Prédiction de l'utilisation des vélos Divvy à Chicago

Impact de la météo et des facteurs temporels

Noé Cramette

M1 Data & IA - Ynov Paris
Machine Learning

19 Janvier 2026

Contexte & Problématique

Divvy : Système de vélos partagés de Chicago

- 5,8M de trajets en 2024
- 700+ stations
- Usage très variable selon conditions

Problématique :

- Comment prédire le nombre de trajets par heure ?
- Quels facteurs influencent le plus l'usage ?
- Peut-on anticiper la demande pour optimiser la distribution ?

Données & Sources

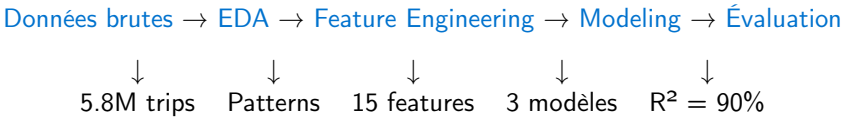
Trois sources de données :

Source	Description	Volume
Divvy Trips	Historique 2024 + 2025	11,3M trajets
Météo	Température, précipitations, vent	730 jours
Calendrier	Evènements US 2024 + 2025	22 jours

Target

Nombre de trajets par heure

Pipeline ML - Vue d'ensemble



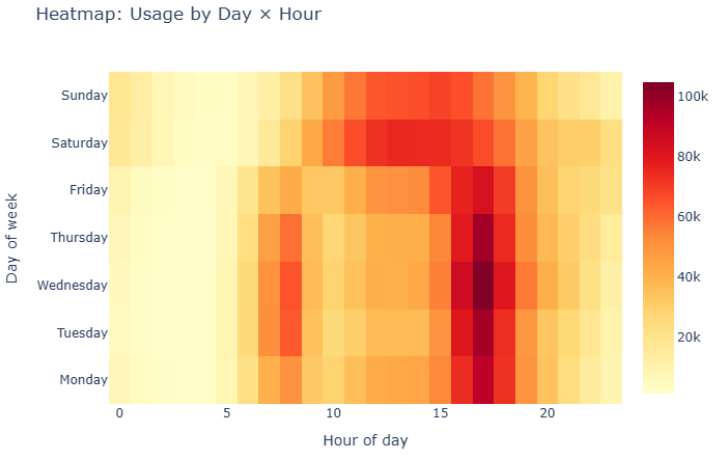
4 étapes clés :

1. **Exploration** : Comprendre les patterns d'usage
2. **Features** : Créer des variables pertinentes
3. **Modélisation** : Tester 3 algorithmes
4. **Validation** : Test sur données 2025

Étape 1 - Exploration des Données

Découvertes clés :

- **Pics horaires** : 8h et 17h
  (trajets domicile-travail)
- **Semaine** ! = **Weekend**
  Patterns différents
- **Saisonnalité forte**
  Été > Hiver
- **Température**
  Corrélation positive forte



Impact des données - Analyse Détaillée

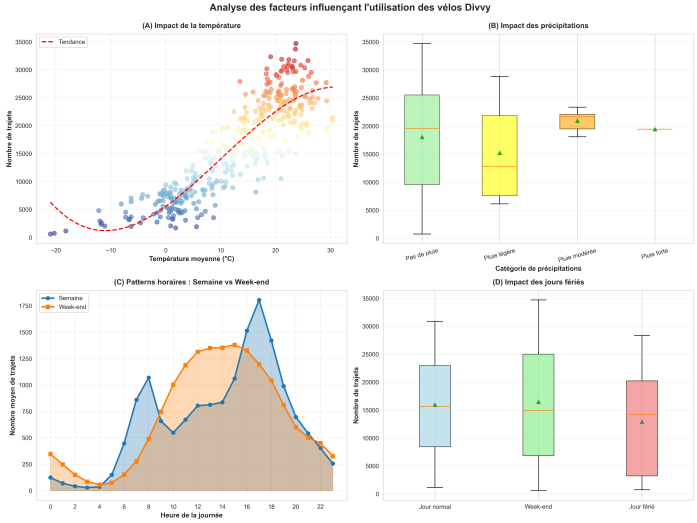
Relations identifiées :

- **Température**
  Relation importante
- **Pluie**
  Impact négatif fort
  -30 à -50% selon intensité
- **Vent**
  Impact modéré

Zone de confort

15-25°C → Usage optimal

Impact Météo - Graphique



Étape 2 - Feature Engineering

15 features créées en 3 catégories

Temporelles (7)

- `hour`
- `day_of_week`
- `month`
- `is_weekend`
- `season_*`
  (one-hot encoded)

Météo (3)

- `temperature`
- `precipitation`
- `wind_speed`

Calendrier (1)

- `is_holiday`

Agrégation : HORAIRE

8760 observations/an → Prédiction heure par heure

Étape 3 - Modélisation

3 modèles testés :

1. **Linear Regression**
   - Baseline simple
   - Relations linéaires uniquement
2. **Random Forest**
   - Ensemble d'arbres de décision
   - Capture les non-linéarités
3. **XGBoost**
   - Gradient Boosting optimisé
   - Amplification de gradient

Split temporel

Train : 2024
(8760 heures)

Test : 2025
(8758 heures)

Agrégation : HORAIRE

2024 est bisextile ⇒ 366 jours
2025 est une année normale ⇒ 365 jours

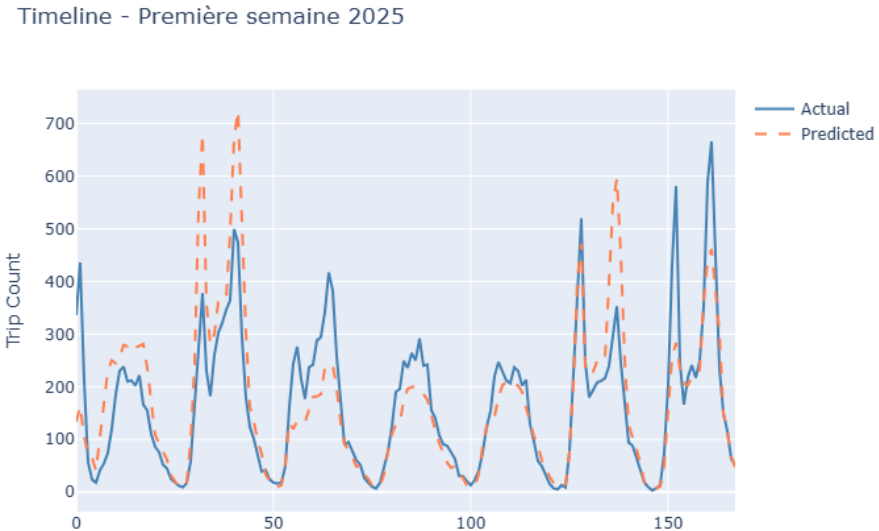
Résultats - Comparaison des Modèles

Performance sur Test 2025

Modèle	R^2	RMSE	MAE	MAPE
Linear Regression	38.6%	507	383	313%
Random Forest	**89.9%**	**205**	**122**	**34.5%**
XGBoost	**89.6%**	**209**	**124**	**36.2%**

Résultat clé : $\sim$90% de variance expliquée !

Visualisation des Prédictions



Visualisation des Prédictions

Prédictions vs Réalité (Random Forest)

Observations :

- Pics correctement prédits
- Tendances saisonnières capturées
- Légères sous-estimations sur événements exceptionnels

Réalité
Prédictions

Feature Importance - Quels facteurs comptent ?

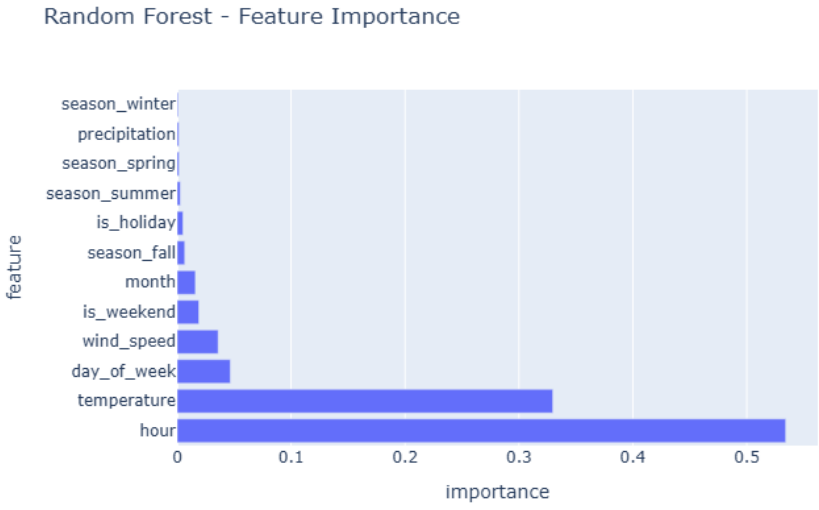
Top 5 des features (Random Forest) :

1. **hour** (53%)
   L'heure de la journée domine
2. **temperature** (32%)
   Facteur météo #1
3. **day_of_week** (4%)
   Cycles hebdomadaires
4. **wind speed** (3%)
5. **precipitation** (0.1%)
   Impact pluie

Insight

Temporel > Météo,
mais météo reste crucial

Feature Importance

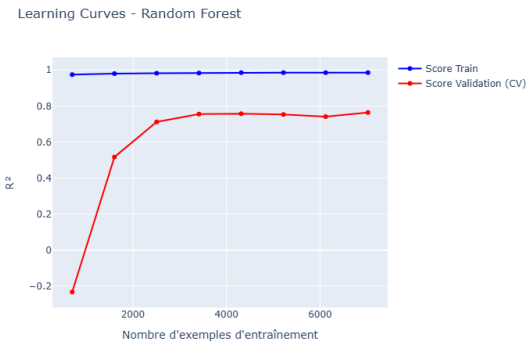


Analyse Critique - Limites

Overfitting détecté

- Train R^2 : 98.4% (RF) et 99.7% (XGB)
- Test R^2 : 89.9% (RF) et 89.6% (XGB)
- Gap : ∼8-10%

→ Modèles apprennent du bruit



Autres limites identifiées :

- Pas d'information sur pannes/maintenance
- Événements exceptionnels sous-prédits

Conclusion

Ce qui a été fait :

Pipeline ML complet

- EDA → Features → Modeling → Evaluation

Prédiction fiable : 90% de précision

- Modèles très satisfaisant

Insights actionnables :

- Température = levier #1 météo
- Patterns temporels primordiaux
- Pluie = impact négatif majeur

Pipeline ML → 5.8M données → 90% précision

Limites & Perspectives

Limites identifiées :

Données

- Météo **journalière** (pas horaire)
  → Perte de précision intra-journée
- Événements incomplets
  → Concerts, festivals, évènements sportifs manquants
- Pas d'info maintenances/pannes
  → Baisse usage inexpliquée

Modèle

- Overfitting détecté (gap 8-10%)
- Pas de features temporelles avancées
- Granularité globale uniquement

Pistes d'amélioration :

Données enrichies

- Météo **horaire**
- Calendrier événements complet
- Logs maintenances système

Features avancées

- Moyennes mobiles (tendances)
- Interactions heure $\times$ météo

Analyse spatiale

- **Prédiction par station**
- Clustering stations similaires

Merci pour votre attention !

Noé Cramette
M1 Data & IA - Ynov Paris