ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Python for Data Analysis

Polish companies bankruptcy data Data Set

- Axel Joly – Clément Jaccarino -

Problème

Plusieurs questions se posent à l'aide de ce dataset :

- Quels sont les indicateurs qu'une entreprise Polonaise risque la faillite ?

- Quels sont les indicateurs qu'une entreprise Polonaise est en bonne santé ?

- Pouvons-nous prédire la faillite d'une entreprise Polonaise à l'aide d'indicateurs clé ?

Traitement des données

Nous avons 5 dataset, 1 par année de 2007 à 2013.

Chaque dataset recense les données financières sur l'année et par la suite les entreprises ont été réévalué à la fin de l'étude afin de constater les sociétés qui avaient fait faillite ou non.

Pour chacune des entreprises nous avons 64 attributs financiers.

Pour faciliter le traitement des données et la réflexion que l'on peut avoir sur le dataset nous souhaitons commencer par réduire le nombre d'attribut avec lesquels nous allons travailler. Pour cela nous commençons par étudier la corrélation des attributs entre eux afin de conserver uniquement les plus pertinents.

Ensuite nous avons choisi de réaliser quelques graphiques en nuage de point pour visualiser à la fois les entreprises ayant fait faillite et les autres en fonction de deux paramètres.

Traitement des données

Les chiffres sur les corrélations entre les variables indépendantes et la variable cible nous permettent de comprendre que la prédiction de la faillite d'une entreprise polonaise dans les années à venir ne dépend d'une ou deux variables indépendantes mais de l'ensemble des 62 attributs. En effet, les faibles corrélations indiquent que modifier un attribut ne favorisera pas la bonne santé ou au contraire la mauvaise santé de l'entreprise et sa probable faillite. Par ailleurs, la métrique R^2 du modèle final choisi (régression logistique) nous confirme cela.

Avec les matrices de corrélations, ces ensembles de données et ces graphiques nous permettent de conclure qu'il n'y a pas vraiment d'indices fort au repérage d'une entreprise polonaise qui pourrait faire faillite dans un avenir proche.

Nous avons ensuite dédié une partie du notebook à la mise en place d'une comparaison de modèles de prédiction, à l'affinage de ces modèles avec des GridSearch et au choix final du modèle et la présentation de ses métriques.

Problème dataset

Nous avons rencontré quelques limitations liées au dataset.

Etant donné que nous avons uniquement des id pour chaque année, nous ne pouvons réaliser un suivi d’une unique entreprise sur les cinq ans de données.

Aussi, les attributs étaient assez abstraits pour des élèves n'ayant que peu d'attrait pour le monde de la finance et de la comptabilité.