Projet n°7 :
“Lead Scoring”

Soutenance de Jury - Projet n°7 - Juin 2022

Ali Naama

Sommaire

I. Description du contexte et exploration du jeu de données

Description du contexte

- Une grande **entreprise française spécialisée dans la fourniture de matériaux de construction** souhaiterait **optimiser sa gestion des prospects (Lead) B2B**.
- Actuellement les **équipes Marketing** de cette entreprise disposent des outils Pardot et Salesforce CRM afin de cibler les prospects par email. Les prospects sont récupérés par différents canaux : Via les commerciaux qui récupèrent des cartes de visites dans des salons spécialisés ou via l'acquisition de fichier de prospection par l'équipe Marketing.
- Les Applications **Pardot et Salesforce CRM** sont intégrées partiellement et il n'y a pas encore d'intégration fluide entre les activités de prospections Pardot et les équipes commerciales utilisant le CRM Salesforce.
- Dans la première phase du projet lancée en 2022, une **interaction optimisée** entre le CRM et Pardot a été mise en œuvre et permet maintenant une visibilité des activités de sollicitations des contacts existants au niveau du CRM.
- En revanche, il manque la partie prospection de « **Lead** » qui n'est pas adressée et qui demeure visible uniquement dans Pardot par les équipes marketing. Il manque également la partie intégration avec la partie Digitale (Site Web, Réseaux Sociaux).
- **A ce stade, les décideurs ont besoin d'un POC sur la partie prospection de Lead afin de lancer les investissements sur cette phase de projet.**

Objectifs :

Proposer une démarche projet adaptée au contexte client, en proposant une **modélisation du taux de conversion** afin d'aider les équipes commerciale à accroître le taux de conversion.

Trouver les **principales variables du modèle** de prédiction qui contribuent à un meilleur taux de conversion d'un prospect et le modèle de prédiction le plus approprié pour augmenter **le taux de conversion de Leads**.

Leads - Définitions utiles

Lead

Une Piste est un prospect ou une opportunité potentielle - une personne qui peut manifester un intérêt pour les produits de la marque

conversion

Contact

Les contacts sont toutes les personnes associées à vos comptes professionnels que vous devez suivre dans Salesforce. Vous pouvez stocker diverses informations pour un contact, telles que les numéros de téléphone, les adresses, les titres et les fonctions.

Génération de Leads :

La génération de prospects est l'initiation de l'intérêt ou de la demande des clients pour les produits ou services de votre entreprise. Les prospects sont créés dans le but de convertir l'intérêt ou la demande en ventes.

Qualification des Leads :

La qualification des prospects fait référence au processus de détermination des clients potentiels les plus susceptibles d'effectuer un achat réel.

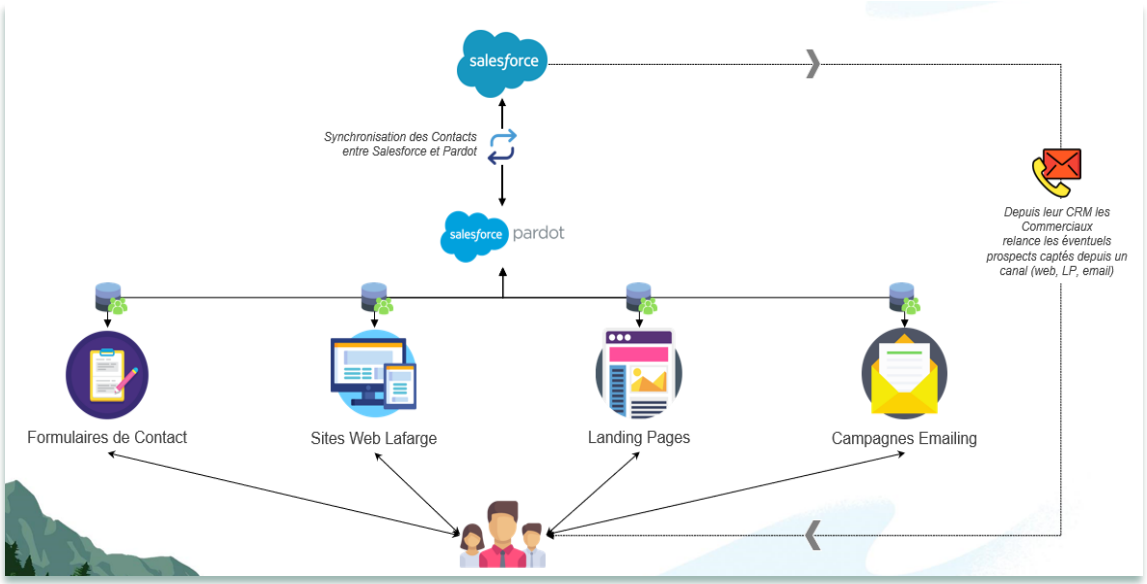
Conversion des Leads :

La conversion des prospects est une phase où vous convertissez enfin un prospect qualifié en client payant. Il s'agit de l'ensemble des pratiques marketing qui stimulent le désir d'acheter un produit ou un service et poussent une avance vers une décision d'achat. Il s'agit d'une phase de monétisation ou de clôture et le résultat de celle-ci définit généralement le succès de la campagne marketing globale.

Leads – Canaux d'acquisition

- Les Leads sont un vecteur clé du développement des entreprises et la **capacité à convertir** les prospects en clients payants est la clé du succès.
- Lors de l'acquisition de Lead, par différents canaux, comme : un service tiers, une campagne marketing, cela intègre les informations telles que : Nom, Prénom, Email, …

Contexte Client



Lead, Contact

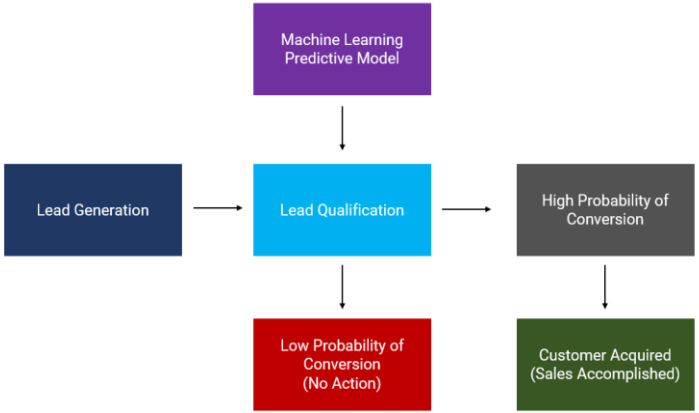
Formulaire de saisie permettant la création du Lead
CRM Salesforce propose une structure similaire permettant d'accueillir ces données

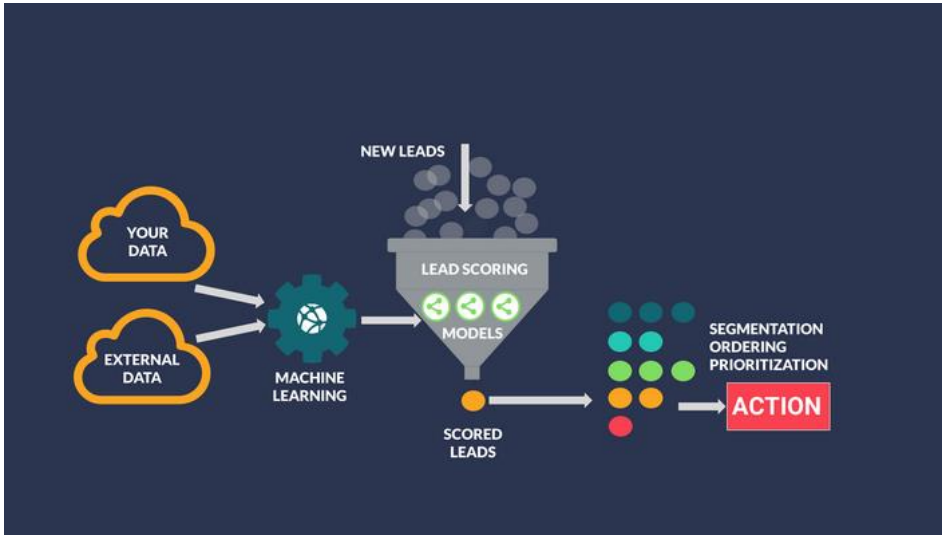
Champ	Valeur d'exemple
Nom	Pierre
Prénom	Legrand
Adresse	Rue de la Paix - Paris
Téléphone	0745124578
Email	pleg@ggmail.com
Source	Fichier, Salon, Web, LinkedIn, Facebook
Produit	GEN35, ..
Campagne	PRINTEMPS22
RGPD (Consentement)	Oui / Non

Leads - Taux de conversion

- **Comment faire pour maximiser le taux de conversion de Lead ?**
  - L'approche optimale est de mettre en place un modèle d'apprentissage automatique (ML) qui utilisera les attributs du client, l'origine du prospect, les références et d'autres détails disponibles afin de déterminer la **probabilité de conversion** ou non du Lead.
  - Le systèmes CRM Salesforce permet de suivre le statut du Lead, ce qui aidera à créer la variable cible : Conversion : Oui / Non

Le processus de gestion des prospects en un coup d'œil





Exemple de Lead (Piste) – Salesforce CRM Client

Piste
Julien Da Silva

+ S'abonner | Modifier | Supprimer | Cloner

Société [[Unknown]] | Fonction | Téléphone (2) | Adresse e-mail | Pardot Campaign: Base contacts BATI FID (mars 2021) | Pardot Score: 0

Nouveau | En cours | Nurturing | Converti | Marquer Statut comme terminé(e)

Champs clés

Société [[Unknown]]
Site Web
Secteur d'activité
Chiffre d'affaires annuel
Nombre d'employés

Modifier Guide de réussite

Renseignez-vous et contactez rapidement votre nouveau prospect.

Utilisez les prospects Salesforce pour séparer les prospects des contacts Salesforce auxquels vous avez déjà vendu.

- Répondez à votre prospect dans les cinq minutes pour augmenter vos chances de convertir l'enregistrement de prospect en opportunité.
- Consultez le site Web de votre prospect pour en savoir plus sur l'activité et le secteur de votre prospect.
- Trouvez votre prospect sur les réseaux sociaux pour déterminer la meilleure façon de vous engager et de vous connecter.

Détails | Chatter

Statut de la piste: Nouveau
Propriétaire de la piste: Marketing
Site Web
Société: [[Unknown]]
Téléphone
Nombre d'employés
Téléphone mobile
Origine de la piste
Service
Fonction
Cote
Devise de la piste: EUR - Euro

Coordonnées

Adresse: 93370

Engagement History (2)

Email Open — 13 days 21 hrs ago
Email: Relance dépôt de facture 52,5R
List: Base clients BATI FID (mars 2021)

Email Sent — 14 days 20 hrs ago
Email: Relance dépôt de facture 52,5R
List: Base clients BATI FID (mars 2021)

Consigne... | Nouvelle ... | Nouvel é... | E-mail

Récapitulez votre appel...

Ajouter

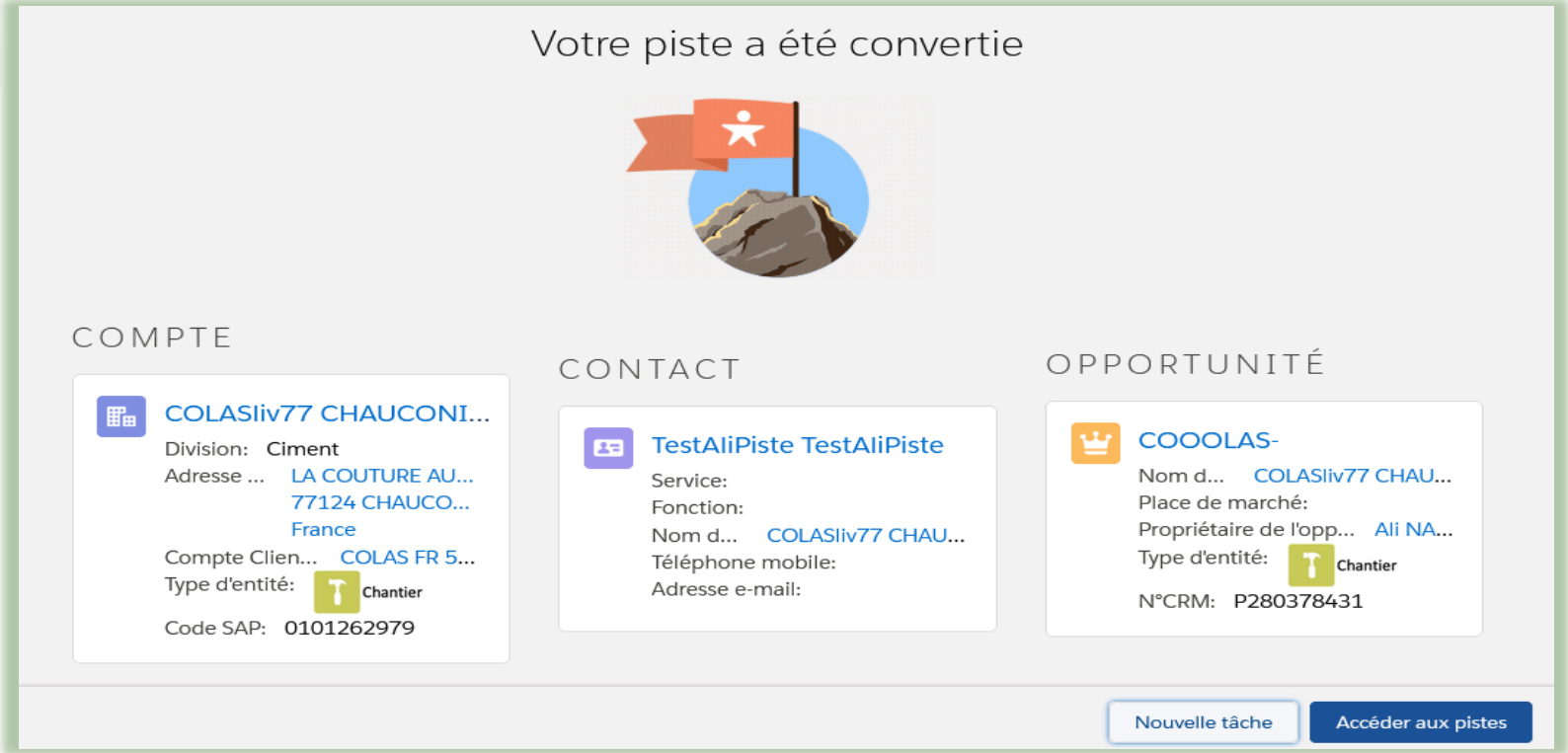
Filtres : Depuis 2 mois • Toutes les activités • Tous les types

Actualiser • Agrandir tout • Afficher tout

À venir et passés

Aucune étape suivante.
Pour continuer, ajoutez une tâche ou définissez une réunion.

Lead Converti



- Lead conterti, le prospect est transformé en client : Contact et Compte Entreprise (B2B)

Jeu de données « Lead » adapté au contexte client

Informations sur les variables :

- ❑ Les données : Lead Conversion sont extraites du site Kaggle et correspondent à un contexte de ventes de cours en lignes
- ❑ Les colonnes vont être adaptées en fonction du contexte client : renommage et suppression de certain attributs, ajout de la colonne CreatedDate, valeur Select substitué par de vrais valeurs par imputation
- ❑ Les valeurs des colonnes : Country et City vont être adaptés au contexte client pour ne prendre que des villes française
- ❑ Colonne Segmentation qui est structurante pour le client remplace la colonne Specialization
- ❑ Le fichier résultant de ces opérations devient le fichier Leads_France.csv

Nombre de colonnes du fichier : 37

Nombre de lignes du fichier : 9240

Fichier	Nb lignes	Nb colonnes	Taux remplissage moyen	Doublons	Description
Leads_France.csv	9240	37	**90.1%**	0	Leads_France.csv

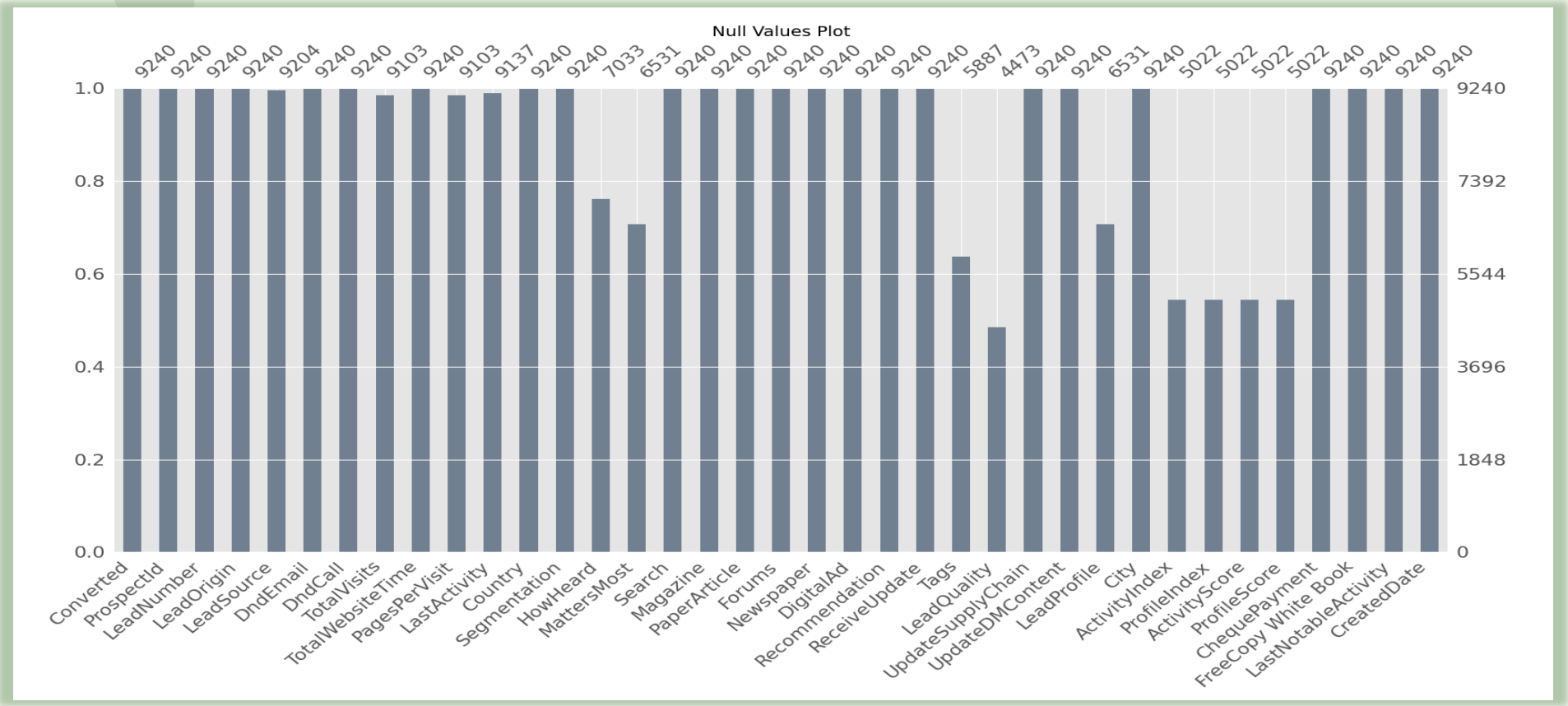
Valeurs clés du jeu de données

Champ	Valeur d'exemple	Type	Description
Prospect ID			Identifiant unique du Prospect
Lead Origin			Origine du Lead :
Lead Source			Source du Lead :
Do Not Email			Indicateur Email Opt Out
Do Not Call			Indicateur Call Opt Out
Converted			Indicateur de Lead Converti
Country			Pays : dans notre cas uniquement France
City			Ville
Segmentation			Segmentation client utilisé pour le ciblage
Total Time Spent on Website			Temps passé sur le site Web
TotalVisits			Nombre total de visite
Page Views Per Visit			Nombre de page vues sur le site
CreatedDate			

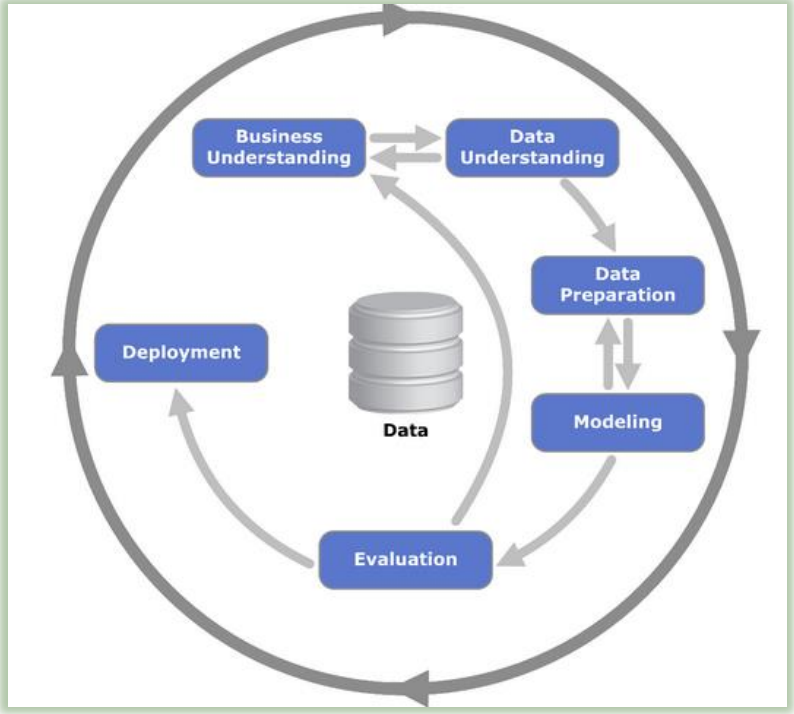
Nombre de colonnes du fichier : 37

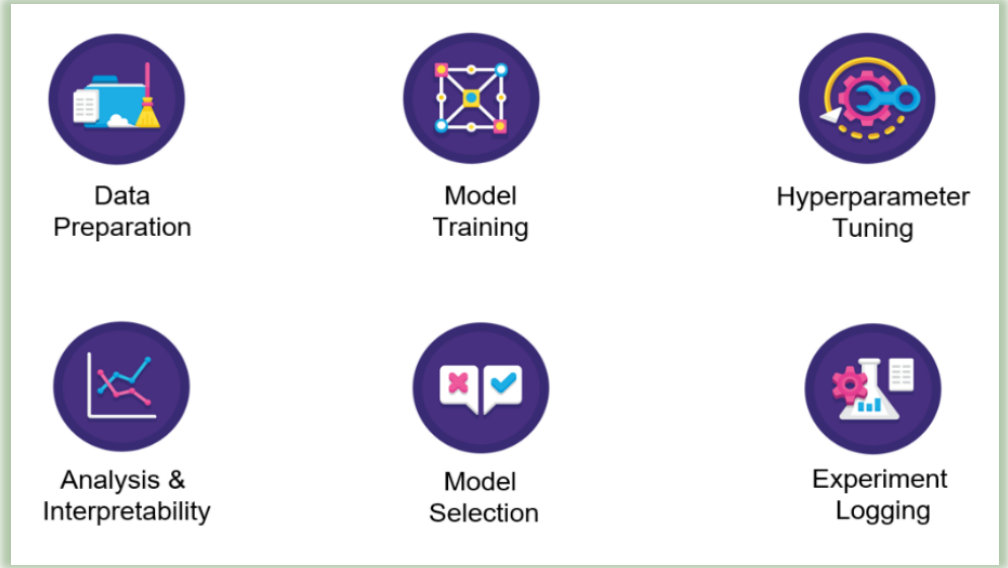
Nombre de lignes du fichier : 9240

Leads – Taux de Remplissage par colonne



Méthodologie de Data Science « CRISP DM »





Outils utilisés pour l’analyse

Nom	Utilisation	Fonctions spécifiques
PyCharm 2021.1 **Jupyter Notebook**	Test et développement	IDE Community Edition, Debug, Synchro Git
Python 3.9.5	Moteur Python Gestionnaire de librairies	Moteur d’exécution
Pandas 1.4.0	Librairie de manipulation de données Représentation des données	Manipulation de Dataframe : création, copie, filtres, tris, description, concaténation, pivotage, autre
Matplotlib 3.5.1, Seaborn 0.11.2 **Numpy 1.22.0** **Sklearn** **scipy** **Missingno** **Flask** **Dash**	Génération de graphiques Gestion des densités de probabilité Machine Learning	Barplot, Scatterplot, lineplot, distplot, heatmap Calcul statistique Flask, Dash : Pour la mise en ouevre de Dashboard et du simulateur de Lead

Stratégie d'analyse

1

Découvrir les Données

- Découverte des données Lead
- Taux de Remplissage par colonne
- Type de variables
- Statistiques des variables clés

2

Analyse Exploratoire / Feature Engineering

- Préparation des données
- Renommage des colonnes
- Travaux de Nettoyage et de feature engineering
- Traitements des valeurs manquantes et imputations
- Transformation des variables catégorielles en variables numériques (get_dummy)
- Standardisation des données

3

Modéliser / Comparer / Tester

- Modélisation
- Comparaison des Modèles
- Test
- Optimisation
- Résultats de performance

4

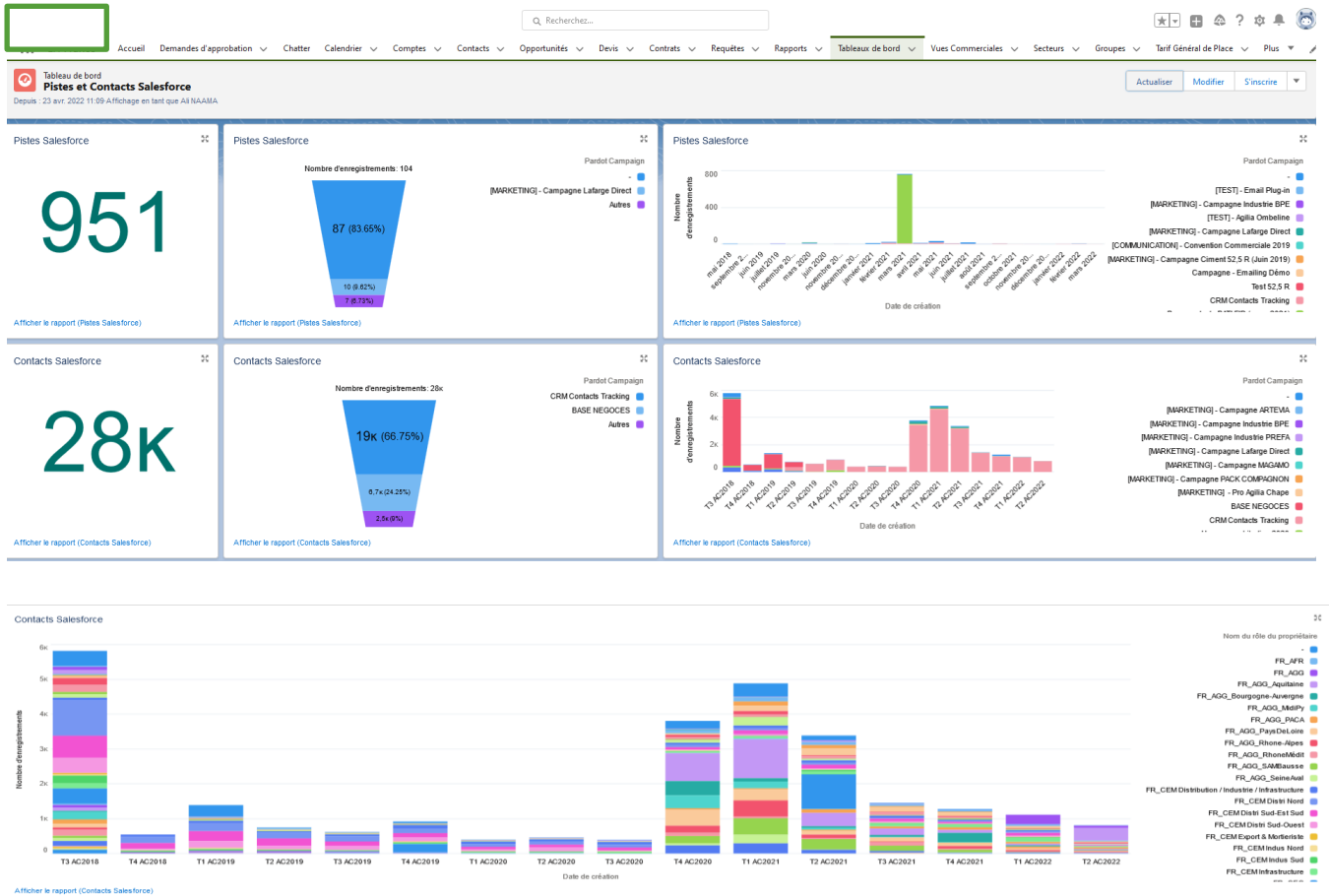
Visualisation / Interprétation / Validation

- Interprétation des Résultats
- Tableau de Bord interactif
- Simulateur
- Revue / Validation Métier

Découverte / Préparation / Analyse / Transformation	Modélisation	Validation
Projet7.py	Projet7_ML.py	Lead_simulator/script.py Lead_dashboard/app.py

II. Analyse Exploratoire

Leads – Dashboard Salesforce



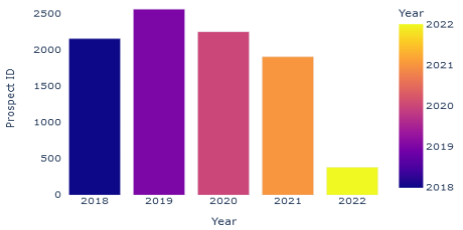
Leads – Dynamic Dashboard - Discovery

Dash by plotly

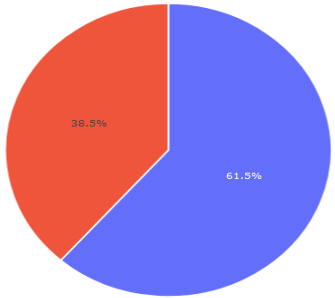
Discovery | Exploration | Machine Learning | Simulation

Leads Data Set Discovery : Number of Records 9240

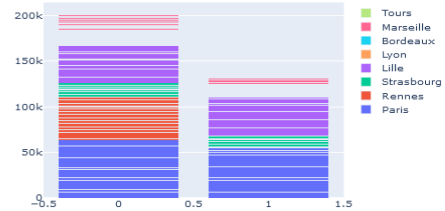
Lead volume per year



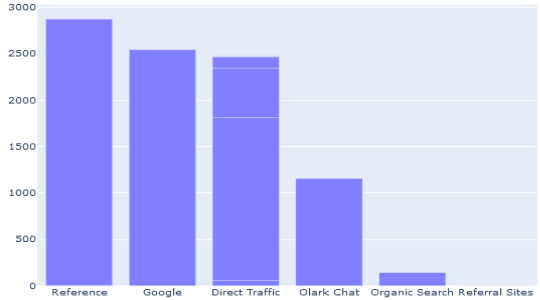
Lead Conversion Rate



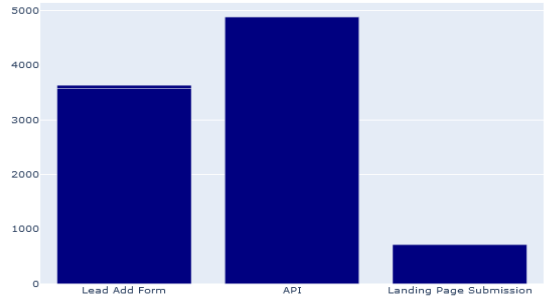
Conversion By City



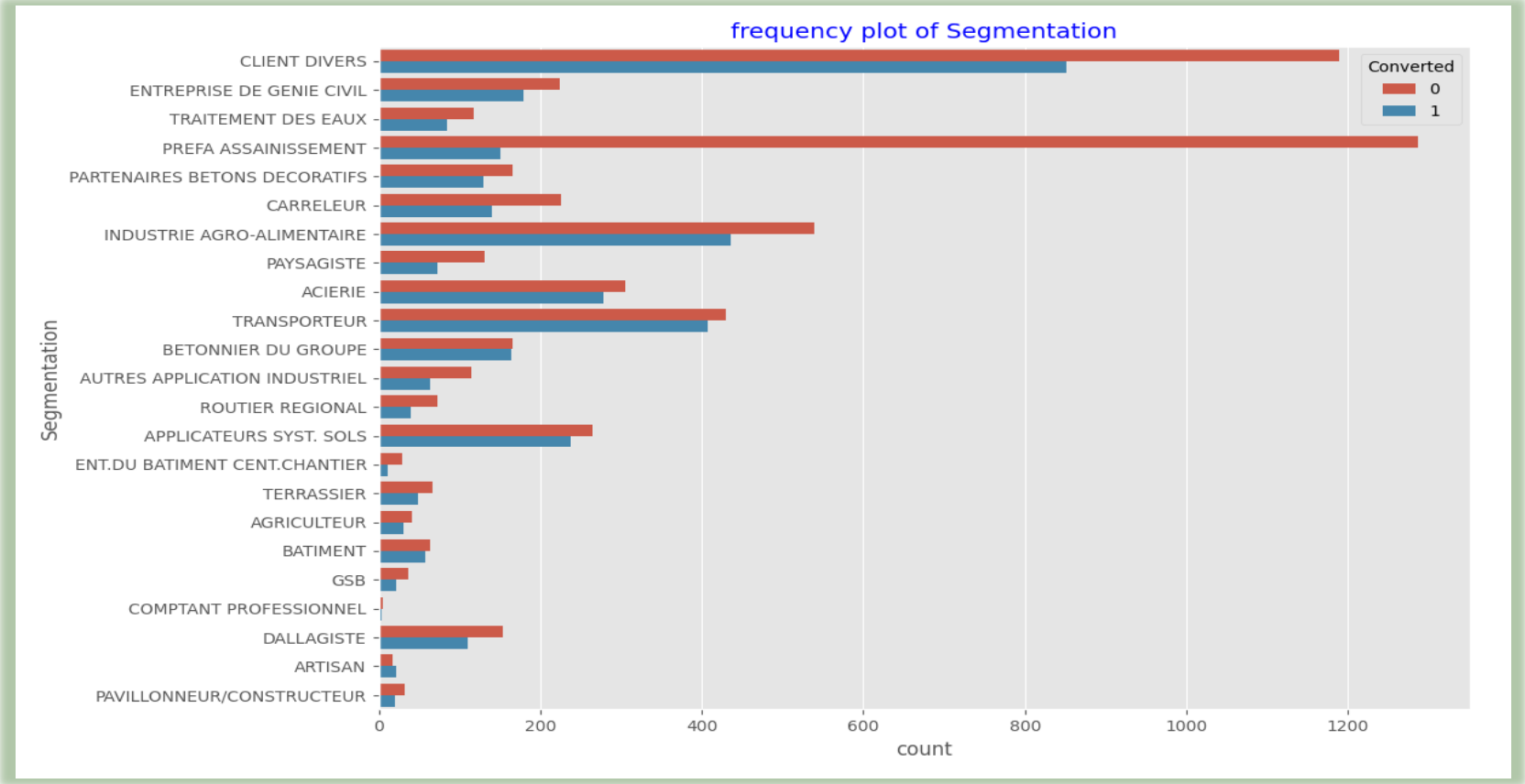
Lead Source



Lead Origin

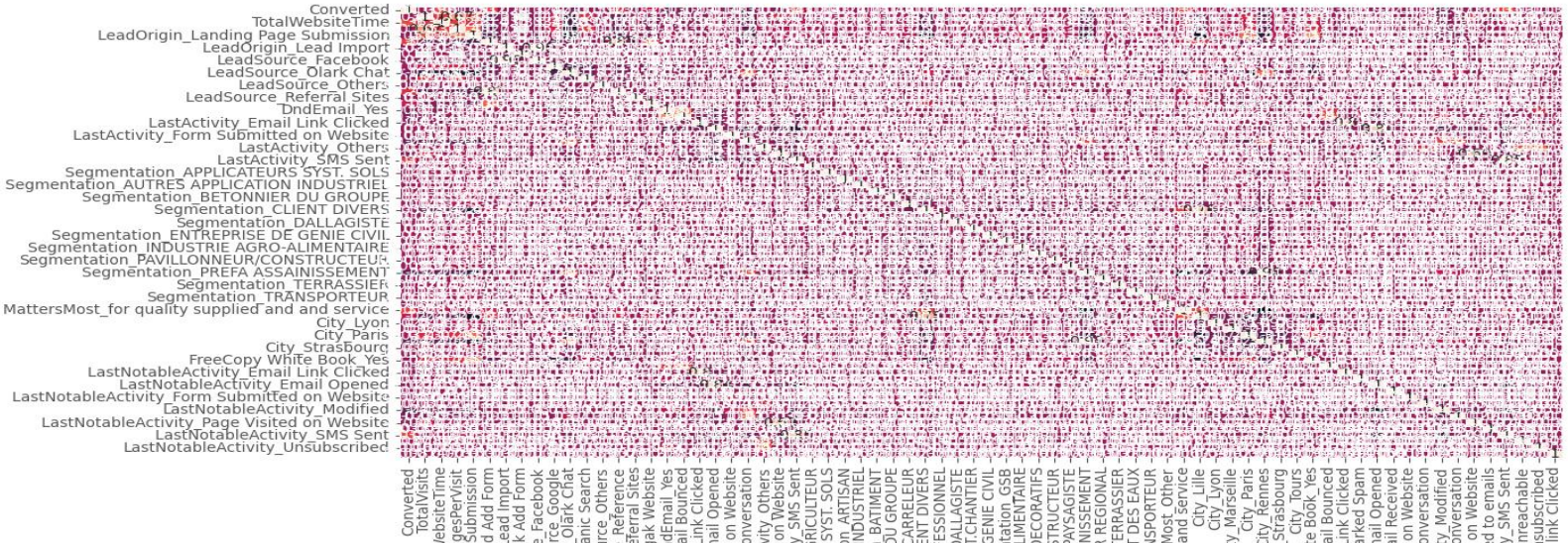


Leads – Dashboard

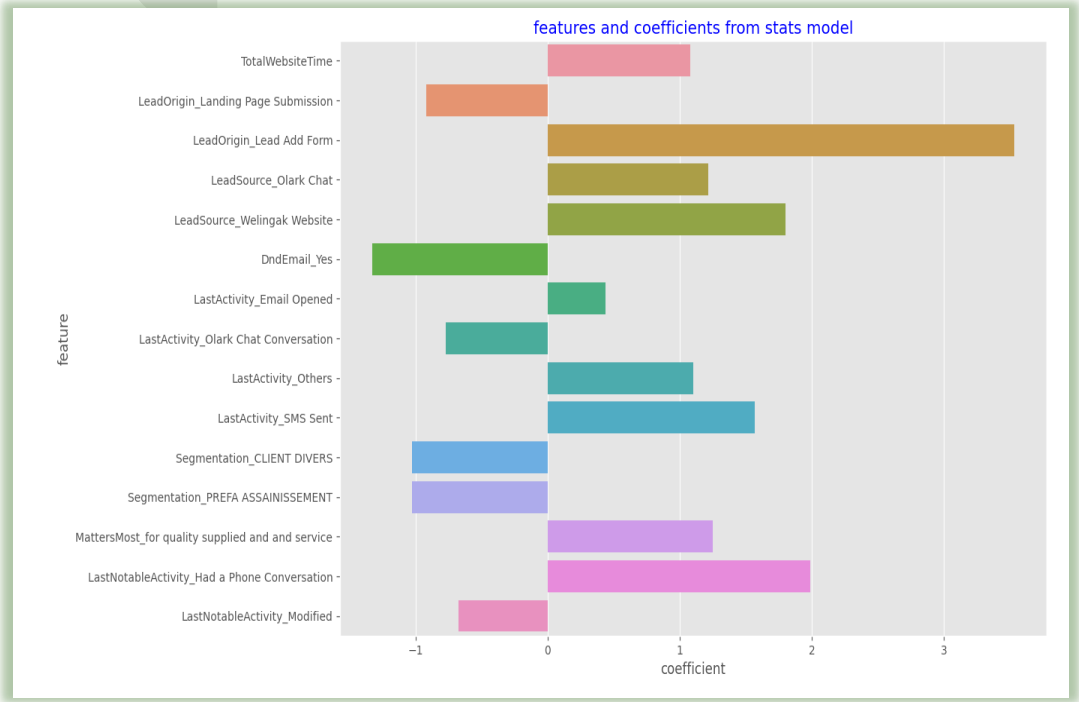


Leads – Dashboard - Exploration

Heat Map - Correlation



Feature importance

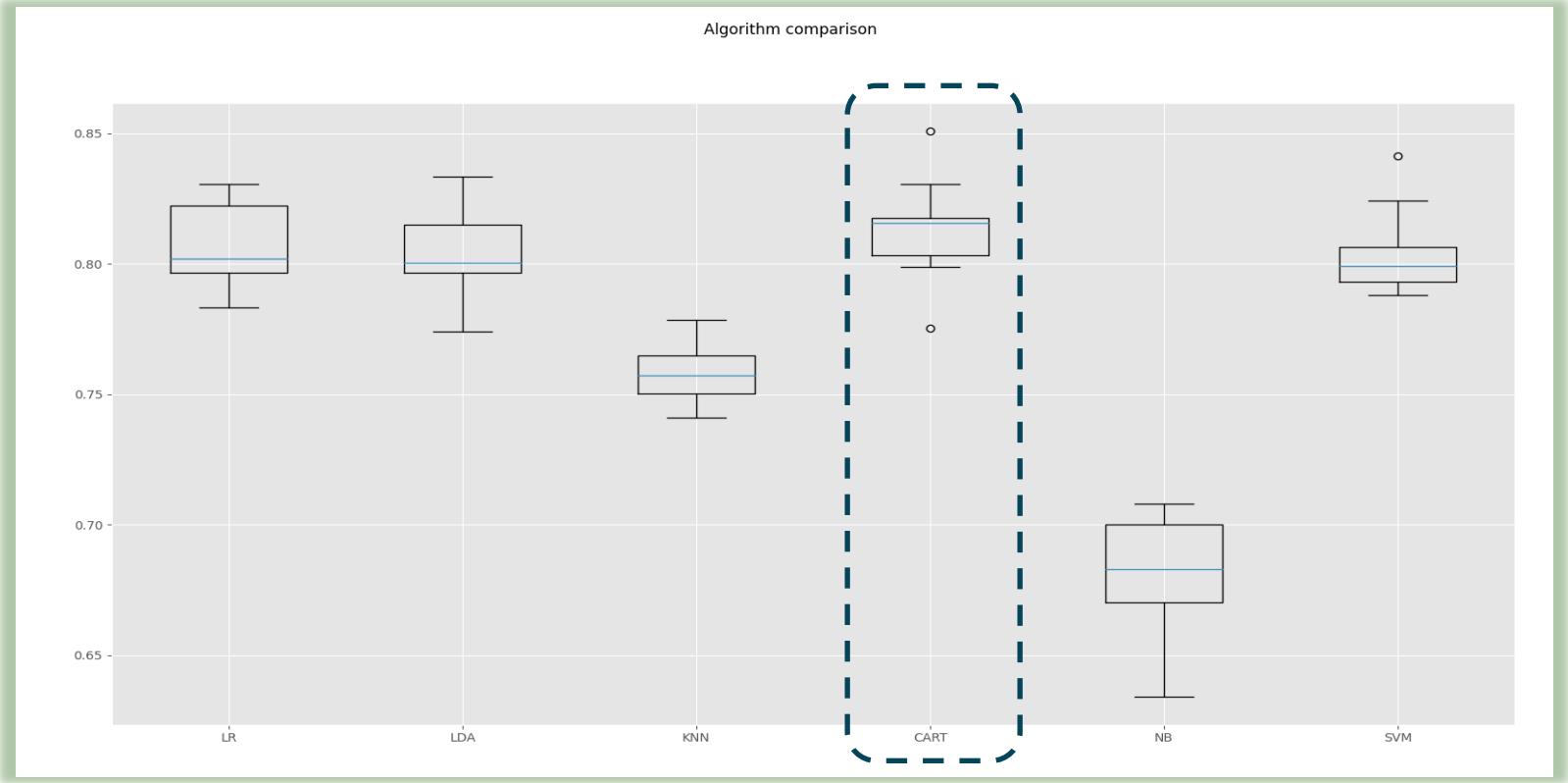



```
*********** feature - coefficients ****************
const                                          -1.4677
TotalWebsiteTime                                1.0789
LeadOrigin_Landing Page Submission             -0.9173
LeadOrigin_Lead Add Form                        3.5335
LeadSource_Olark Chat                           1.2181
LeadSource_Welingak Website                     1.8034
DndEmail_Yes                                   -1.3296
LastActivity_Email Opened                       0.4386
LastActivity_Olark Chat Conversation           -0.7718
LastActivity_Others                             1.1013
LastActivity_SMS Sent                           1.5665
Segmentation_CLIENT DIVERS                     -1.0254
Segmentation_PREFA ASSAINISSEMENT              -1.0252
MattersMost_for quality supplied and and service    1.2487
LastNotableActivity_Had a Phone Conversation    1.9910
LastNotableActivity_Modified                   -0.6726
dtype: float64
***************************************************
```

- Observations : Les features les plus importantes sont : LeadOrigin_Lead Add Form, LastNotableActivity and LeadSource_Welingak Website

III. Modélisation

Comparatif des Algorithmes



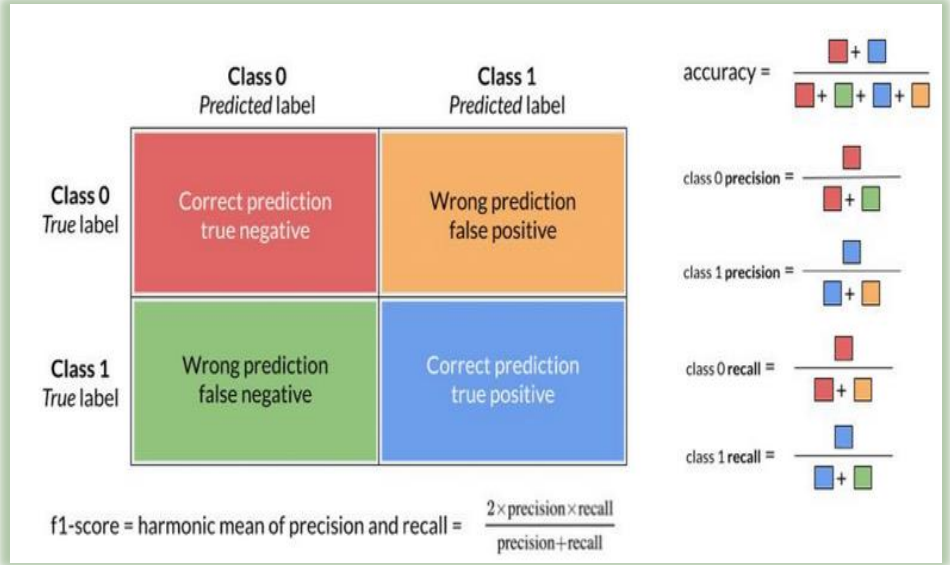
- Le principe du modèle est de modéliser la variable « Converted » en fonction des autres variables en utilisant plusieurs algorithmes
- Utiliser une cross validation via sklearn

Métriques / Performance du Modèle

- **Accuracy :** Taux d'erreur
- **Precision** : Le taux d'erreur, proportion d'individus mal classés doit être le plus bas possible
- **Recall** (Sensitivity) : **doit être > 0,5 pour être correct**
- **F1-Score :** Le F1-score évalue la capacité d'un modèle de classification à prédire efficacement les individus positifs, en faisant un compromis entre la precision et le recall.

```
print(classification_report(y_test,predictions))
```

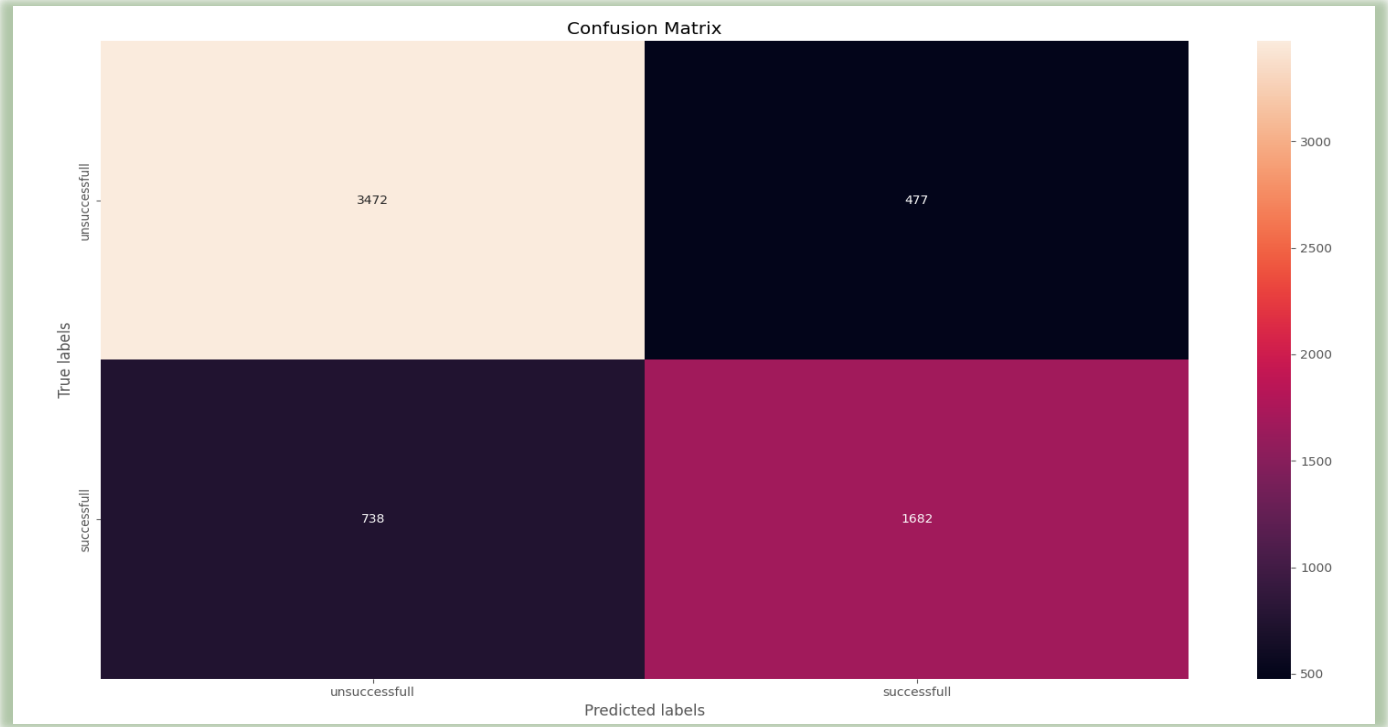
	precision	recall	f1-score	support
0	0.67	0.97	0.79	1690
1	0.80	0.22	0.35	1040
accuracy			0.68	2730
macro avg	0.74	0.59	0.57	2730
weighted avg	0.72	0.68	0.62	2730



precision	Precision is the fraction of true positive examples among the examples that the model classified as positive. In other words, the number of true positives divided by the number of false positives plus true positives.
recall	Recall, also known as sensitivity, is the fraction of examples classified as positive, among the total number of positive examples. In other words, the number of true positives divided by the number of true positives plus false negatives.

Matrice de Confusion – Modèle LR

- La matrice de confusion permet de visualiser les performances des modèles d'apprentissage automatique de classification. Elle donne une meilleure idée des performances du modèle : ici le modèle de type Logistique Régression. Pour notre modèle de classification de classe binaire : Lead Converti : True, Lead non converti : False (0 ou 1)



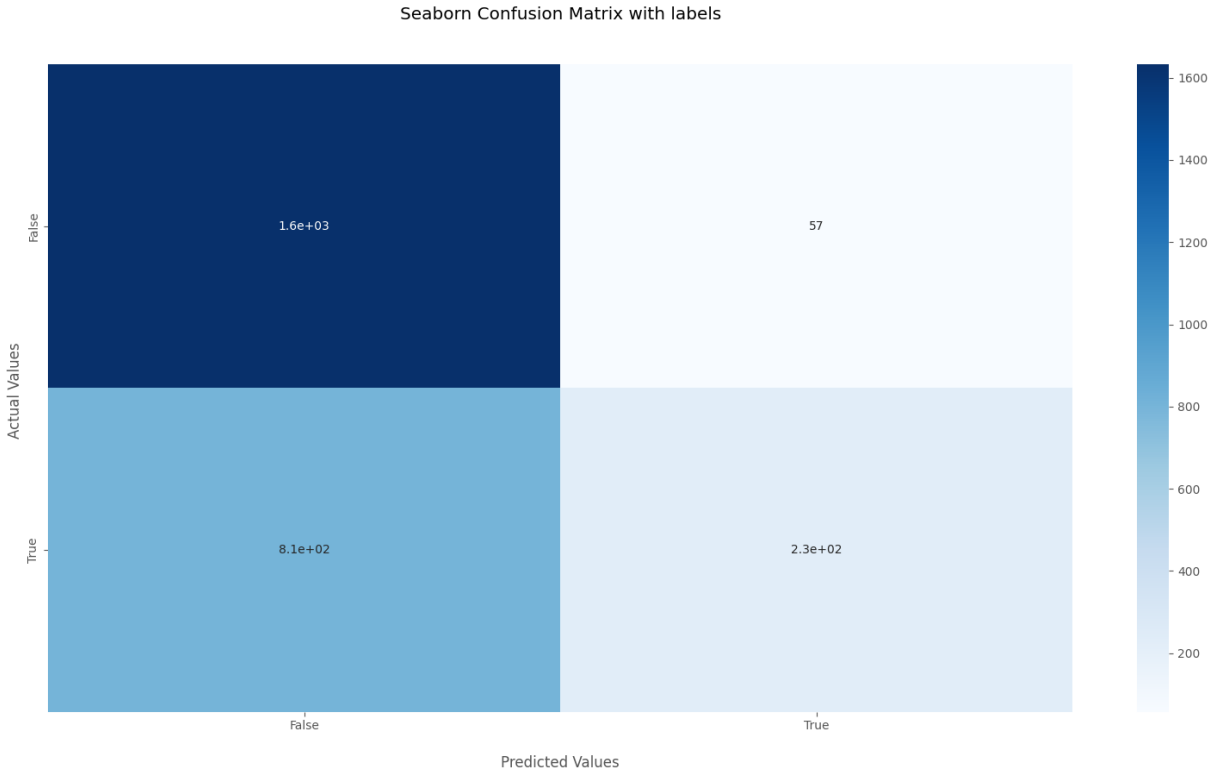
- Accuracy score: 0.8092322185586435
- Sensitivity score: 0.6950413223140496
- f1-score: 0.7346582223192837
- Precision score: 0.779064381658175
- Recall score: 0.6950413223140496….

Matrice de Confusion – Modèle SVM

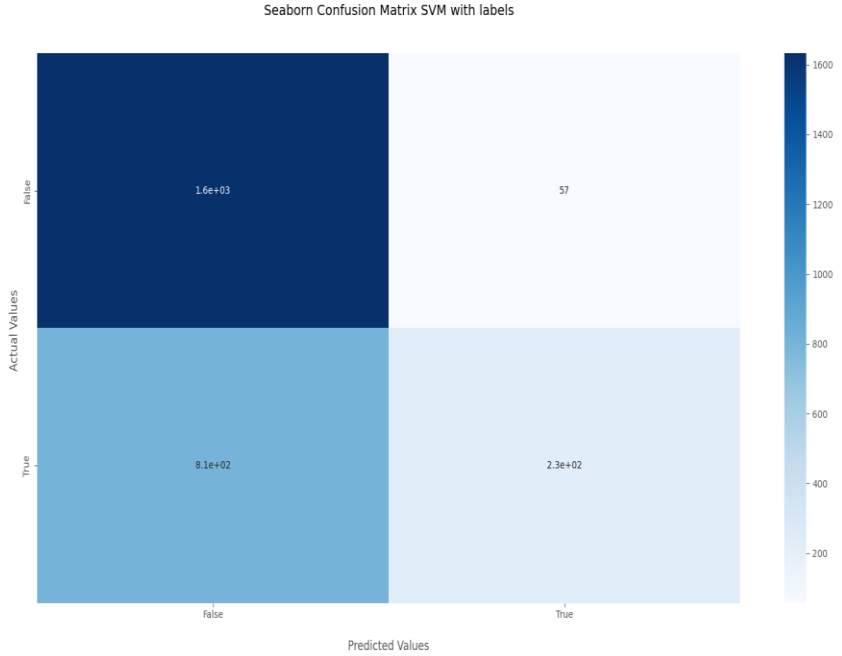
- La matrice de confusion permet de visualiser les performances des modèles d'apprentissage automatique de classification. Elle donne une meilleure idée des performances du modèle : ici le modèle de type SVM. Pour notre modèle de classification de classe binaire : Lead Converti : True, Lead non converti : False (0 ou 1)

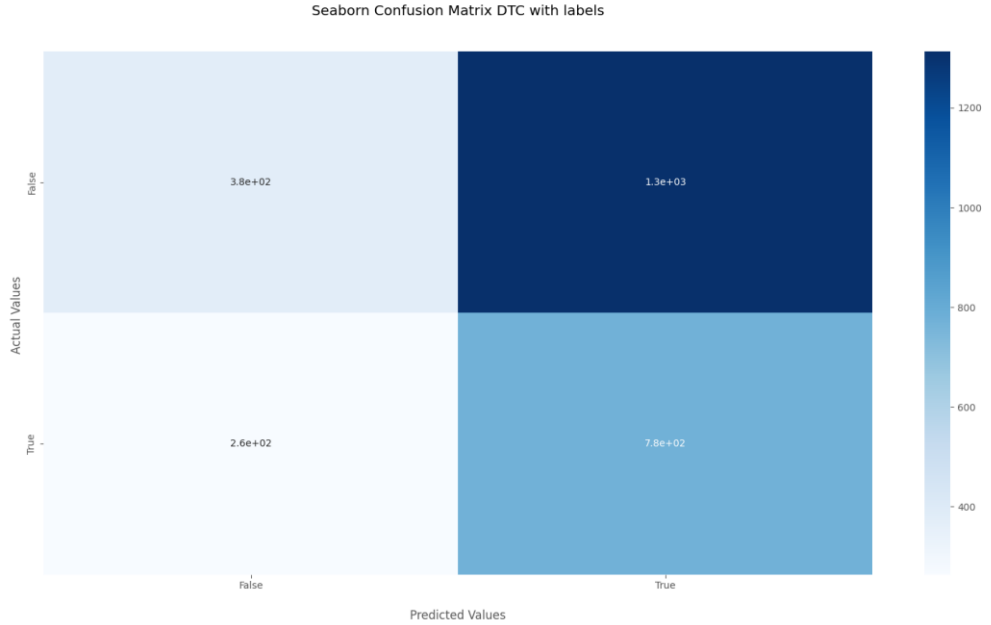
```
print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.67	0.97	0.79	1690
1	0.80	0.22	0.35	1040
accuracy			0.68	2730
macro avg	0.74	0.59	0.57	2730
weighted avg	0.72	0.68	0.62	2730



Matrice de Confusion – SVM & DTC





Leads – Dashboard - Modélisation

Feature Importance

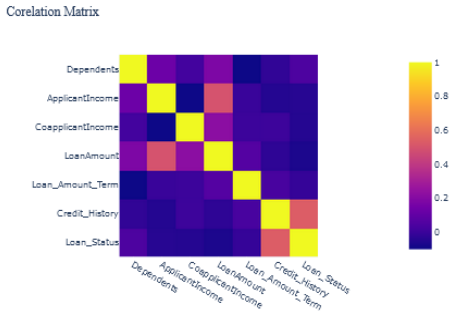
Correlation Matrix

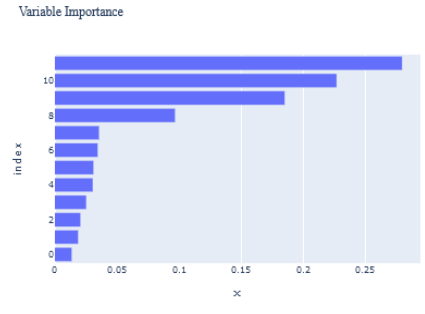
ROC Curve

Score F1, Accuracy Table

Algo Compare Perf

Confusion Matrix for best Model

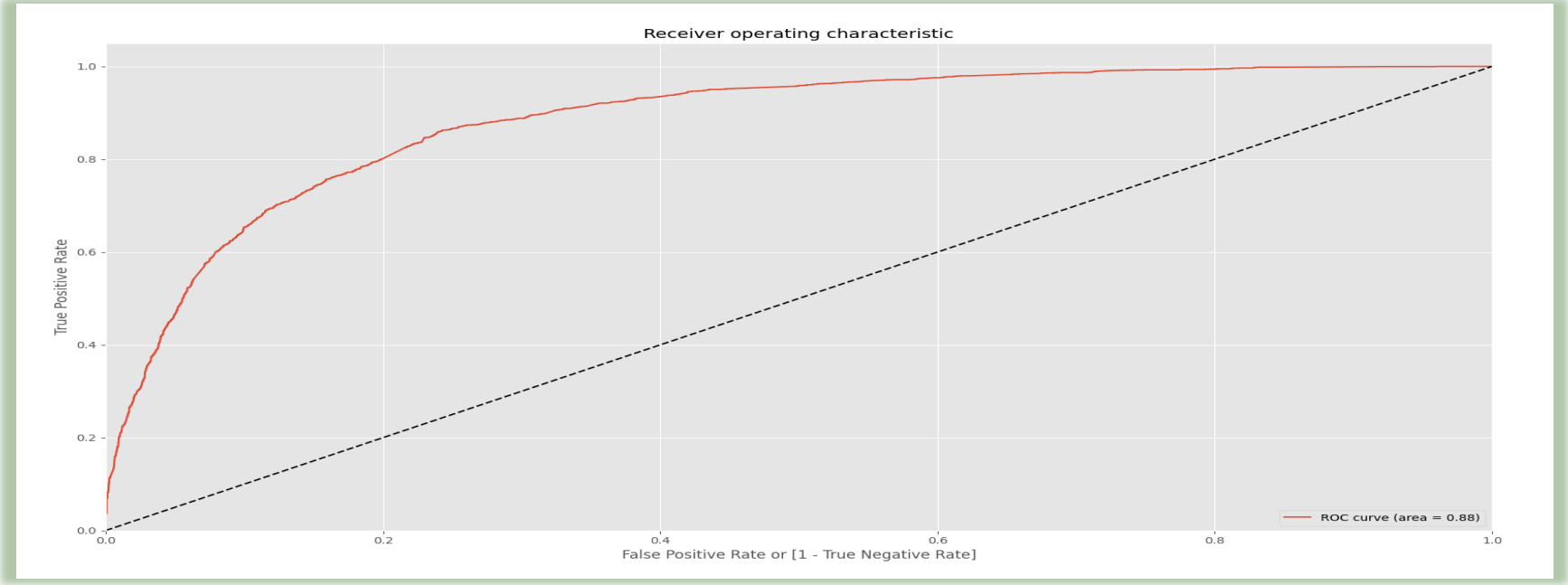




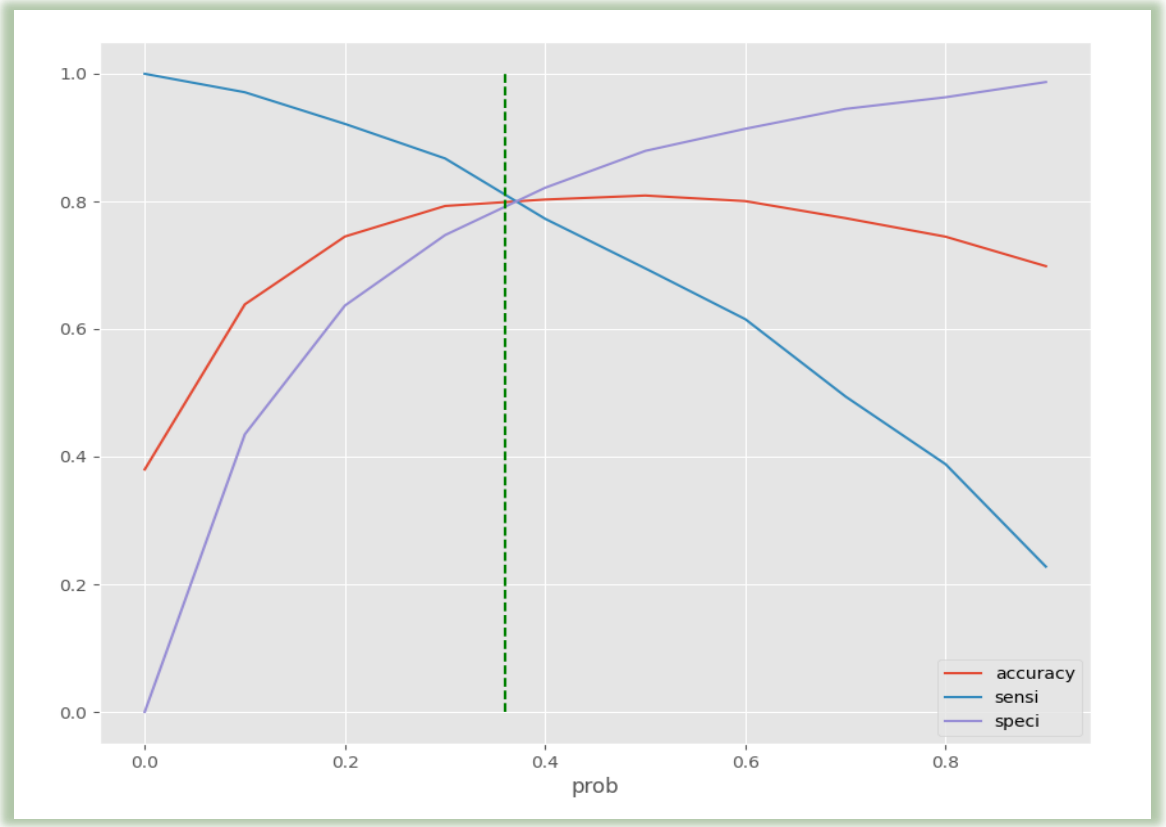
Courbe ROC

- Courbe ROC (Receiver Operating Characteristic) représente l'évolution du taux du vrai positif (TPR) en fonction du taux du faux positif (FPR) en faisant varier un seuillage sur la confiance (probabilité) qu'un exemple soit dans la classe positive. Pour évaluer globalement la performance d'un modèle, on calcule l'aire sous la courbe Precision-Recall, nommée **AUC Precision-Recall**.

```
from sklearn.metrics import roc_auc_score, roc_curve, precision_recall_curve
```



Seuil de Robustesse « Accuracy » - Seuil de coupure



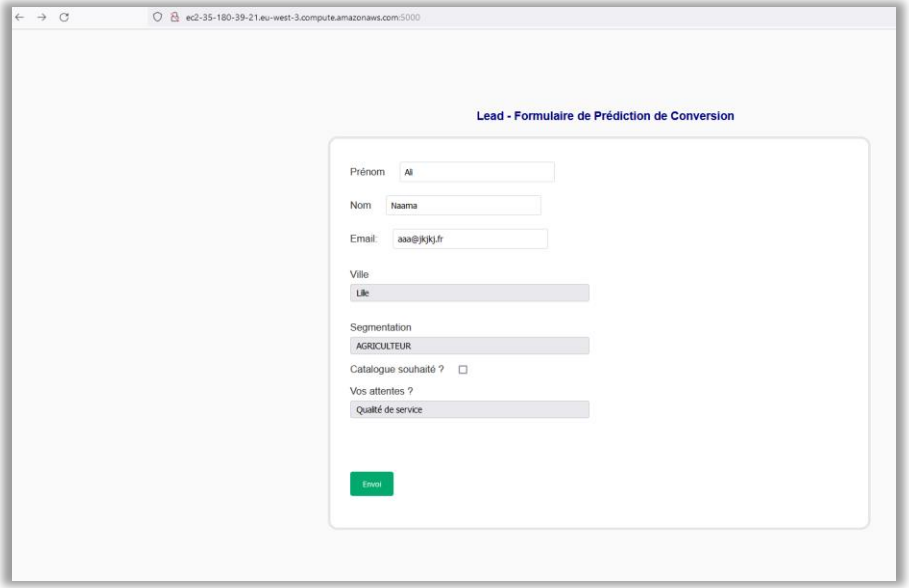
V. Résultats et recommandations

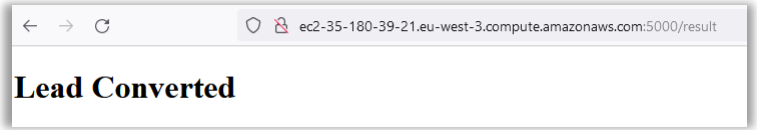
Interprétation « Lead Scoring»

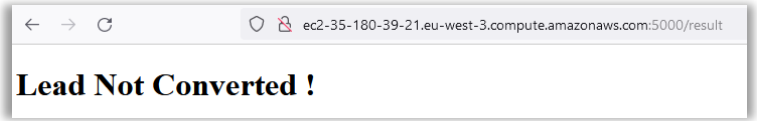
- L'origine du prospect, la dernière activité notable et la source du prospect sont les **principaux prédicteurs du modèle** de conversion des Leads
- **Ciblage des segmentation Divers à prévoir : Actions Emailing ou d'appel à prévoir**
- **En commercialisant davantage sur le site Web, les professionnels qui travaillent augmenteront les chances de conversion des prospects**
- **Le marketing avec Lead Source en tant que formulaire d'ajout rapide augmentera également les chances de taux de conversion**
- Avec un seuil de 0,6, nous avons une précision de test de 80,04 %, une sensibilité de 61 % et une spécificité de 92,3 %
- La probabilité de coupure doit être définie sur 0,35 pour le taux de conversion, c'est-à-dire que la sensibilité du modèle doit être de 80 %
- Pour une conversion élevé, le seuil doit être défini sur 0,3, ce qui augmente la sensibilité du modèle sans trop compromettre la précision du modèle

Test du modèle via un formulaire de captation de Lead

- Mise en œuvre d'un formulaire de captation de Lead permettant de tester le modèle.
- Principe, si l'on saisie les informations dans le formulaire, les données sont testées en temps réel par l'algorithme et l'on reçoit immédiatement une réponse quand à la conversion potentielle du Lead
- Cette application Python Flask, intègre le modèle de type DTC (Arbre de Décision) mis en place dans cette étude
- 2 cas résultats possibles :
  - Lead Converted
  - Lead Not Converted
- Déploiement de l'application sur une instance EC2 Amazon





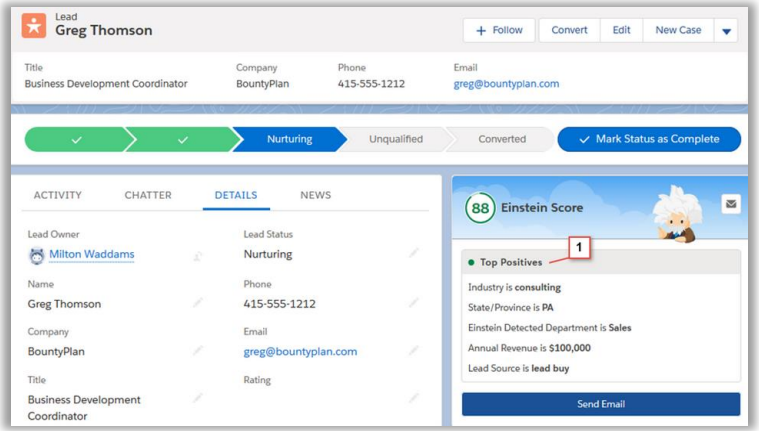


VI. Conclusion et Recommandations

Conclusion

- Cette étude permet de voir les différentes possibilités de modélisation des Scoring de Lead qui nous a permis de construire un tableau comparatif de solution
- Salesforce Sales Einstein, Scoring paramétrable, Scoring externalisé
- Le module de gestion de Lead doit être construit rigoureusement et inclure l'intégration avec différents canaux : Site Web, Outil Marketing, Email, Phone Call afin de suivre les différentes interactions
- Un ensemble d'automatisme sont à réaliser afin par exemple de notifier les commerciaux de nouveaux Leads à traiter. Ces Leads seront affectés automatiquement en fonction de critères géographiques (City) et permettront garder les Leads « chauds ».

Solutions	Avantages	Inconvénients
Acquisition du Module Salesforce Sales Einstein	Scoring Natif Refresh du modèle / 10 Jours	Coût d'acquisition Boite noire
Modèle de Scoring développé et externalisé	Coût d'acquisition	Maintenance et infrastructure prévoir une mise à jour mensuelle du modèle
Modèle de Scoring développé en spécifique et hébergé dans Salesforce sous forme de feature Weighting (Pondération)	Coût d'acquisition	Maintenance à prévoir / mise à jour mensuelle du modèle



Git

Merci de votre attention