Anticipation du nombre de passagers dans le métro New-Yorkais

AUMAGY Léa
BOUKHALFA Abdeldjallil
BOUCHNEB Hedi

M2 EBDS

January, 2024

amse
école d'économie d'aix-marseille
aix-marseille school of economics

Sommaire

1. Introduction
2. Matériels et méthodes
   - La base de données Subway
   - Méthodologie
3. Analyse exploratoire des variables
   - Tendance de la variable explicative 'Ridership'
   - Effet temporels et météorologique sur la variable 'Ridership'
4. Analyse des prédictions
   - Méthodes paramétriques
   - Méthodes non paramétriques basées sur des arbres de décisions
   - Analyse comparative des méthodes de prédiction
   - Prédiction sur le long terme
5. Conclusion

Introduction

- La mobilité urbaine, un enjeu majeur de ces dernières années.
- Hausse de l'utilisation des transports en communs.
- NYC, une ville dynamique, des milliers de passagers chaque jour par stations.

La base de données Subway

1. Introduction
2. Matériels et méthodes
   - La base de données Subway
   - Méthodologie
3. Analyse exploratoire des variables
   - Tendance de la variable explicative 'Ridership'
   - Effet temporels et météorologique sur la variable 'Ridership'
4. Analyse des prédictions
   - Méthodes paramétriques
   - Méthodes non paramétriques basées sur des arbres de décisions
   - Analyse comparative des méthodes de prédiction
   - Prédiction sur le long terme
5. Conclusion

La base de données Subway

Sources :

- `https://new.mta.info/open-data`
- `https://weatherdownloader.oikolab.com/`

Informations sur la base de données :

- du 01/02/2022 au 31/01/2023
- 4034110 observations
- 6 variables temporelles
- 5 variables de la MTA
- 8 variables météorologique

```
RangeIndex: 4034110 entries, 0 to 4034109
Data columns (total 24 columns):
 #   Column                        Dtype
---  ------                        -----
 0   Date                          datetime64[ns]
 1   Hour                          int64
 2   Ridership                     int64
 3   Station_complex_id            int64
 4   Station_complex               object
 5   Booth                         object
 6   Line                          object
 7   Borough                       object
 8   Structure                     object
 9   Latitude                      float64
 10  Longitude                     float64
 11  Temperature_Celsius           float64
 12  Dewpoint_Temperature_Celsius  float64
 13  Humidex_Index_Celsius         float64
 14  Wind_Speed_ms                 float64
 15  Surface_Solar_Radiation_Wm2   float64
 16  Total_Cloud_Cover             float64
 17  Total_Precipitation_mm        float64
 18  Snowfall_mm                   float64
 19  Month                         int64
 20  Day                           int64
 21  Day_Of_Week                   object
 22  Season                        object
 23  Holidays                      object
```

Méthodologie

1. Introduction
2. Matériels et méthodes
   - La base de données Subway
   - Méthodologie
3. Analyse exploratoire des variables
   - Tendance de la variable explicative 'Ridership'
   - Effet temporels et météorologique sur la variable 'Ridership'
4. Analyse des prédictions
   - Méthodes paramétriques
   - Méthodes non paramétriques basées sur des arbres de décisions
   - Analyse comparative des méthodes de prédiction
   - Prédiction sur le long terme
5. Conclusion

Méthodologie

- Analyse exploratoire des données
- Modèles paramétriques
  - Régression Linéaire
  - Lasso avec l'interprétabilité SHAP
- Modèles non-paramétriques basés sur des arbres de décision
  - Arbre de décision
  - Fôret aléatoire
  - XGBoost avec GridSearchCV
- Métrique pour l'évaluation des modèles
  - Coefficient de détermination : R^2
  - Calculs des MSE
  - Analyse de l'ajustement des modèles

Tendance de la variable explicative 'Ridership'

1. Introduction
2. Matériels et méthodes
   - La base de données Subway
   - Méthodologie
3. Analyse exploratoire des variables
   - Tendance de la variable explicative 'Ridership'
   - Effet temporels et météorologique sur la variable 'Ridership'
4. Analyse des prédictions
   - Méthodes paramétriques
   - Méthodes non paramétriques basées sur des arbres de décisions
   - Analyse comparative des méthodes de prédiction
   - Prédiction sur le long terme
5. Conclusion

Tendance de la variable explicative 'Ridership'

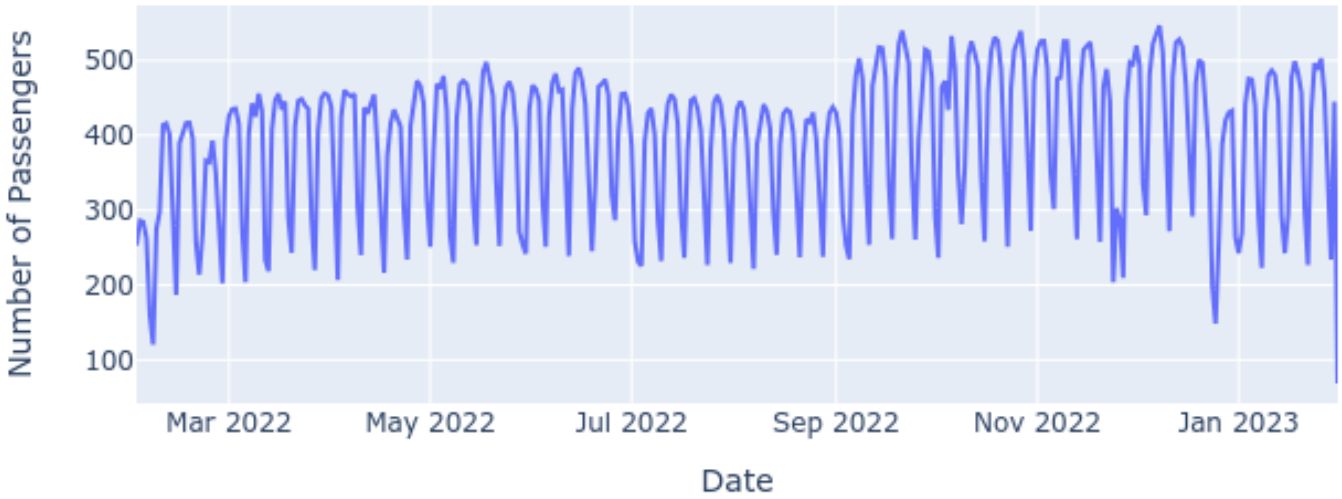


Figure: Évolution du nombre de passager dans le métro par jour

Effet temporels et météorologique sur la variable 'Ridership'

1. Introduction
2. Matériels et méthodes
   - La base de données Subway
   - Méthodologie
3. Analyse exploratoire des variables
   - Tendance de la variable explicative 'Ridership'
   - Effet temporels et météorologique sur la variable 'Ridership'
4. Analyse des prédictions
   - Méthodes paramétriques
   - Méthodes non paramétriques basées sur des arbres de décisions
   - Analyse comparative des méthodes de prédiction
   - Prédiction sur le long terme
5. Conclusion

Analyse exploratoire des variables

Figure: Moyenne du nombre de passagers selon le jour de la semaine

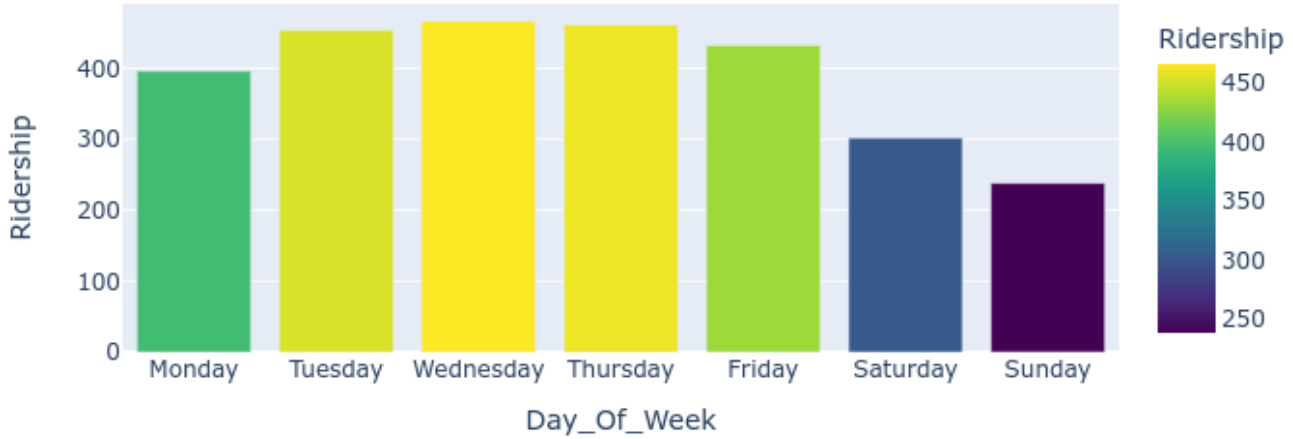
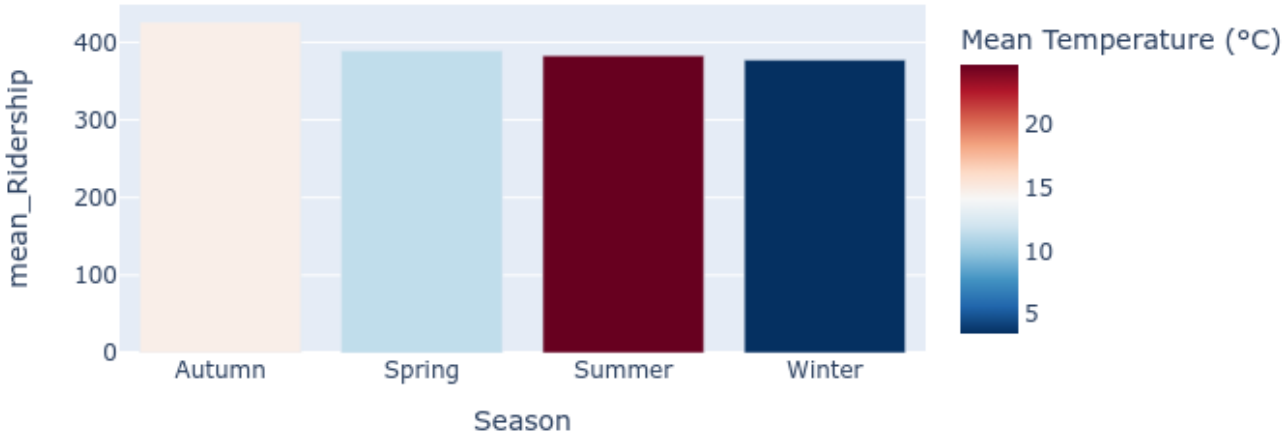


Figure: Moyenne du nombre de passagers par saison et selon la température



Méthodes paramétriques

1. Introduction
2. Matériels et méthodes
   - La base de données Subway
   - Méthodologie
3. Analyse exploratoire des variables
   - Tendance de la variable explicative 'Ridership'
   - Effet temporels et météorologique sur la variable 'Ridership'
4. Analyse des prédictions
   - Méthodes paramétriques
   - Méthodes non paramétriques basées sur des arbres de décisions
   - Analyse comparative des méthodes de prédiction
   - Prédiction sur le long terme
5. Conclusion

Modèle de régression linéaire : Matrice de corrélation

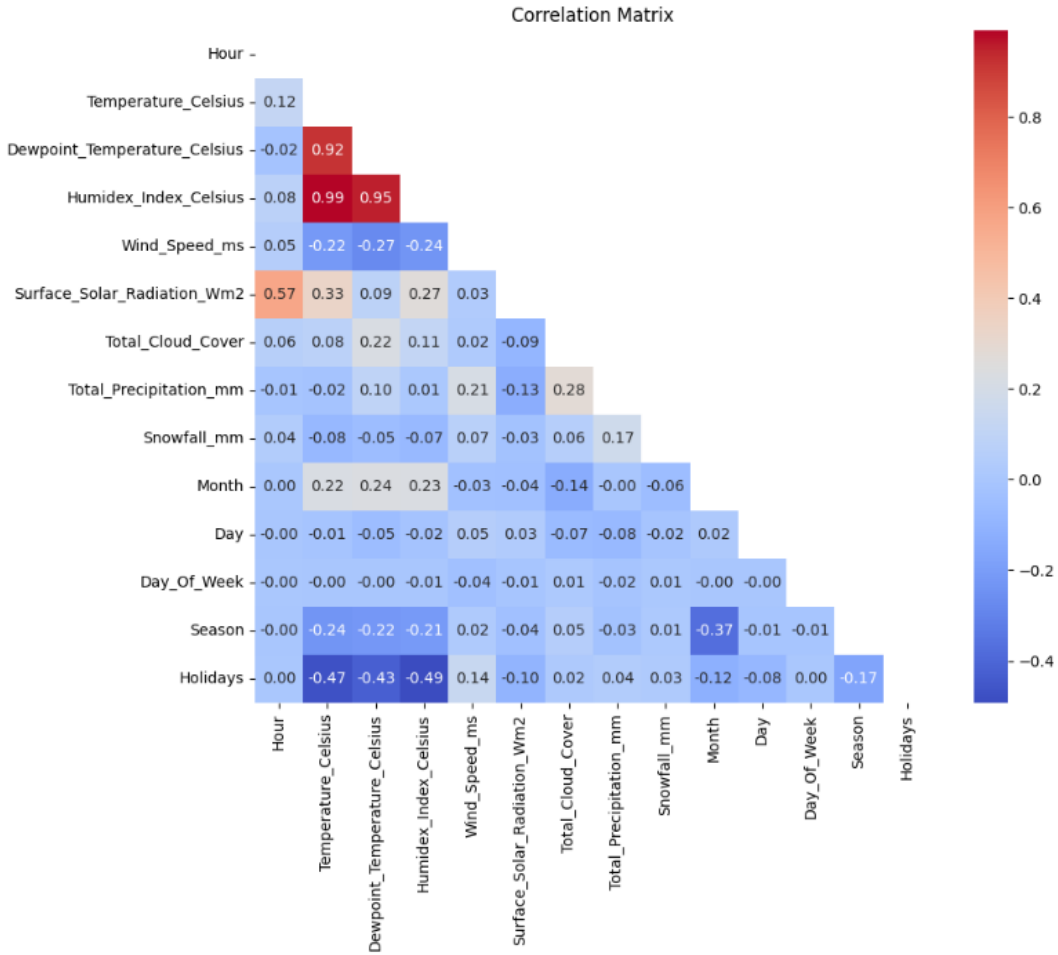


Figure: Matrice de corrélation

Modèle de régression linéaire : Test ANOVA

Table: Anova test

Variables	Df	Sum Sq	F	PR(>F)	
Line	1	2.032544×10^{12}	27922.7644	< 2e-16	***
Temperature_Celsius	1	6.997178×10^{10}	961.2611	< 2e-16	***
Wind_Speed_ms	1	2.861850×10^{10}	393.1564	< 2e-16	***
Surface_Solar_Radiation_Wm2	1	6.573409×10^{11}	9030.4449	< 2e-16	***
Total_Cloud_Cover	1	2.217556×10^{10}	304.6443	< 2e-16	***
Total_Precipitation_mm	1	1.005597×10^{10}	138.1474	< 2e-16	***
Snowfall_mm	1	8.425009×10^{6}	0.1157	0.7337	
Month	1	6.214168×10^{10}	853.6925	< 2e-16	***
Day	1	4.899301×10^{5}	0.0067	0.9346	
Day_Of_Week	1	4.346407×10^{10}	597.1025	< 2e-16	***
Hour	1	1.088130×10^{11}	1494.8560	< 2e-16	***
Season	1	1.453564×10^{9}	19.9688	< 2e-16	***
Holidays	1	5.943270×10^{9}	81.6477	< 2e-16	***
Residual	288255	2.098255×10^{13}			*

Modèle de régression linéaire : Résultats

Table: Résultats de la régression linéaire

Variable	Coefficient	Std Error	t-value	P> |t|	
Line	-248.1484	1.936	-128.180	< 2e-16	***
Temperature_Celsius	8.7950	2.370	3.711	< 2e-16	***
Wind_Speed_ms	53.3168	10.774	4.949	< 2e-16	***
Surface_Solar_Radiation_Wm2	7.4786	0.102	73.544	< 2e-16	***
Total_Cloud_Cover	1166.4839	49.654	23.492	< 2e-16	***
Total_Precipitation_mm	248.1586	46.435	5.344	< 2e-16	***
Month	315.1912	5.561	56.679	< 2e-16	***
Day_Of_Week	315.7359	9.230	34.207	< 2e-16	***
Hour	146.0407	3.357	43.502	< 2e-16	***
Season	576.4821	16.602	34.724	< 2e-16	***
Holidays	2050.4478	42.937	47.755	< 2e-16	***
R-squared		0.364			
Adjusted R-squared		0.364			
F-statistic		1.052e+04			
Prob (F-statistic)		0.000			
No. Observations		201788			

Table: Résultats des métriques

Metric	Value
R-squared (R2)	0.1399
Mean Squared Error (Training Set)	73,427,412.26
Mean Squared Error (Validation Set)	75,162,060.17
Difference in MSE	1,734,647.91

Méthode de réduction de l'information : Lasso

Figure: Valeurs SHAP

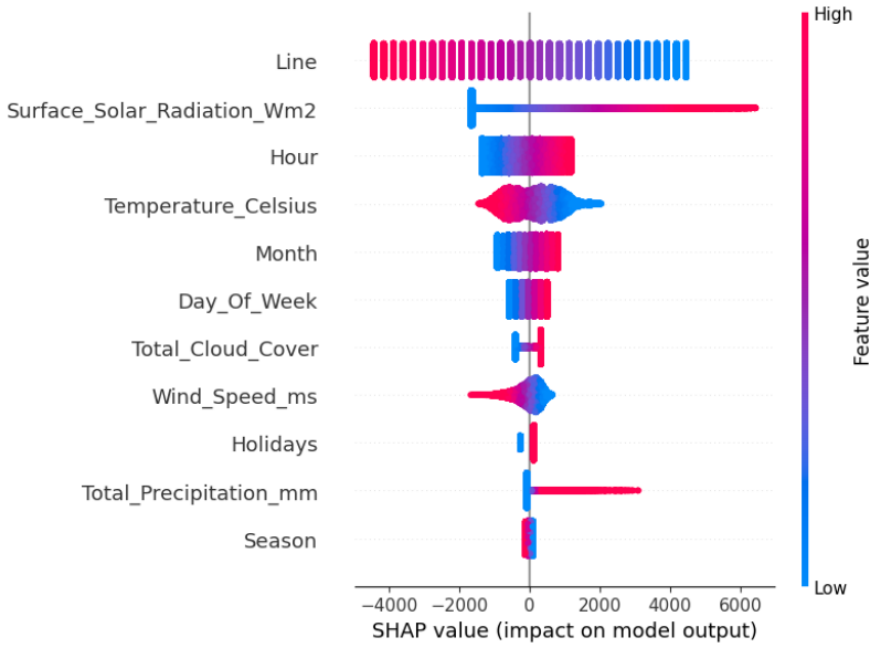


Table: Résultats des métriques

Metric	Value
R-squared (R2)	0.1531
Mean Squared Error (Training Set)	72,265,076.42
Mean Squared Error (Validation Set)	74,014,456.83
Difference in MSE	1,749,380.41

Méthodes non paramétriques basées sur des arbres de décisions

1. Introduction
2. Matériels et méthodes
   - La base de données Subway
   - Méthodologie
3. Analyse exploratoire des variables
   - Tendance de la variable explicative 'Ridership'
   - Effet temporels et météorologique sur la variable 'Ridership'
4. Analyse des prédictions
   - Méthodes paramétriques
   - Méthodes non paramétriques basées sur des arbres de décisions
   - Analyse comparative des méthodes de prédiction
   - Prédiction sur le long terme
5. Conclusion

Arbre de décision

Figure: Decision Tree

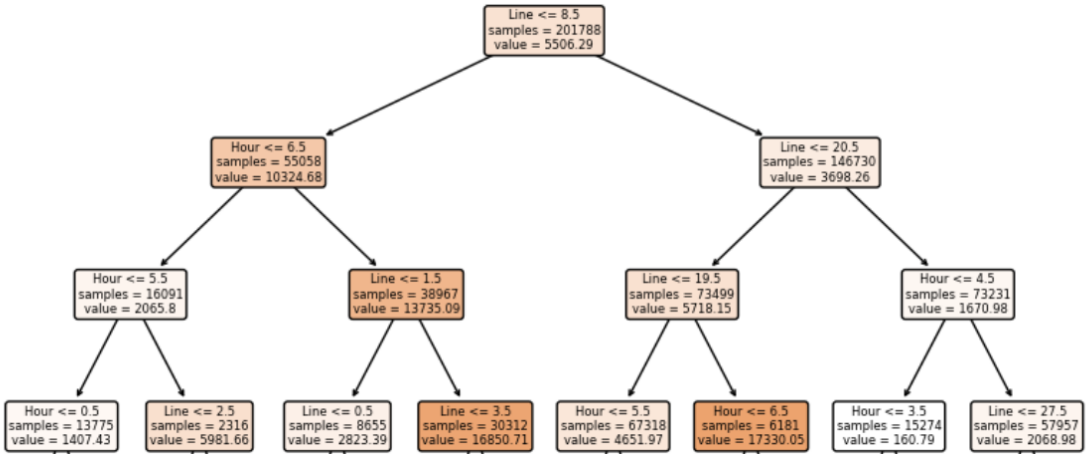


Table: Résultats des métriques

Metric	**Value**
R-squared (R2)	0.448
Mean Squared Error on training set	46836557.0131
Mean Squared Error on validation set	48227885.596
The difference between the two MSE	1391328.583

Forêts aléatoires

Figure: Variables importantes

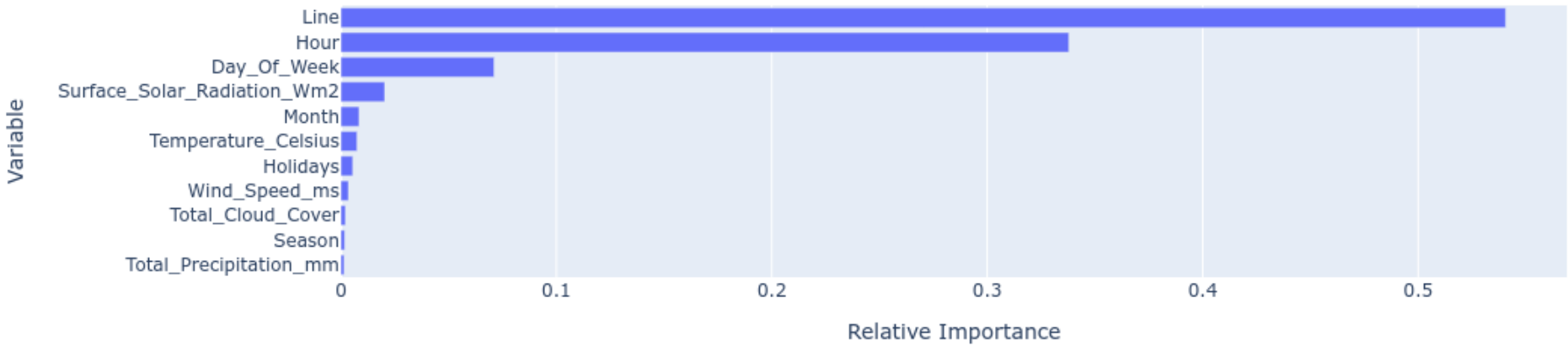


Table: Résultats des métriques

Metric	Value
R-squared (R2)	0.982
Mean Squared Error (Training Set)	289476.611
Mean Squared Error (Validation Set)	1582498.791
Difference in MSE	1293022.180

Extreme Gradient Boosting (XGBoost)

Choix des hyperparamètres selon GridSearchCV :

- Taux d'apprentissage : 0.2
- Nombre d'arbres : 300
- Profondeur des arbres : 6

Table: Résultats des métriques

Métriques	Valeurs
R-squared (R2)	0.979
Mean Squared Error (Training Set)	1469527.953
Mean Squared Error (Validation Set)	1833577.636
Difference in MSE	364049.683

Analyse comparative des méthodes de prédiction

1. Introduction
2. Matériels et méthodes
   - La base de données Subway
   - Méthodologie
3. Analyse exploratoire des variables
   - Tendance de la variable explicative 'Ridership'
   - Effet temporels et météorologique sur la variable 'Ridership'
4. Analyse des prédictions
   - Méthodes paramétriques
   - Méthodes non paramétriques basées sur des arbres de décisions
   - Analyse comparative des méthodes de prédiction
   - Prédiction sur le long terme
5. Conclusion

Analyse des métriques d'évaluation

Table: Comparaison des métriques

Modèles	**R^2**	**MSE_train**	**MSE_test**	**Ajustement**
Régression Linéaire	0.14	73427412.26	75162060.17	1734647.91
Lasso	0.15	72265076.42	74014456.83	1749380.41
Arbre de décision	0.45	46836557.01	48227885.60	1391328.58
Forêts aléatoires	0.98	290870.44	1627931.11	1337060.67
XGBoost	0.98	1469527.95	1833577.64	364049.68

Analyse des métriques d'évaluation

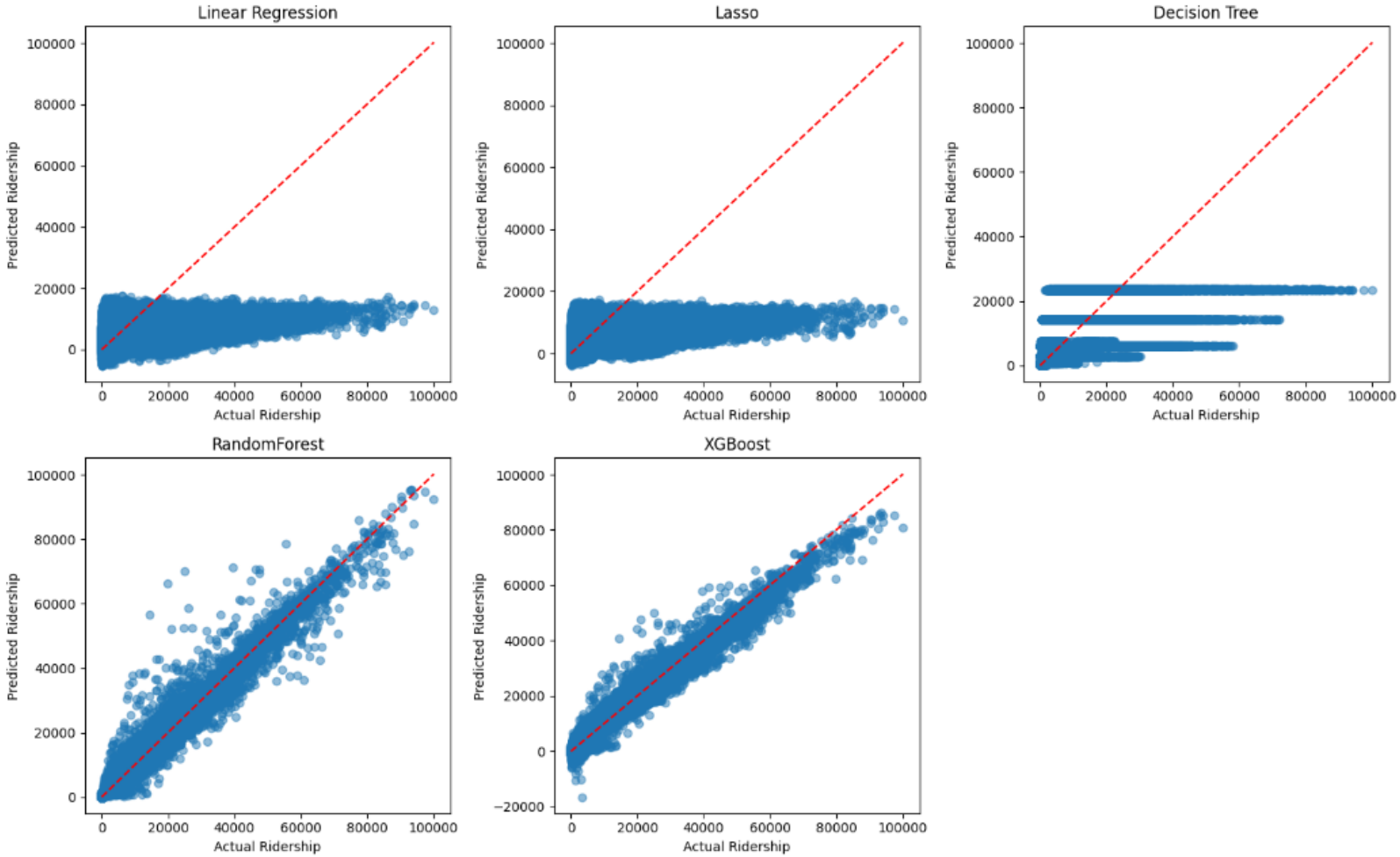
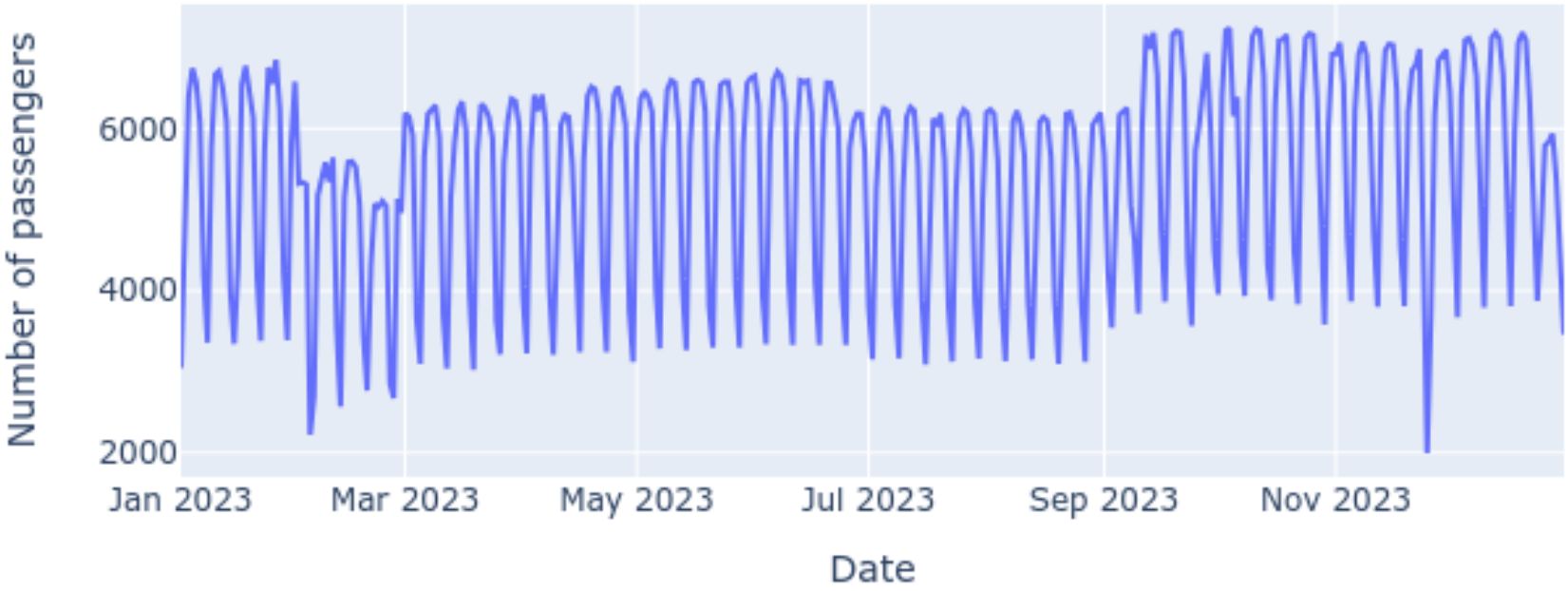


Figure: Valeurs réelles vs. prédites

Prédiction sur le long terme

1. Introduction
2. Matériels et méthodes
   - La base de données Subway
   - Méthodologie
3. Analyse exploratoire des variables
   - Tendance de la variable explicative 'Ridership'
   - Effet temporels et météorologique sur la variable 'Ridership'
4. Analyse des prédictions
   - Méthodes paramétriques
   - Méthodes non paramétriques basées sur des arbres de décisions
   - Analyse comparative des méthodes de prédiction
   - Prédiction sur le long terme
5. Conclusion

Prédiction sur l'année 2023



Conclusion