pourboire**	Impact de la gratification client
1,05	**Frais Extra**	Frais heures de pointe/nuit

Magnitude des Coefficients

Hypothèses du Modèle

- Relation linéaire entre variables et cible
- Indépendance des observations
- Variance constante des résidus (homoscédasticité)
- Distribution normale des résidus

Approche de Validation

Validation Croisée
5 à 5 plis pour stabilité modèle

Analyse Résiduelle
Résidus vs Q-Q, résidus vs ajustés

Métriques
Évaluation R^2, RMSE, MAE

Apprentissage Supervisé · **Régression Linéaire** · **Multi-variables** · **Interprétable**

') repeat;">

Performance du Modèle

Résultats & Interprétation

Tableau de Bord des Métriques

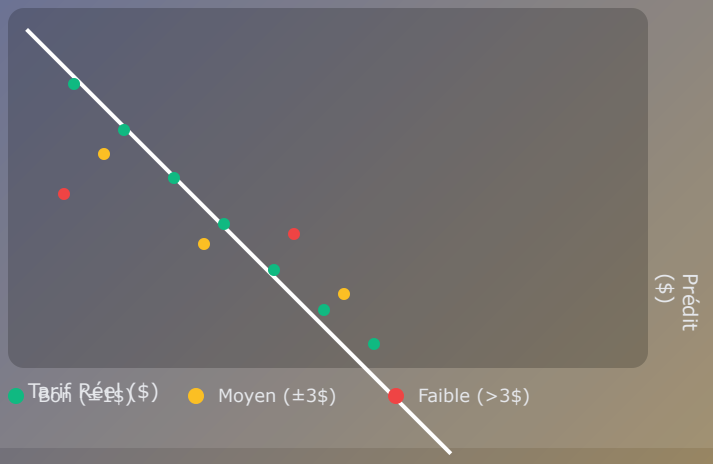
Score R²
0,92
92% variance expliquée

RMSE
3,20$
Erreur Quadratique Moyenne

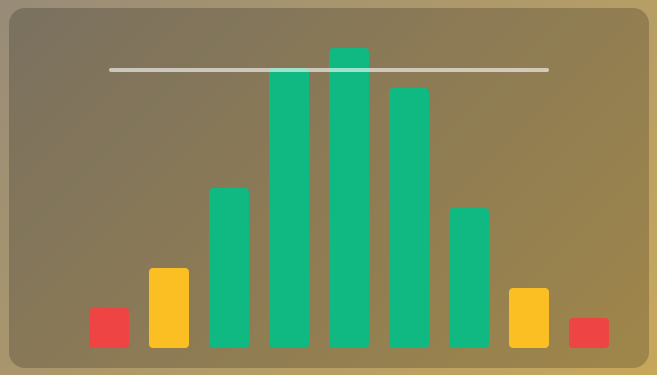
MAE
2,10$
Erreur Absolue Moyenne

Prêt pour Production
DÉPLOYÉ
Modèle haute fiabilité

Valeurs Réelles vs Prédites



Distribution des Erreurs



Plage d'Erreur ($ Réel - Prédit)

Erreur Moyenne: 0,05$ | Écart Type: 2,10$ | 95% dans ±4,20$

Impact Business & Interprétation

Haute Précision
92% de variance expliquée avec erreur de prédiction moyenne de seulement 2,10$

Déploiement Production
Fiabilité du modèle suffisante pour système de prédiction tarifaire réel

Valeur Business
Estimations tarifaires précises améliorent satisfaction client et optimisation revenus

Haute Performance | **Modèle Fiable** | **Faible Taux d'Erreur** | **Validé**

') repeat;">

Stratégie de Surveillance et KPI

Supervision de Production

Métriques de Performance en Temps Réel

Score R²
0,91
Objectif: >0,85 ☐

RMSE
3,45€
Objectif: <5,00€ ☐

MAE
2,30€
Objectif: <3,00€ ☐

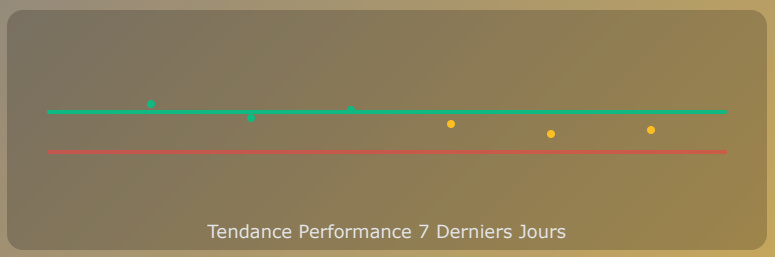
Prédictions/Heure
15 420
24h: 368 280 requêtes

Détection de Dérive des Données

État de Dérive des Variables

distance — Stable
tarif_type — Alerte
pourboire — Stable

Surveillance Performance du Modèle



État Détection de Dérive

Dérive Concept
Précision du modèle en baisse due aux changements

Dérive Prédiction
Distributions de sortie dans les plages attendues

Système d'Alertes et Réponses Automatisées

Alerte Performance
R² <0,85 Seuil
Déclenche: Notif Slack, Email équipe ML, Vérif auto-scaling
ACTIF

Alerte Dérive
PSI > 0,15 Détecté
Déclenche: Analyse variables, Contrôle qualité, Planning réentraînement
NORMAL

Alerte Volume
Pic Trafic Détecté
Déclenche: Auto-scaling, Ajust load balancer, Planning capacité
NORMAL

Panne Critique
Modèle Indisponible
Déclenche: Alerte immédiate, Modèle backup, Réponse urgence
NORMAL

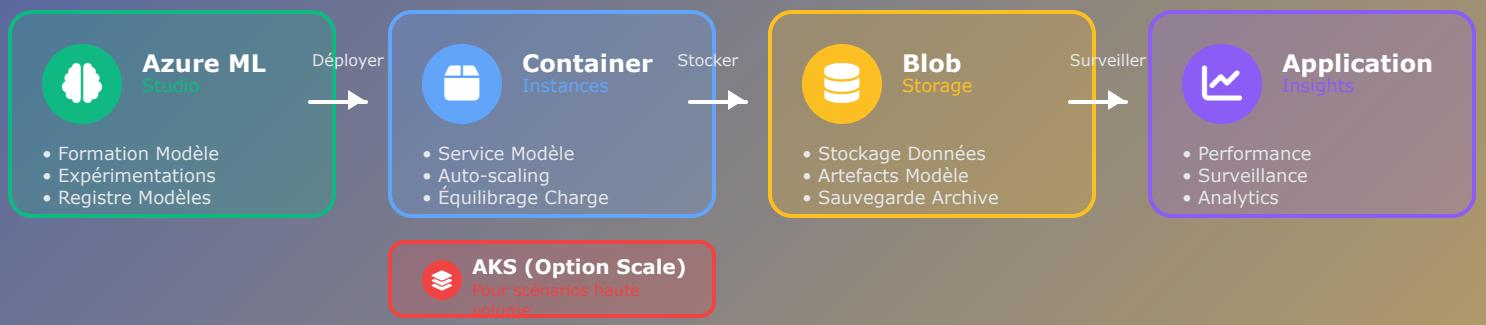
Grafana · Elasticsearch · PagerDuty · Slack

') repeat;">

Déploiement Azure & Structure des Coûts

Architecture Cloud & Coûts

Architecture Azure de Production



Répartition Coûts Mensuels

Scénario Léger (~49€/mois)

Container Instances	Blob Storage	App Insights	Azure ML
25€	12€	7€	5€
1 vCPU, 2GB RAM	100GB stockage	Suivi basique	Niveau base

Scénario Moyen (~180€/mois)

Cluster AKS	Blob Storage	App Insights	Azure ML
120€	25€	20€	15€
Cluster 3 nœuds	500GB stockage	Logs enrichis	Niveau standard

Stratégie de Montée en Charge & Optimisation

Stratégies d'Optimisation des Coûts

Auto-scaling
Réduction automatique pendant faible usage. Azure Functions pour charges sporadiques

Instances Réservées
Économisez jusqu'à 72% avec engagements 1-3 ans pour charges prévisibles

Infrastructure Prête pour la Production

Sécurité & Conformité
- Key Vault pour secrets
- Réseaux virtuels (VNet)
- Points de terminaison privés
- Contrôle d'accès basé rôles
- Chiffrement données au repos

Haute Disponibilité
- Déploiement multi-régions
- Équilibrage de charge
- Basculement automatique
- Vérifications état
- SLA cible 99,9%

DevOps & CI/CD
- Pipelines Azure DevOps
- Tests automatisés
- Déploiements bleu-vert
- Infrastructure as Code
- Workflows GitOps

Surveillance & Alertes
- Tableaux de bord temps réel
- Métriques personnalisées
- Alertes proactives
- Agrégation logs
- Analytics performance

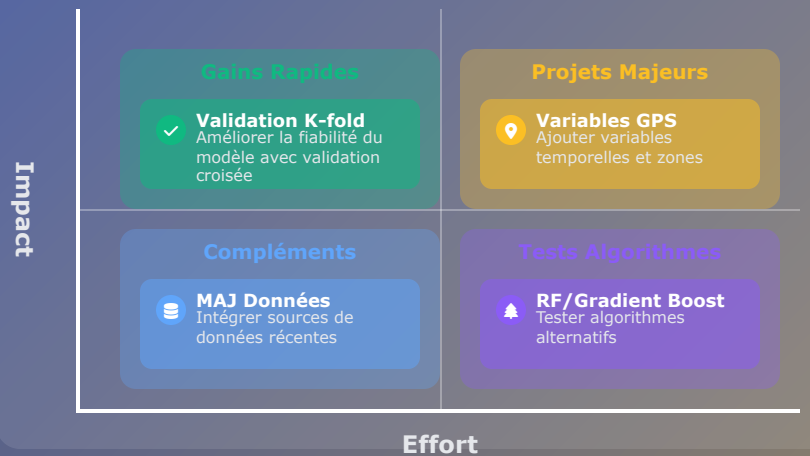
Rentable | Auto-Scalable | Cloud Native | Entreprise Ready

') repeat;">

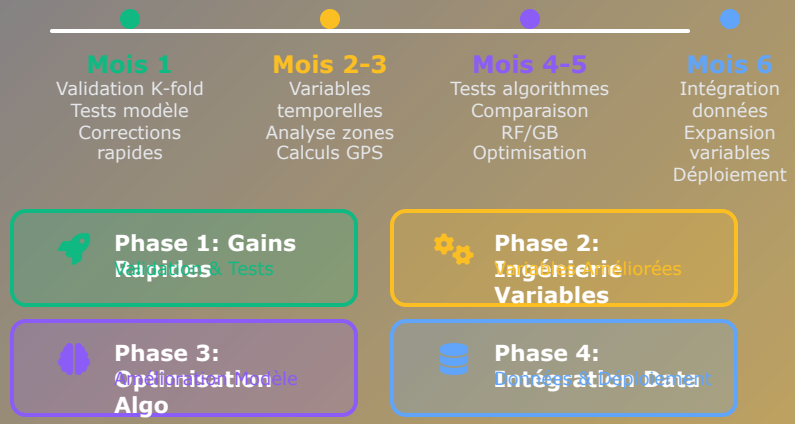
Prochaines Étapes & Améliorations

Feuille de Route Future

Matrice de Priorité des Améliorations



Timeline de la Feuille de Route 6 Mois



Actions Détaillées & Indicateurs de Progression

Améliorations Modèle

Ajouter Variables Temporelles
- Motifs heure/jour de la semaine
- Indicateurs heures de pointe
- Ajustements saisonniers

25%

Calcul Distance GPS
- Implémentation formule Haversine
- Métriques efficacité trajet
- Analyse motifs trafic

10%

Variables par Zones
- Tarification spécifique arrondissement
- Majorations aéroport/monuments
- Indicateurs quartiers d'affaires

0%

Amélioration Algorithmes

Tester Random Forest
- Comparaison avec Régression Linéaire
- Analyse importance variables
- Benchmark performance

75%

Modèles Gradient Boosting
- Implémentation XGBoost
- Optimisation hyperparamètres
- Prévention surapprentissage

40%

Validation Croisée K-fold
- Configuration validation 5-fold
- Évaluation stabilité modèle
- Analyse biais-variance

90%

Stratégie Données

Intégrer Données Récentes
- Registres taxis 2023-2024
- Structures tarifaires mises à jour
- Motifs post-pandémie

20%

Développer Ingénierie Variables
- Intégration données météo
- Ajustements événements
- Facteurs tarification dynamique

15%

Automatisation Pipeline Données
- Ingestion données automatisée
- Calcul variables temps réel
- Surveillance qualité

5%

$R^2 > 0,92$ | < 100ms Réponse | RMSE < 3,00$ | 95% Disponibilité