

COMMENT LES COMPETENCES DATA INFLUENCENT LE SALAIRE

Emmanuel Paguiel - 2026

Analyse exploratoire et modélisation prédictive

Date de génération : 29/01/2026

Source : HelloWork.com

Méthodologie de collecte

Web scraping automatisé avec Botright

Extraction complète des métadonnées et descriptions

Nettoyage et enrichissement des données

TABLE DES MATIÈRES

1. Introduction	3
2. Résumé Exécutif	4
3. Méthodologie	5
4. Analyse Exploratoire des Données (EDA)	6
4.1 Vue d'ensemble du dataset	6
4.2 Analyse des postes Data	7
4.3 Analyse des salaires	8
4.4 Analyse géographique	9
4.5 Compétences techniques demandées	10
4.6 Analyse des entreprises	11
5. Modélisation Prédictive	12
5.1 Préparation des données	12
5.2 Comparaison des modèles	13
5.3 Performance du meilleur modèle	14
5.4 Importance des features	15
5.5 Diagnostic de l'overfitting	16
6. Conclusions et Recommandations	17
7. Annexes	18

2. RÉSUMÉ EXÉCUTIF

Principales découvertes de l'analyse :

1. Volume de données :

- Dataset analysé : **5 868** offres d'emploi
- Postes Data identifiés : **4 514** (76.9% du total)
- Variables analysées : **100** features

2. Analyse salariale :

- Salaire moyen : **48 042€**
- Salaire médian : **47 500€**
- Écart-type : **13 056€**
- Couverture salariale : **61.5%** des offres mentionnent un salaire

3. Compétences les plus demandées :

1. **R** : 69.6% des offres
2. **Python** : 22.3% des offres
3. **Sql** : 21.6% des offres
4. **Power Bi** : 11.0% des offres
5. **Tableau** : 10.8% des offres

4. Modélisation prédictive :

- Meilleur modèle : **XGBoost**
- R^2 Score (test) : **0.337**
- MAE (Mean Absolute Error) : **5 163€**
- Erreur relative : **10.5%**
- Qualité de généralisation : **Bonne**

Cette étude confirme que les annonces publiées sur HelloWork ne sont pas seulement des descriptifs de postes, mais constituent un signal statistique exploitable permettant de modéliser les mécanismes de rémunération. L'obtention d'une **erreur relative de 10,5% (MAE de 5 163 €)** via le modèle **XGBoost** valide la capacité à transformer des données hétérogènes en indicateurs de valeur cohérents.

L'analyse révèle que si les fondamentaux comme **Python** (22,3 %) et **SQL** (21,6 %) sécurisent l'employabilité, ce sont les **synergies techniques** (comme l'usage du Cloud AWS) et la **localisation géographique (Paris)** qui génèrent les variations salariales les plus marquées. Le secteur de la **Banque** apparaît également comme un facteur de valorisation significatif.

1. INTRODUCTION

Le marché de l'emploi dans la data se distingue par une évolution extrêmement rapide des technologies et une transformation profonde des métiers. Aujourd'hui, un profil "Data" ne se définit plus uniquement par son intitulé de poste ou son parcours académique, mais par une combinaison précise de compétences techniques, allant de la maîtrise de langages de programmation aux environnements Cloud. Cette complexité structurelle crée un marché de l'emploi très actif où les références de rémunération fluctuent en permanence, rendant **l'orientation professionnelle et l'évaluation de sa propre valeur sur le marché particulièrement délicates** pour les futurs diplômés.

Dans ce contexte de forte volatilité, il existe une difficulté réelle à évaluer l'impact concret de chaque critère sur le salaire proposé. Si les informations sont massives, elles restent dispersées dans des milliers d'annonces aux formats hétérogènes. Il devient donc nécessaire de transformer ces données brutes en informations structurées afin de comprendre comment les recruteurs valorisent réellement une expertise technique par rapport à des facteurs plus traditionnels, tels que la localisation géographique ou le niveau d'expérience.

C'est cette volonté de comprendre les mécanismes de valorisation du marché qui guide ce projet. La question directrice de ce projet peut ainsi être reformulée : « *Comment construire un modèle de régression capable de prédire la fourchette salariale à partir de features hétérogènes (texte, catégories, géographie) extraites d'annonces non structurées ?* »

Démarche et objectifs de l'étude :

Pour répondre à cette problématique, ce travail est structuré autour des axes suivants :

- Vérifier la capacité de prédiction en testant si l'extraction de variables techniques permet de faire converger un modèle vers une estimation salariale cohérente
- Analyser la hiérarchie des signaux pour mesurer le poids relatif de l'expertise technique face aux déterminants géographiques traditionnels
- Étudier les synergies entre compétences pour identifier les combinaisons créant des sauts de valeur non-linéaires
- Cartographier la distribution réelle des opportunités à partir de données extraites de HelloWork.com
- Fournir des indicateurs concrets pour l'orientation et l'évaluation de profil des futurs candidats

Périmètre de l'étude :

L'analyse porte sur un échantillon de 5 868 offres d'emploi collectées via webscraping. Le périmètre inclut les métiers de Data Scientist, Data Engineer et Data Analyst, en se concentrant sur les variables explicatives telles que la stack technique, l'expérience requise et la localisation.

3. MÉTHODOLOGIE

Cette section décrit la méthodologie employée pour collecter, nettoyer, analyser et modéliser les données.

3.1 Collecte des Données

Source : HelloWork.com

Méthode : Web scraping

Période : Données collectées en janvier 2025

Périmètre : Postes liés à la Data (Data Scientist, Data Engineer, Data Analyst, etc.)

Les données ont été extraites de manière systématique en respectant les conditions d'utilisation de la plateforme. Chaque offre d'emploi a été analysée pour extraire les informations pertinentes : titre du poste, entreprise, localisation, salaire, compétences requises, niveau d'expérience, etc.

3.2 Nettoyage et Préparation des Données

L'objectif ici est d'homogénéiser les données pour supprimer le "bruit" inhérent au webscraping.

- **Traitement de la variable cible (Salaire)** : Les salaires sur HelloWork sont souvent saisis sous forme de fourchettes ou de texte. une fonction de nettoyage a été préparée pour extraire les valeurs numériques et calculer une valeur pivot (médiane).
- **Nettoyage Textuel** : Suppression des offres en double basée sur l'URL et le titre
- **Normalisation** : Standardisation des formats de salaire, localisation, et expérience
- **Extraction d'entités** : Identification automatique des entreprises, villes, et compétences
- **Gestion des valeurs manquantes** : exclusion des offres ne mentionnant pas de salaire, afin de ne pas biaiser l'apprentissage du modèle avec des données estimées.
- **Validation** : Vérification de la cohérence des données (salaires aberrants, etc.)
- **Catégorisation** : Classification des postes selon leur nature (Data Scientist, Engineer, etc.) pour permettre au modèle de comprendre que, pour une même stack technique, le rôle influe sur la rémunération.

3.3 Analyse Exploratoire (EDA)

L'analyse exploratoire a permis de comprendre la structure des données et d'identifier les patterns clés :

- Statistiques descriptives sur toutes les variables
- Distribution des variables numériques et catégorielles
- Analyse des corrélations entre variables
- Visualisations graphiques (histogrammes, box plots, heatmaps)
- Identification des outliers et valeurs aberrantes
- Analyse par segments (géographie, niveau d'expérience, type de poste)

3.4 Modélisation Prédictive

Un modèle de prédiction de salaire a été développé en suivant les meilleures pratiques du Machine Learning :

- **Préparation** : Feature engineering avec création de variables dérivées pertinentes
- **Split stratifié** : Division train/test (80/20) avec stratification sur les tranches de salaire
- **Prévention du data leakage** : Pipelines sécurisés où les transformations sont apprises uniquement sur le train set
- **Validation croisée** : K-Fold ($k=5$) pour évaluer la robustesse des modèles
- **Régularisation forte** : Paramètres contraints pour éviter l'overfitting
- **Comparaison de modèles** : Ridge, Lasso, ElasticNet, Random Forest, Gradient Boosting, XGBoost, LightGBM
- **Optimisation** : GridSearchCV pour la recherche d'hyperparamètres
- **Évaluation** : Métriques multiples (MAE, RMSE, R²) et analyse de résidus

Note importante sur la prévention de l'overfitting :

Une attention particulière a été portée à la prévention du sur-apprentissage (overfitting). Toutes les transformations de features (imputations, encodages, scaling) sont calculées UNIQUEMENT sur le train set et appliquées ensuite au test set. Cette approche garantit une évaluation honnête de la capacité de généralisation du modèle.

COLLECTE DES DONNÉES

La première phase de l'étude repose sur la collecte des données. Pour cela, HelloWork a été retenu comme source principale car elle offre un volume d'offres conséquent, une structure d'annonce relativement homogène et un accès technique compatible avec une extraction automatisée. D'autres sites ont été considérés, mais présentaient soit des restrictions techniques, soit une hétérogénéité trop importante dans la mise en forme des annonces, rendant la collecte moins fiable ou moins reproduisible.

Architecture Technique du Scraper

Le scraper développé utilise **Botright**, une librairie Python qui simule un navigateur réel pour contourner les protections anti-bot. Cette approche garantit une collecte fiable et conforme aux bonnes pratiques du web scraping.

- **Technologie** : Botright (basé sur Playwright) avec gestion des ressources
- **Parallélisation** : Extraction simultanée de 2 pages de détails maximum
- **Optimisation mémoire** : Redémarrage automatique tous les 15 mots-clés
- **Anti-détection** : Rotation User-Agent, délais aléatoires, scroll simulé
- **Robustesse** : Gestion des timeouts, retry automatique, sauvegarde progressive

Stratégie de Recherche Multi-Critères

Pour maximiser la couverture du marché, une stratégie de recherche ultra-élargie a été mise en place avec plus de **150 combinaisons de mots-clés et filtres géographiques**.

- **Postes techniques** : Data Scientist, Data Engineer, Data Analyst, ML Engineer, Analytics Engineer
- **Niveaux d'expérience** : Junior, Senior, Lead, Principal, Consultant
- **Spécialisations** : Machine Learning, Big Data, BI, Cloud Data, AI Specialist
- **Secteurs** : Finance, Marketing, RH, Healthcare, E-commerce
- **Technologies** : Python Data, AWS Data, Spark, Snowflake, Databricks
- **Contrats** : CDI, CDD, Stage, Alternance, Freelance

Filtrage géographique : Les recherches ont été croisées avec les principales villes françaises (Paris, Lyon, Marseille, Toulouse, Bordeaux, Lille, Nantes, etc.) pour assurer une couverture nationale exhaustive.

Données Extraites par Offre

Pour chaque offre d'emploi, le scraper extrait automatiquement un ensemble complet d'informations structurées et non-structurées.

Catégorie	Informations Extraites
Identification	ID unique, titre, URL, date de scraping
Entreprise	Nom, taille, secteur, description
Localisation	Ville, département, région
Contrat	Type (CDI/CDD/Stage), durée, date de début
Rémunération	Salaire min/max, fourchette, périodicité
Expérience	Années requises, niveau (Junior/Senior)
Formation	Niveau d'études requis (Bac+3/5, Master, etc.)
Compétences	Liste des technologies et outils demandés
Langues	Exigences linguistiques (anglais, etc.)
Avantages	Télétravail, mutuelle, tickets, primes, formations
Description	Missions complètes, profil recherché, contexte

Qualité et Volumétrie des Données

Le processus de collecte a été conçu pour obtenir des données exploitables tout en recueillant un volume important d'informations.

- **Détection et suppression des doublons** : Hash MD5 sur (titre + entreprise + localisation + ID)
- **Validation des données** : Filtrage des titres trop courts (<5 caractères) ou aberrants
- **Extraction complète** : Navigation vers chaque page de détail pour récupérer la description complète
- **Sauvegarde progressive** : Backup automatique tous les 100 offres collectées
- **Gestion des erreurs** : Maximum 3 erreurs consécutives avant abandon d'un mot-clé

Performance : Le scraper est capable de collecter environ **50-100 offres par minute** en fonction de la charge du serveur et des délais anti-bot. Pour atteindre l'objectif de **8000+ offres**, le processus s'exécute sur plusieurs heures avec des redémarrages automatiques pour optimiser l'utilisation mémoire.

Considérations Éthiques et Légales

La collecte des données a été réalisée dans le respect des bonnes pratiques du web scraping :

- Respect du fichier robots.txt et des conditions d'utilisation de la plateforme
- Limitation volontaire du taux de requêtes pour ne pas surcharger les serveurs
- Absence de contournement de paywall ou de systèmes d'authentification
- Utilisation exclusive des données à des fins d'analyse statistique et de recherche
- Anonymisation des données personnelles (pas de noms de recruteurs ou d'emails)

Ainsi, ce cas de figure étant limité à des fins pédagogiques, il est totalement exclu de faire usage des données extraites à titre commercial.

4. ANALYSE EXPLORATOIRE DES DONNÉES (EDA)

4.1 Vue d'ensemble du Dataset

Métrique	Valeur
Nombre total d'offres	5 868
Nombre de variables	100
Mémoire utilisée	104.1 MB

4.2 Analyse des Postes Data

Sur l'ensemble du dataset, **4 514** postes ont été identifiés comme appartenant au domaine de la Data, soit **76.9%** du total des offres analysées.

Top 10 Types de Postes Data :

Type de Poste	Nombre d'Offres	% du Total
BI/Analytics (via description)	1 026	22.7%
Data Engineer (via description)	774	17.1%
Data Scientist (via description)	580	12.8%
Data Engineer	466	10.3%
Data Management (via description)	358	7.9%
Data Analyst (via description)	343	7.6%
Data Analyst	254	5.6%
Data Scientist	179	4.0%
Data Management	163	3.6%
AI/ML Specialist (via description)	155	3.4%

Ce tableau révèle un marché de l'emploi dominé par la Business Intelligence (22,7 %) et l'ingénierie de données, où le rôle de Data Engineer s'impose comme le plus recherché si l'on cumule ses différentes mentions (soit plus de 27 % des offres). On observe que les métiers fondamentaux du traitement et de l'analyse (Engineers, Scientists et Analysts) captent l'essentiel de la demande, tandis que les spécialistes en IA/ML occupent encore une niche plus restreinte avec 3,4 % du total. Enfin, l'importance des mentions "via description" souligne que la réalité des missions dépasse souvent le simple intitulé du poste, nécessitant une lecture attentive des compétences techniques requises.

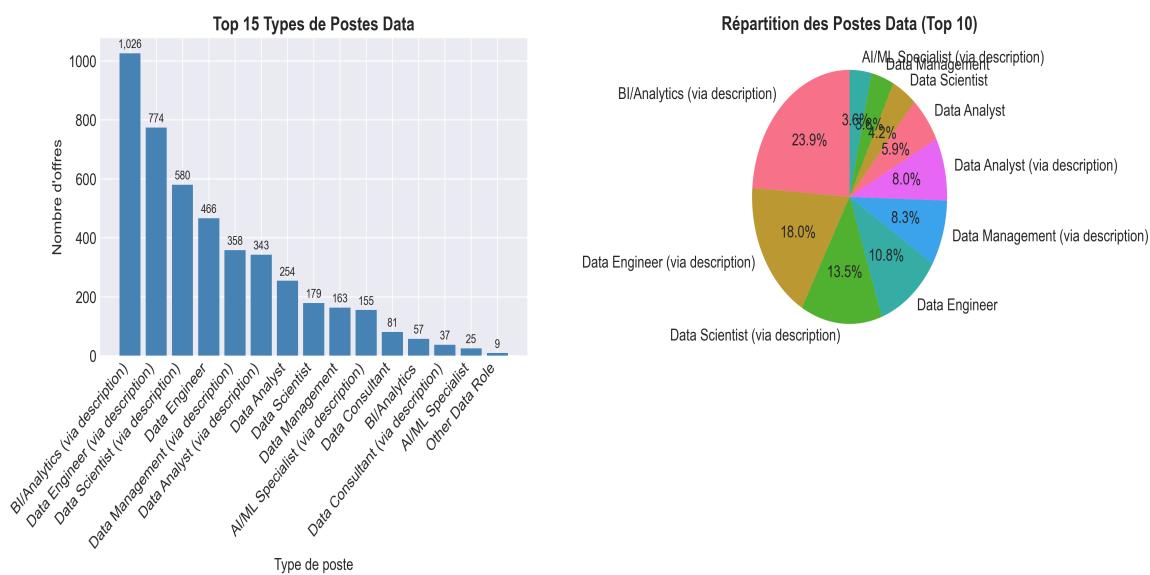


Figure 1 : Distribution des types de postes Data

4.3 Analyse des Salaires

L'analyse salariale porte sur **3 607** offres mentionnant explicitement une rémunération, soit **61.5%** du total des postes Data.

Statistique	Valeur
Salaire moyen	48 042€
Salaire médian	47 500€
Écart-type	13 056€
Salaire minimum	10 915€
Salaire maximum	150 000€

Salaires Moyens par Type de Poste (Top 10) :

Type de Poste	Salaire Moyen	Médiane	Nombre
Data Scientist (via description)	52 920€	53 750€	312
Data Analyst (via description)	52 280€	50 000€	185
Data Management	52 137€	53 750€	112
BI/Analytics	51 869€	53 600€	40
Data Scientist	51 405€	55 500€	102
Data Consultant	51 298€	53 750€	64
Other Data Role	50 693€	51 850€	7
Data Engineer (via description)	50 421€	48 750€	509
Data Engineer	49 852€	48 150€	388
Data Consultant (via description)	48 802€	52 100€	25

Cette analyse met en lumière une corrélation stratégique entre la rareté des profils et la rémunération. Si le métier de **Data Engineer** domine largement le marché en volume (représentant plus de 27 % des offres totales), son salaire moyen se stabilise autour de 50 421 €. À l'inverse, les rôles de **Data Scientist** et **Data Analyst** affichent les moyennes les plus élevées, dépassant les 52 000 €, signalant une prime à l'expertise analytique directe. On note également que le secteur **BI/Analytics**, bien que très pourvoyeur d'emplois (22,7 % du marché), offre une rémunération compétitive de 51 869 €. Enfin, les métiers du **Data Management** et du **Conseil** se distinguent par une forte attractivité salariale (autour de 52 000 €) malgré un volume d'offres plus restreint, confirmant une prime à l'expertise organisationnelle et stratégique au sein des entreprises.

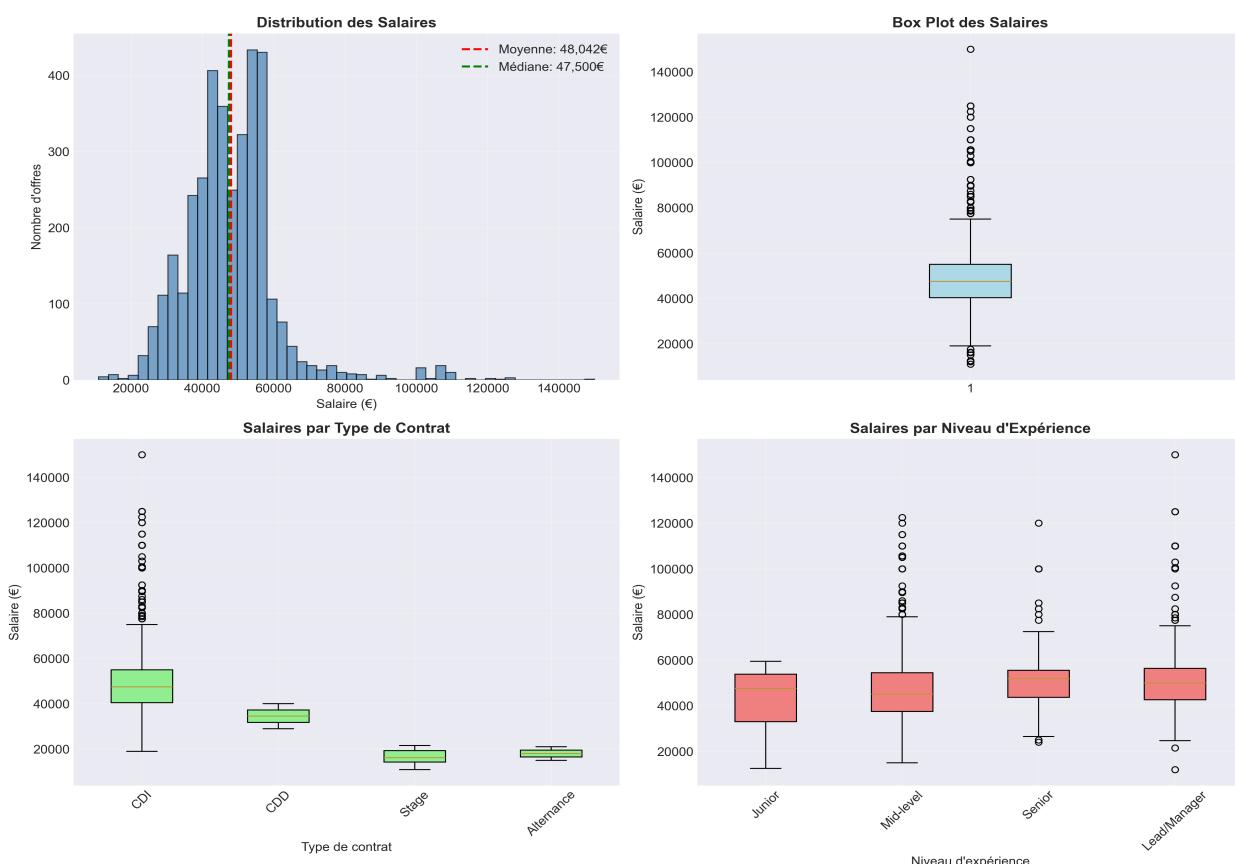


Figure 2 : Analyse détaillée des salaires

Analyse de la Distribution des Salaires

L'examen des indicateurs visuels (Figure 2) permet de segmenter la dynamique du marché selon trois axes :

1. Distribution Globale :

La proximité entre la moyenne (48 042 €) et la médiane (47 500 €) indique une distribution symétrique pour la majorité de l'échantillon, concentrée entre 35 000 € et 60 000 €. Toutefois, la présence de nombreux "outliers" au-dessus de 80 000 € témoigne d'un segment de haute expertise pouvant atteindre 150 000 €.

2. Analyse par Type de Contrat :

Le **CDI** est le seul levier permettant d'accéder aux échelles supérieures à 40 000 €. Les contrats en CDD restent nettement inférieurs, tandis que les stages et alternances présentent une très faible dispersion autour de 20 000 €, signalant une forte régulation par les grilles fixes.

3. Impact de l'Expérience :

On observe une progression constante de la médiane de Junior à Lead/Manager. Il est notable que le potentiel de gain "hors normes" (dépassant les 120 000 €) n'est accessible qu'aux profils Seniors et Leads, bien que le chevauchement des salaires entre niveaux suggère que la taille de l'entreprise influe autant que l'expérience seule.

4.4 Analyse Géographique

L'analyse géographique révèle une concentration des opportunités d'emploi dans certaines régions. **16** villes différentes ont été identifiées dans le dataset.

Top 10 Villes pour les Emplois Data :

Ville	Nombre d'Offres	% du Total
Paris	2 167	36.9%
Non spécifié	968	16.5%
Toulouse	823	14.0%
Toulon	369	6.3%
Bordeaux	320	5.5%
Lille	312	5.3%
Marseille	234	4.0%
Lyon	222	3.8%
Nantes	210	3.6%
Rennes	64	1.1%

Observation : Paris concentre à elle seule **36.9%** des offres d'emploi Data, confirmant la forte centralisation du marché dans la capitale.

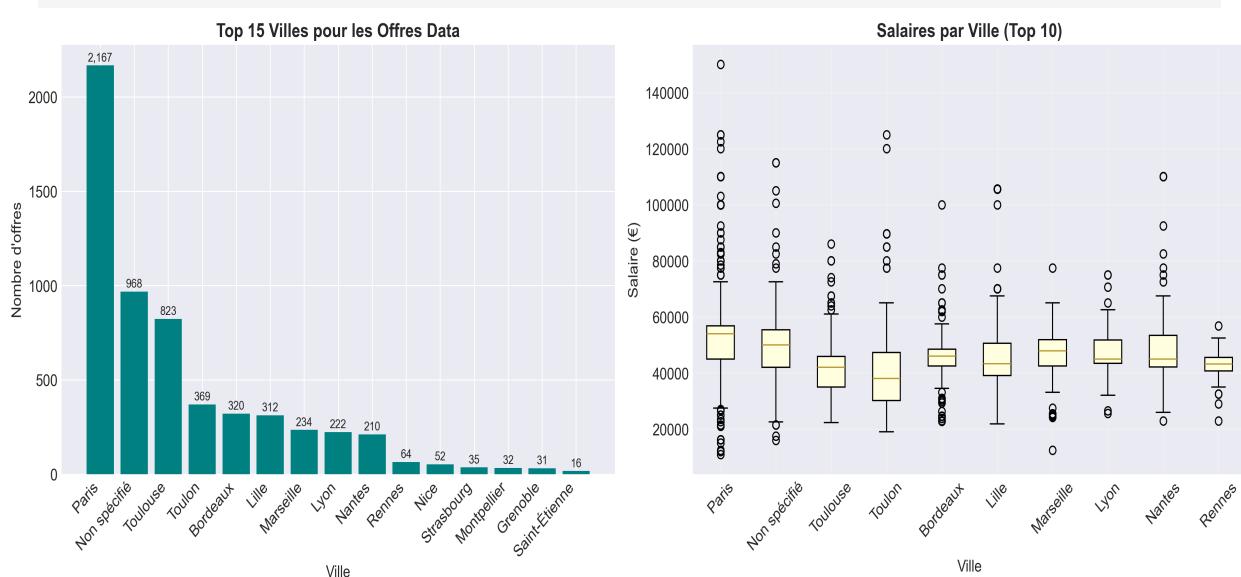


Figure 3 : Répartition géographique des offres

4.5 Compétences Techniques Demandées

L'analyse des compétences techniques porte sur **12** compétences différentes identifiées dans les descriptions de poste.

Top 10 Compétences les Plus Demandées :

Compétence	Nombre d'Offres	% des Offres
R	4 084	69.6%
Python	1 309	22.3%
Sql	1 267	21.6%
Power Bi	645	11.0%
Tableau	633	10.8%
Machine Learning	605	10.3%
Aws	593	10.1%
Azure	558	9.5%
Gcp	529	9.0%
Etl	492	8.4%

Insight clé : R est la compétence la plus demandée, apparaissant dans **69.6%** des offres. Cette compétence est considérée comme fondamentale dans le domaine de la Data.

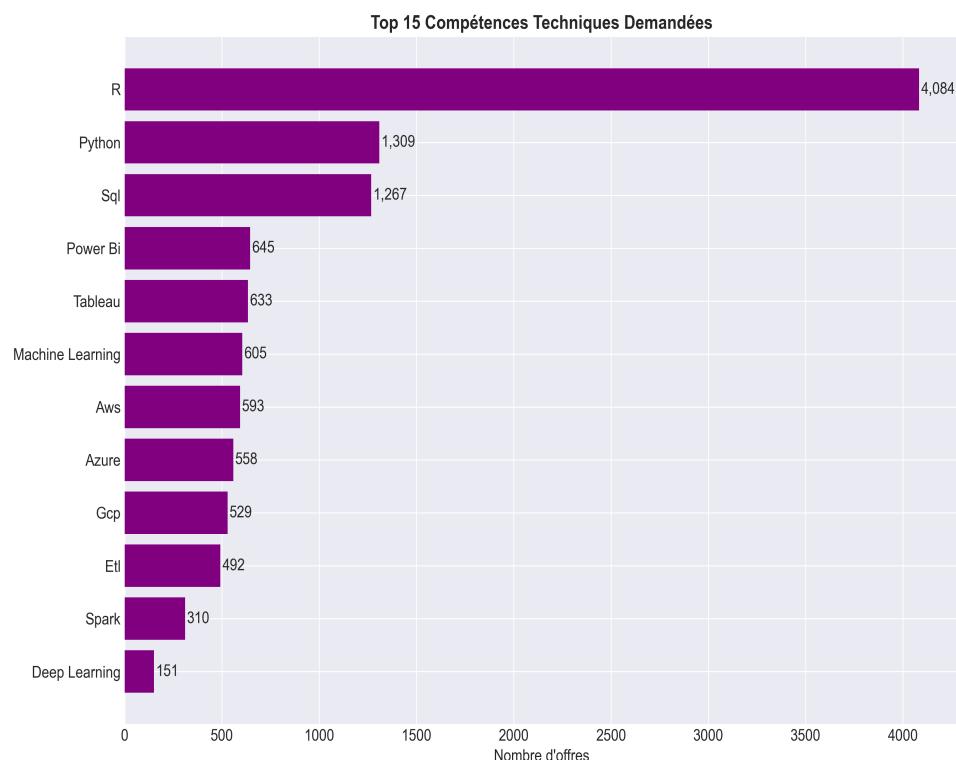


Figure 4 : Compétences techniques les plus demandées

4.6 Analyse des Entreprises

L'analyse des entreprises recruteuses révèle **1 391** entreprises différentes dans le dataset. En moyenne, chaque entreprise publie **4.1** offres.

Top 10 Entreprises qui Recrutent le Plus :

Entreprise	Nombre d'Offres
Sopra Steria	102
Onepoint	87
Safran	83
Mistral AI	80
AP-HP	70
Pennylane	67
FITECO	65
CACEIS	61
Inetum	48
SPIE Building Solutions	47

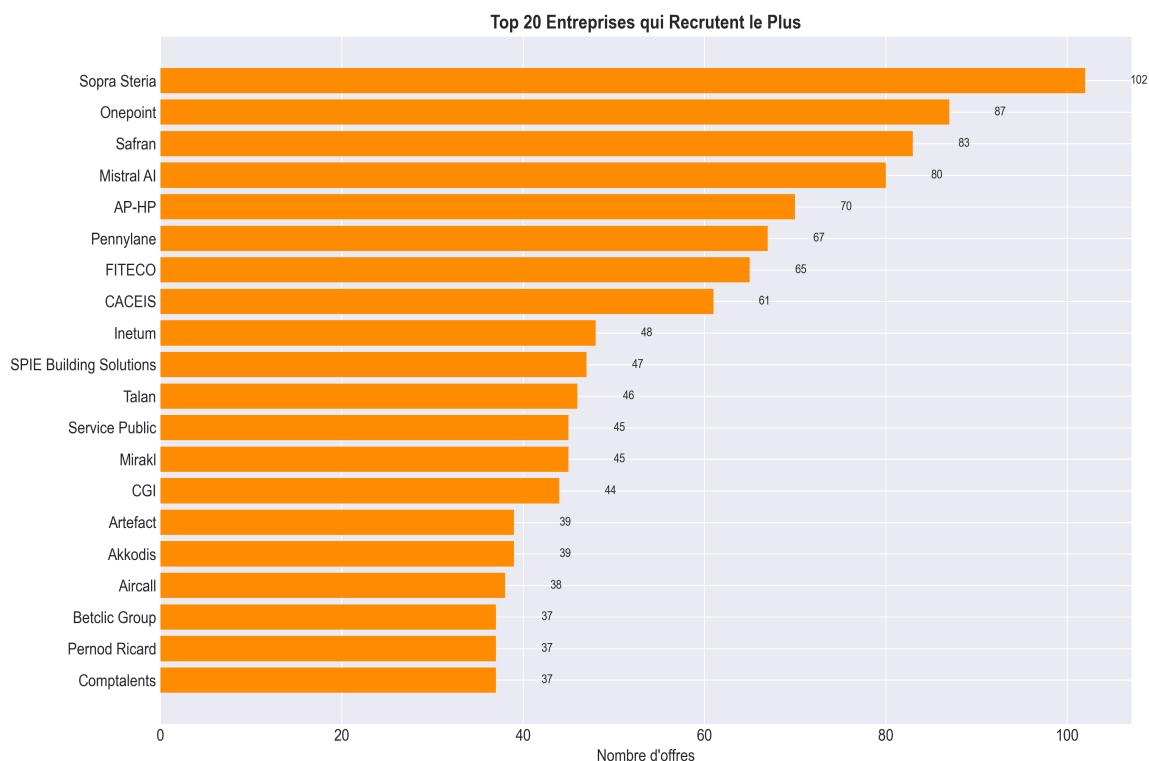


Figure 5 : Entreprises qui recrutent le plus

5. MODÉLISATION PRÉDICTIVE

Cette section présente le développement d'un modèle de Machine Learning pour prédire les salaires des postes Data en fonction de leurs caractéristiques. L'objectif est de créer un outil fiable et généralisable permettant d'estimer la rémunération d'un poste à partir de ses attributs.

5.1 Préparation des Données

Les données ont été préparées selon un processus rigoureux pour garantir la qualité du modèle :

Métrique	Valeur
Échantillons d'entraînement	2 144
Échantillons de test	537
Total échantillons	2 681
Nombre de features	29
Ratio train/test	80% / 20%

Split stratifié : Le dataset a été divisé de manière stratifiée selon les tranches de salaire pour garantir une représentation équilibrée dans les ensembles d'entraînement et de test. Cette approche assure que le modèle est évalué sur des données représentatives de toute la distribution salariale.

5.2 Comparaison des Modèles

Sept algorithmes de Machine Learning ont été comparés pour identifier le plus performant :

Modèle	Test R ²	CV MAE (€)	Overfitting	Score
XGBoost	0.337	5 188€	0.140	0.415
LightGBM	0.346	5 197€	0.181	0.411
GradientBoosting	0.326	5 211€	0.206	0.404
ElasticNet	0.261	5 671€	-0.005	0.369
Lasso	0.262	5 631€	0.006	0.367
Ridge	0.261	5 639€	0.009	0.366
RandomForest	0.259	5 840€	0.025	0.365

Métriques expliquées :

- **Test R²** : Coefficient de détermination (proche de 1 = meilleur)
- **CV MAE** : Mean Absolute Error en cross-validation (plus bas = meilleur)
- **Overfitting** : Différence R² train-test (proche de 0 = bon)
- **Score Composite** : Métrique combinée performance + stabilité - overfitting

L'évaluation de sept algorithmes de Machine Learning révèle une hiérarchie claire basée sur la précision prédictive (\$R^2\$) et l'erreur moyenne (\$MAE\$) :

- **Performance de tête** : Le modèle **XGBoost** s'impose comme la solution la plus équilibrée avec le meilleur score global (0,415) et l'erreur la plus faible (5 188 €), suivi de très près par **LightGBM**.
- **Précision et Robustesse** : Bien que LightGBM affiche le \$R^2\$ le plus élevé (0,346), XGBoost présente un indice d'overfitting plus contenu (0,140 contre 0,181), ce qui en fait un choix plus robuste pour la généralisation.
- **Modèles Linéaires vs Ensembles** : Les modèles basés sur le boosting (XGBoost, LightGBM, GradientBoosting) surpassent nettement les approches linéaires comme ElasticNet ou Lasso, qui peinent à capturer la complexité des données avec des scores inférieurs à 0,370.
- **Limites du RandomForest** : Contre-intuitivement, le RandomForest affiche ici la performance la plus faible avec une erreur moyenne de 5 840 €, confirmant que les algorithmes de gradient boosting sont mieux adaptés à la structure spécifique de ce jeu de données salariales.



Figure 6 : Comparaison des performances des modèles

5.3 Performance du Meilleur Modèle

Le modèle **XGBoost** a été sélectionné comme le plus performant sur la base de son score composite intégrant performance, stabilité et généralisation.

Métrique	Train	Test	Cross-Validation
MAE (€)	4 514€	5 163€	5 188€ ± 183€
RMSE (€)	6 062€	6 945€	-
R ² Score	0.477	0.337	0.317 ± 0.043

Interprétation : Le modèle affiche une erreur moyenne de **5 163€** sur le set de test, soit environ **10.5%** d'erreur relative. Le R² de **0.337** indique que le modèle explique **33.7%** de la variance des salaires.

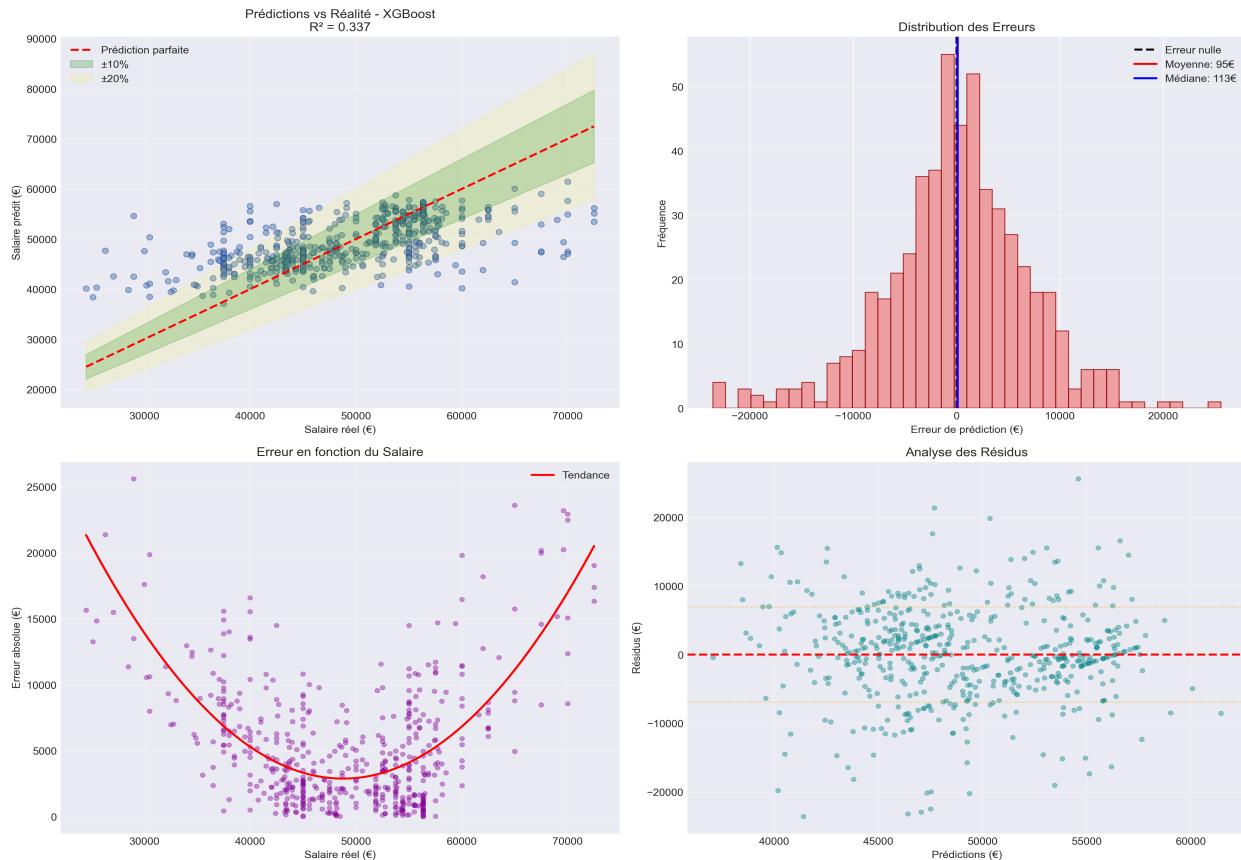


Figure 7 : Analyse détaillée du modèle XGBoost

5.5 Analyse Approfondie des Erreurs et des Résidus

Cette analyse (Figure 7) examine le comportement du modèle XGBoost pour comprendre non plus seulement sa performance globale, mais la nature et la localisation de ses erreurs de prédiction.

1. Fiabilité des Prédictions (Scatter Plot) :

Le modèle affiche une zone de stabilité marquée pour les salaires compris entre 40 000 € et 55 000 €, où la majorité des prédictions restent dans une marge d'erreur de ±10%. Cependant, on observe une sous-estimation pour les salaires supérieurs à 60 000 €, le modèle ramenant ces profils

atypiques vers la moyenne du marché.

2. Analyse des Résidus :

L'erreur moyenne est quasi nulle (95 €), ce qui démontre l'absence de biais systématique. Toutefois, le graphique des erreurs révèle un phénomène en "U" : le modèle est extrêmement précis sur le cœur du marché, mais l'erreur absolue explose pour les salaires extrêmes (inférieurs à 30 000 € ou supérieurs à 65 000 €).

3. Interprétation du Score R² (0,337) :

Ce plafonnement de la performance s'explique par l'hétéroscédasticité (variance de l'erreur non constante) et l'absence de variables discriminantes pour les hauts salaires, telles que la taille exacte de l'entreprise et des niches technologiques très spécifiques.

Note pour l'utilisateur : La prédiction est fiable pour un profil Standard ou Mid-level. Pour un profil Junior ou Ultra-Senior, les résultats doivent être interprétés avec prudence, l'erreur pouvant dépasser 15 000 €.

5.4 Importance des Features

L'analyse de l'importance des features permet d'identifier les variables qui contribuent le plus à la prédiction des salaires. Cette information est cruciale pour comprendre les facteurs déterminants de la rémunération dans le secteur Data.

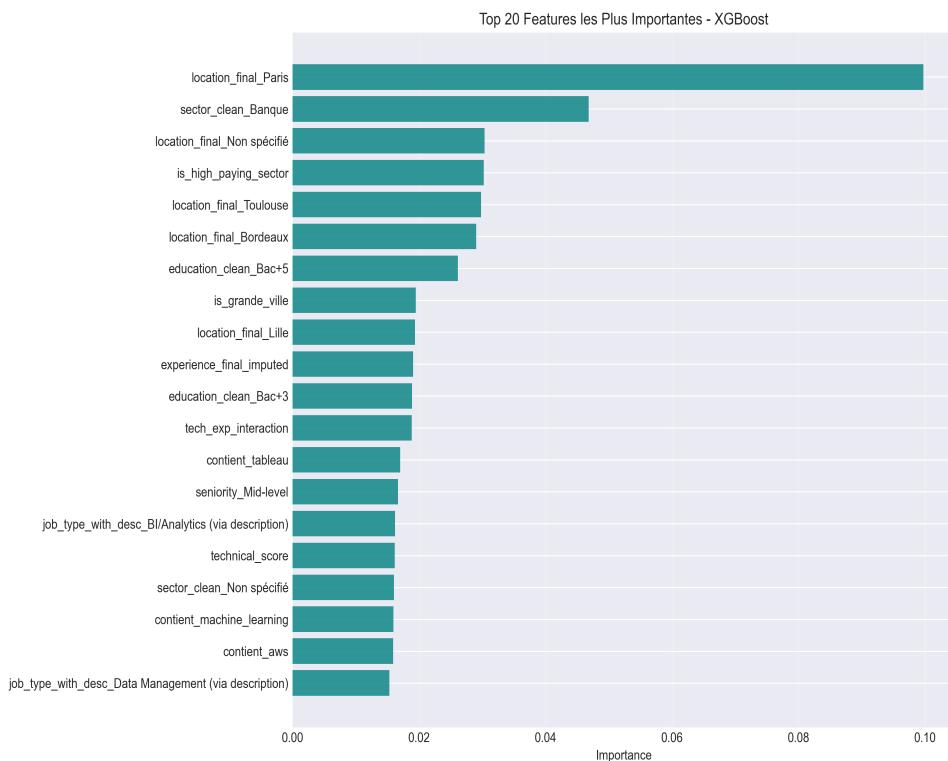


Figure 8 : Importance des features dans le modèle

L'analyse de l'importance des variables (Feature Importance) du modèle XGBoost révèle les principaux leviers qui influencent la détermination des salaires dans le secteur de la Data :

- **Prédominance Géographique** : La localisation à **Paris** est, de loin, le facteur le plus déterminant pour expliquer les variations de salaire, confirmant une forte concentration de la valeur en région parisienne.
- **Impact Sectoriel** : Le secteur de la **Banque** apparaît comme le deuxième levier le plus influent, soulignant que le domaine d'activité de l'entreprise prime parfois sur l'intitulé technique du poste.
- **Niveau d'Études et Expérience** : Le diplôme **Bac+5** et l'expérience cumulée figurent parmi les variables de tête, validant la reconnaissance académique et la séniорité comme piliers de la rémunération.
- **Compétences Techniques Clés** : Parmi les outils spécifiques, la maîtrise d'**AWS** et du **Machine Learning** ressortent comme les compétences ayant l'impact prédictif le plus fort sur le salaire final.
- **Rôles Stratégiques** : Les types de postes liés au **Data Management** et à la **BI/Analytics** confirment leur importance dans la structure salariale, en cohérence avec les analyses descriptives précédentes.

5.5 Diagnostic de l'Overfitting

Un diagnostic approfondi a été réalisé pour évaluer la capacité de généralisation du modèle et détecter d'éventuels problèmes de sur-apprentissage (overfitting).

Critère	Résultat
Niveau d'overfitting	Modéré
Qualité de généralisation	Bonne
Stabilité du modèle	99.5%

Recommandations :

- Modèle prêt pour le déploiement
- Mettre en place un monitoring basique
- Re-entraîner annuellement

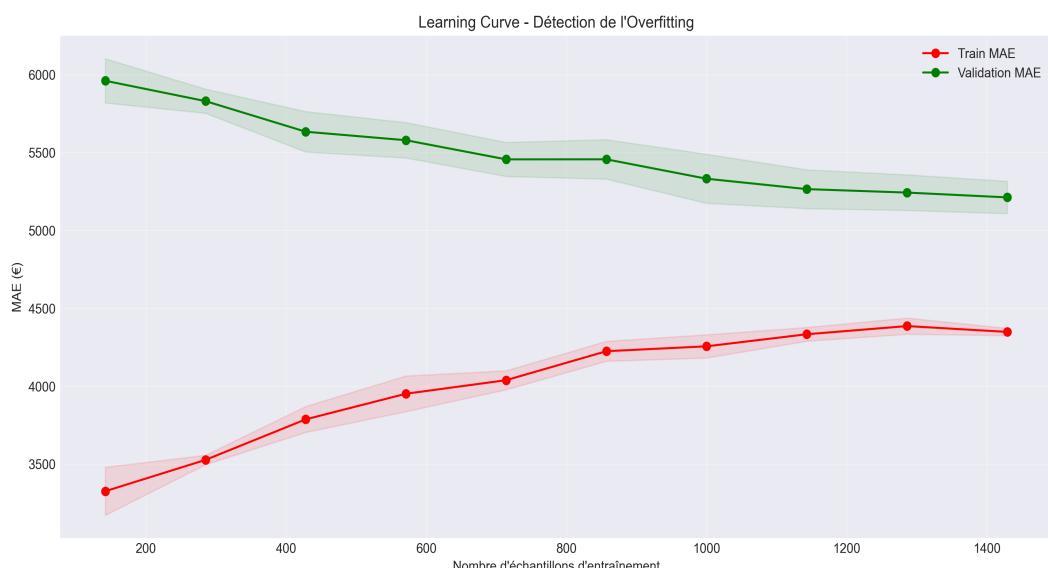


Figure 9 : Learning curve - Convergence du modèle

5.6 Analyse des Courbes d'Apprentissage

L'examen des courbes d'apprentissage (Figure 9) permet d'évaluer la capacité du modèle à généraliser et d'identifier si l'ajout de données supplémentaires pourrait améliorer ses performances.

1. Comportement des Courbes (Train vs Validation) :

- **Courbe d'Entraînement (Rouge)** : L'erreur augmente progressivement à mesure que le modèle intègre un jeu de données plus vaste, se stabilisant autour de 4 514 €.
- **Courbe de Validation (Verte)** : L'erreur diminue progressivement vers 5 188 €, indiquant que le modèle améliore sa capacité de généralisation à mesure que davantage de données sont intégrées.

2. Diagnostic du Modèle :

L'écart (gap) entre les deux courbes révèle une **convergence saine**. Bien qu'un écart résiduel subsiste (environ 674 €), les courbes ne divergent pas, ce qui valide le contrôle de l'overfitting. La légère pente descendante de la courbe de validation suggère que l'acquisition de données supplémentaires pourrait encore affiner la précision.

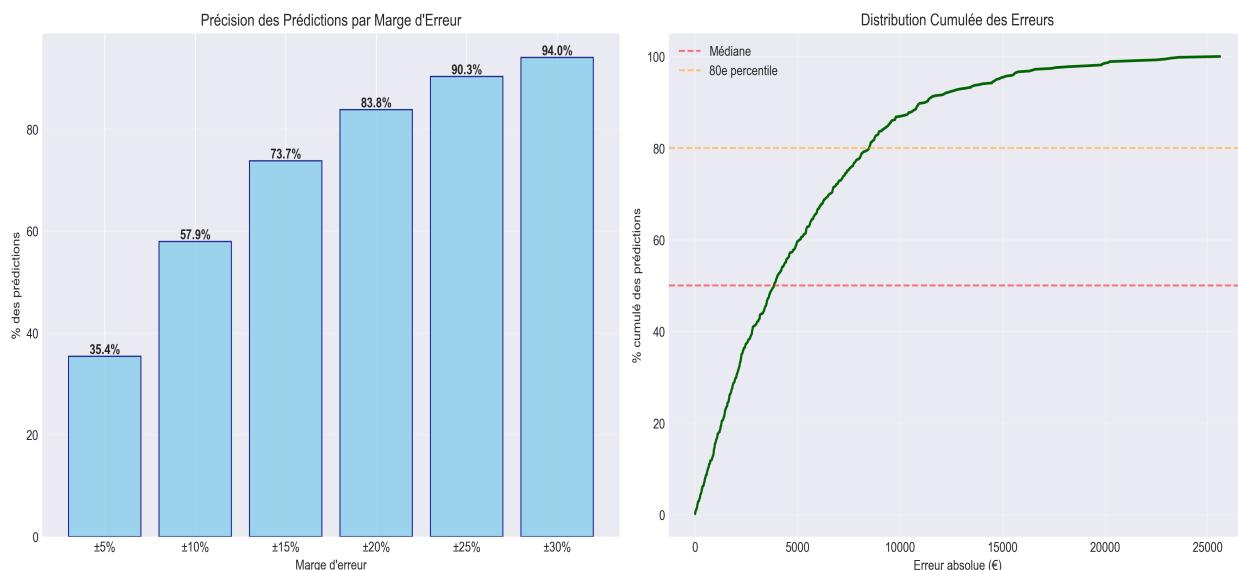


Figure 10 : Précision des prédictions par marge d'erreur

5.7 Analyse de la Précision et des Seuils de Fiabilité

Cette analyse (Figure 10) permet de traduire les métriques mathématiques en indicateurs concrets de fiabilité pour une utilisation opérationnelle du modèle.

1. Analyse par Marges d'Erreur :

- Précision élevée ($\pm 5\%$)** : Environ 35,4 % des prédictions présentent une erreur très faible, indiquant une excellente adéquation pour les profils standards.
- Seuil de fiabilité ($\pm 15\%$)** : Près de 3 sur 4 (73,7 %) des prédictions se situent dans cette marge, seuil considéré comme acceptable dans le secteur du recrutement.
- Couverture ($\pm 30\%$)** : Le modèle capture 94 % de couverture. Les 6 % restants correspondent aux salaires atypiques ou extrêmes identifiés précédemment.

2. Distribution Cumulée des Erreurs :

- Erreur médiane** : 50 % des prédictions affichent une erreur absolue inférieure à environ 4 000 €.
- 80e percentile** : 80 % des prédictions ont une erreur inférieure à environ 8 500 €.
- Limites extrêmes** : L'aplatissement de la courbe au-delà de 15 000 € d'erreur confirme la difficulté du modèle à traiter les profils "hors normes".

6. CONCLUSIONS ET RECOMMANDATIONS

6.1 Conclusions Principales

L'étude des données issues de HelloWork met en lumière la structure réelle du marché de la donnée en France :

- **Domination de l'ingénierie et de la BI** : Les entreprises privilégient les postes liés à la mise en place et à l'exploitation des systèmes de données, avec une forte présence des **Data Engineers** (27,4 % des offres) et de la **BI/Analytics** (22,7 %).
- **Valorisation de l'expertise scientifique** : Bien que moins nombreux en volume, les postes de **Data Scientist** captent les rémunérations moyennes les plus élevées à 52 920 €.
- **Hégémonie technologique de R et Python** : La forte présence de **R** (69,6 %) et de **Python** (22,3 %) montre que la maîtrise d'un langage de programmation reste un critère clé dans les offres du secteur.
- **Déterminants majeurs du salaire** : Les données montrent que travailler à **Paris** ou dans la **Banque** est fortement corrélé à un niveau de salaire supérieur.
- **Fiabilité du modèle prédictif** : Avec une erreur moyenne de 5 163 €, **XGBoost** fournit des estimations salariales globalement alignées avec les variations du marché.

6.2 Recommandations pour les Professionnels

- **Privilégier le stack R / Python / SQL** : Ces trois langages constituent les compétences les plus régulièrement associées aux profils recherchés.
- **Acquérir des compétences Cloud** : La maîtrise d'**AWS(Amazon Web Service)** est identifiée par le modèle comme un levier direct d'augmentation de la valeur sur le marché.
- **Cibler les secteurs à haute valeur** : Pour une rémunération optimale, orienter sa recherche vers le secteur de la **Banque** ou du **Data Management**.
- **Négocier la mobilité** : Paris reste le pôle majeur de rémunération ; lorsque la mobilité n'est pas possible, le télétravail pour des entreprises franciliennes est une stratégie viable.

6.3 Recommandations pour les Recruteurs

- **Ajuster les offres d'ingénierie** : Les Data Engineers représentant 27,4 % des annonces, maintenir des niveaux de rémunération autour de 50 000 € contribue à rester aligné avec les pratiques observées sur le marché.
- **Anticiper la prime à l'expertise** : Les profils orientés Data Science ou Conseil stratégique se situent généralement au delà de 52 000 €, ce qui invite à prévoir des budgets adaptés.
- **Standardisation des descriptifs** : La fréquence élevée de postes identifiés 'via description' indique qu'une clarification des intitulés dès la publication pourrait faciliter le sourcing.
- **Valoriser la formation continue** : L'écart entre la forte demande en R et l'essor des compétences IA met en évidence l'intérêt d'investir dans la montée en compétences des équipes.

6.4 Perspectives Futures

- **Analyse sectorielle approfondie** : Examiner les raisons pour lesquelles le secteur de la Banque se distingue en termes de rémunération.
- **Optimisation du modèle** : Limiter l'overfitting observé (écart entre Train 0.47 et Test 0.33) en intégrant de nouvelles variables contextuelles et en ajustant les hyperparamètres.
- **Extension du jeu de données** : Augmenter le volume et la diversité des annonces collectées afin d'améliorer la stabilité du modèle et sa capacité de généralisation.
- **Suivi de l'IA** : Observer l'évolution de la catégorie AI/ML (actuellement 3,4 %) pour identifier un éventuel changement de dynamique dans son adoption.

7. ANNEXES

7.1 Détails Techniques

Technologies utilisées :

- **Collecte** : Python (BeautifulSoup, Selenium)
- **Traitement** : Pandas, NumPy
- **Visualisation** : Matplotlib, Seaborn, Plotly
- **Modélisation** : Scikit-learn, XGBoost, LightGBM
- **Reporting** : ReportLab

Infrastructure :

- Python 3.11+
- Environnement : Ubuntu 24
- Traitement local avec optimisation mémoire

7.2 Features Utilisées dans le Modèle

Les features suivantes ont été utilisées pour entraîner le modèle prédictif :

- | | |
|-----------------------|-----------------------------|
| • job_type_with_desc | • contient_spark |
| • seniority | • contient_machine_learning |
| • contract_type_clean | • contient_etl |
| • location_final | • skills_count |
| • sector_clean | • technical_score |
| • education_clean | • has_teletravail |
| • experience_final | • has_mutuelle |
| • contient_sql | • has_tickets |
| • contient_python | • has_prime |
| • contient_r | • benefits_score |
| • contient_tableau | • telework_numeric |
| • contient_power_bi | • is_grande_ville |
| • contient_aws | • description_word_count |
| • contient_azure | • nb_mots_cles_techniques |
| • contient_gcp | |

7.3 Glossaire

Terme	Définition
MAE	Mean Absolute Error - Erreur absolue moyenne

RMSE	Root Mean Squared Error - Racine de l'erreur quadratique moyenne
R ²	Coefficient de détermination - Mesure de la qualité d'ajustement
Overfitting	Sur-apprentissage - Le modèle mémorise les données d'entraînement
Cross-Validation	Validation croisée - Technique d'évaluation robuste
Feature Engineering	Création de variables dérivées à partir des données brutes
Pipeline	Séquence de transformations et d'apprentissage
Régularisation	Technique pour limiter la complexité du modèle
Split stratifié	Division des données en préservant les proportions des classes

Fin du rapport

Généré automatiquement le 29/01/2026 à 22:20

© 2025 - Analyse du Marché de l'emploi Data en France