Mise en place d’un modèle de pévision de la désactivation des entreprises du fichier de l’identifiant financier unique (IFU)

Ouedraogo G Fabrice

3 aoÃ»t 2024

## Introduction

* **Objectif :** Développer un modèle prédictif pour estimer la date de la première désactivation d’une entreprise dans le fichier IFU.
* **Importance :** Optimiser la gestion des entreprises, anticiper les risques financiers et améliorer la prise de décision des autorités fiscales.
* **Outil utilisé :** Une application Shiny sera développée pour visualiser et utiliser les prévisions du modèle.

## Revue de Littérature

* **Étude des modèles prédictifs existants :** Examen des techniques et des approches utilisées pour la prédiction de la désactivation des entreprises.
* **Identification des variables pertinentes :** Analyse des facteurs influençant la survie et la désactivation des entreprises dans divers contextes.

## Questions de Recherche

* Quels sont les principaux facteurs explicatifs de la désactivation des entreprises dans le fichier IFU ?
* Quel modèle prédictif offre la meilleure performance pour estimer la date de la première désactivation ?
* Comment les variables telles que l’âge de l’entreprise, le secteur d’activité et les redressements fiscaux influencent-elles la probabilité de désactivation ?

## Hypothèses

* L’âge de l’entreprise, le secteur d’activité et le nombre de déclarations fiscales sont des indicateurs significatifs de la désactivation des entreprises.
* La présence d’exonérations fiscales et le nombre de redressements subis augmentent la probabilité de désactivation.
* Les modèles de régression et les techniques de machine learning fournissent des prédictions plus précises que les modèles statistiques simples.

## Méthodologie

### Collecte de Données

* **Sources de données :**
  + Base de données du fichier IFU
  + Données fiscales et financières des entreprises
  + Bases de données publiques et administratives
* **Période d’étude :** Sélectionner une période pertinente pour obtenir des données historiques suffisantes.

### Préparation des Données

* **Nettoyage des données :** Traitement des valeurs manquantes, suppression des duplications, correction des anomalies.
* **Encodage des variables :** Utiliser des techniques appropriées pour les variables catégorielles (e.g., one-hot encoding).
* **Normalisation :** Standardiser les variables numériques pour une meilleure performance des modèles.

### Exploration des Données

* **Analyse descriptive :** Statistiques de base pour chaque variable, distribution des données.
* **Visualisation :** Graphiques et plots pour identifier les patterns et les corrélations.

### Modélisation

* **Sélection des modèles :**
  + Régression linéaire
  + Régression logistique
  + Arbres de décision
  + Forêts aléatoires
  + Gradient Boosting (e.g., XGBoost, LightGBM)
  + Modèles de survie (e.g., modèle de Cox)
* **Entraînement et validation :**
  + Division des données en ensembles d’entraînement (70-80%) et de test (20-30%)
  + Utilisation de la validation croisée (k-fold cross-validation)

### Évaluation des Modèles

* **Métriques de performance :**
  + RMSE (Root Mean Squared Error) pour les modèles de régression
  + Précision, rappel, F1-score, et AUC-ROC pour les modèles de classification
* **Comparaison des modèles :** Sélection du modèle offrant la meilleure performance selon les métriques choisies.

### Optimisation des Modèles

* **Tuning des hyperparamètres :** Utilisation de techniques comme Grid Search ou Random Search.
* **Ensemble learning :** Combinaison de plusieurs modèles pour améliorer les performances globales.

### Développement de l’Application Shiny

* **Objectif :** Créer une application interactive permettant de visualiser les prédictions et d’explorer les données.
* **Fonctionnalités :**
  + Interface utilisateur pour saisir de nouvelles données d’entreprises
  + Visualisation des prédictions et des facteurs influençant les désactivations
  + Options pour ajuster les paramètres du modèle et observer les résultats
* **Technologies utilisées :** R avec le package Shiny, intégration possible avec des librairies de visualisation comme ggplot2.

## Résultats Attendus

* Identification des variables les plus influentes sur la désactivation des entreprises.
* Développement d’un modèle prédictif robuste et précis.
* Création d’une application Shiny fonctionnelle et utile pour les parties prenantes.
* Recommandations pour les autorités fiscales sur la gestion proactive des entreprises à risque.

## Discussion et Conclusion

* **Interprétation des résultats :** Analyse des facteurs influençant la désactivation des entreprises.
* **Implications pratiques :** Utilisation des résultats pour améliorer la gestion des risques et la prise de décision.
* **Limites de l’étude :** Contraintes des données, limites des modèles utilisés.
* **Recommandations pour les recherches futures :** Suggestions pour approfondir l’étude et explorer d’autres approches.

## Plan de Mise en Œuvre

* **Étapes du projet :** Détails des phases du projet avec un calendrier précis.
* **Ressources nécessaires :** Identification des ressources humaines, matérielles et financières requises.
* **Plan de communication :** Stratégies pour communiquer les résultats aux parties prenantes.