École Nationale des Ponts et Chaussées

Fouille de graphes basée sur l'Open Data pour faire évoluer les services bancaires

Marwane HARIAT
Martin THUMMEL
Tong ZHAO
Victor MARCHAIS
Promotion 018

Encadrant : M. Phillipe LAURIER

18 mai 2017

Table des matières

1 Introduction

1.1 Remerciements

Nous souhaiterions remercier notre professeur encadrant Phillipe Laurier. Nos échanges téléphoniques réguliers avec lui nous ont été d'une très grande aide.

Nous exprimons également notre profonde gratitude à l'équipe du Crédit Agricole qui a accepté de réaliser ce projet avec nous. Nous tenons à remercier tout particulièrement Emmanuel Toulemonde-Lin pour sa bienveillance et sa pédagogie.

Enfin, nous souhaiterions remercier tous les professeurs de l'École des Ponts ParisTech qui ont contribué à l'avancement du projet. Nous remercions notamment Guillaume Obozinski pour ses précieux conseils.

1.2 Contexte

Pourquoi est-il important de pouvoir anticiper le nombre de défaillances d'entreprises ? En quoi est-ce une donnée pertinente pour les banques/assurances que de connaître le nombre de défaillances d'entreprises des prochains mois ? Pourquoi est-ce utile ?

De nos jours, tout le monde a entendu parler du "Machine Learning" et du "Big Data". Ces expressions sont très connotés. Si bien que peu de gens perçoivent l'utilité et le fonctionnement réel de ces notions.

Qu'est-ce donc que le Machine Learning ?

Tout d'abord, "Machine Learning" est traduit en français par "apprentissage automatique". Le machine learning s'inspire du processus d'apprentissage par l'exemple, des êtres-humains. Il entend donc reproduire le mimétisme de l'homme. L'idée est de créer un programme qui, pour agir, se réfère à une base de données d'exemples. Plus cette base de données est grande plus l'ensemble des situations fournies par les exemples est varié et meilleure est l'action. On voit donc, à la lumière de ce qui a été dit, que la notion de Big Data est très importante.

Les institutions bancaires ainsi que les assurances commencent tout juste à prendre la mesure de la puissance du machine learning. Il est vrai que les analyses prédicatives que l'apprentissage automatique leur permettrait de réaliser comblerait une énorme manque à gagner et réduirait considérablement leur dépendance au risque.

L'une des problématiques essentielles que l'on rencontre dans le domaine des banques et des assurances est le risque de contrepartie. C'est à dire le risque qu'un emprunteur ne puisse pas rembourser une partie du prêt contracté. C'est une thématique fondamentale compte tenu de l'explosion de l'offre de crédits (croissance de 1 500 000 de contrats entre 1999 et 2001) et des consommateurs de plus en plus enclin et nombreux à recourir à un crédit. Il est donc primordial de développer des techniques d'analyses de risques individualisées. L'une d'entre elle s'appelle le scoring bancaire.

Le scoring bancaire est une méthode consistant à attribuer un score à chaque personne souhaitant contracter un prêt. Il est censé refléter la fiabilité d'une personne. Ce score est conditionné par les

valeur prises par un ensemble de variables pouvant être à visée personnelle (situation économique, situation familiale, âge, adresse,...) et/ou à visée macro-économiques (taux de chômage, indices boursiers, ...). Ces variables ont des degrés d'importances différents.

Le machine learning intervient dans la sélection des variables pertinentes ainsi que dans l'arbitrage quant aux poids accordés à ces dernières. Pour ce faire, l'algorithme consulte une base de données réunissant l'ensemble des individus ayant contracté un prêt ainsi que leurs informations personnelles et les éventuels défauts de paiement.

1.3 Objectif

L'objectif de ce projet de département est de tirer profit d'une base de données libre d'accès afin d'établir un modèle mathématique prédisant le nombre de défaillances d'entreprises un certain nombre de mois à l'avance. Nous nous plaçons donc dans la perspective de la gestion du risque de crédit et de ses enjeux. Mais afin de pallier l'impossibilité d'obtenir des données personnelles sur des particuliers, nous appliquons la démarche du scoring bancaire aux entreprises, sur lesquelles nous avons des données.

La base de données que l'on utilise est fournie par l'INSEE. Elle regroupe, dans un tableau, un ensemble de variables économiques ("Nombre de demandeurs d'emplois", "Nombre de créations d'entreprises",...) et détaille leurs valeurs pour les 145 derniers mois.

Le modèle que l'on souhaite obtenir est une fonction dépendant des valeurs de cette base de données et retournant le nombre de défaillances d'entreprises prévues. Cette fonction doit être le résultat de l'apprentissage d'un algorithme bien choisi.

2 Intérêt économique

Pourquoi est-il utile d'un point de vue économique de pouvoir prédire le nombre de défaillances d'entreprises ?

Tout d'abord, on peut voir le "nombre de défaillances d'entreprises" comme une variable macro-économique. C'est en effet un agrégat réunissant l'ensemble des entreprises qui ont été contraintes de faire un dépôt de bilan. A cet égard, le nombre de défaillances d'entreprises, au même titre que le cours de la bourse ou le PIB, est un indicateur de santé économique pertinent. D'autant plus que le nombre de défaillances d'entreprises, une fois sectorisé, peut s'avérer être un véritable outil de diagnostic et permettre de localiser d'éventuelles sphères de l'industrie, en difficultés.

Ainsi, arriver à donner une prédiction fiable de ce nombre de défaillances d'entreprises améliorerait la transparence de l'information. Cela permettrait aux agents économiques d'avoir accès à plus de renseignements concernant les facteurs significatifs du marché et de prendre de meilleurs décisions.

En outre, anticiper des défaillances d'entreprises peut vraisemblablement permettre aux banques d'obtenir un meilleur jugement des entreprises auxquelles elles accordent des prêts.C'est donc une source potentielle de rentabilité conséquente pour les dites banques mais également pour les particuliers qui profite alors d'un système bancaire moins sujet au risque et plus fiable.

En effet, beaucoup de problèmes financiers sont le résultat de prêts accordés à des entreprises qui ont fait par la suite faillite et qui ont donc été incapables de rembourser. C'est en particulier ce qu'il s'est passé dans les années 1980-1990 et lors de la crise des subprimes de 2008. A cause de mauvaises appréciations quant à la capacité de certaines entreprises à rembourser leur prêt, des banques ont subi d'énorme pertes.

Ces catastrophes financières, très traumatisantes, ont poussé de nombreuses instituions bancaires à investir dans des travaux de détection précoce de défaillances d'entreprises. C'est le cas par exemple de La Banque de France, qui a ainsi mise en place un modèle de scoring applicable aux PME.

Enfin, citons le domaine du Capital d'investissement, plus communément appelé "Private Equity", qui peut assurément tirer profit d'une information concernant la prévision du nombre de défaillances d'entreprises. En effet, le but d'un fond d'investissement est d'investir dans des sociétés qu'il aura choisi judicieusement, dans le but d'obtenir ensuite un retour sur investissement. On voit donc que si la banque d'investissement est capable d'anticiper les défaillances de certaines entreprises, elle pourra alors réduire considérablement son risque de sélectionner des mauvaises sociétés et augmenter simultanément ses bénéfices.

3 Gestion du Projet

Nous avons rencontré l'équipe du Crédit Agricole dans leur locaux, ils nous ont présenté plusieurs problématiques possibles qui allaient de l'extraction de data jusqu'à la création de modèles d'apprentissage. Le sujet initial était "Amélioration des services bancaires par fouille de graphe", cependant il n'était pas possible pour le Crédit Agricole de nous transmettre leur données privées même anonymisés. Parmi les sujets proposés nous avons donc choisis la prédiction.

Notre encadrant au Crédit Agricole nous a fourni un script R permettant d'extraire des données de l'INSEE. Nous avons ainsi pu récupérer des données macro-économiques mensuelles sur la période 2005-20017. Il y avait finalement assez peu de données manquantes a l'exception de certaines séries qui n'était pas renseignées dès 2005, nous les avons ignorées. Cependant cela nous a laissé avec plus de 10000 variables ou features donnant des indications générales notamment des indicateurs de consommation, de production et d'import/export. Par ailleurs nous faisons confiance à l'INSEE quand à la fiabilité des données. Parmi toutes ces séries, certaines donnent le nombre de défaillances d'entreprise sur l'ensemble du territoire par domaine (agricole, industriel, etc..). Ce sont ces séries qui vont nous intéresser et que nous allons essayer de prédire.

La principale difficulté du sujet a été le faible nombre de réalisations et la grande quantité de variables. Après avoir travaillé sur le traitement et l'analyse des données de sortie, nous avons implémenté des méthodes de régressions. Cependant le sur-apprentissage étant trop important pour améliorer les résultats nous nous sommes axés sur des techniques de sélection de variables et de clustering. Enfin nous avons combiné nos différents prédicteurs afin d'en créer un meilleur.

4 Technique

4.1 Traitement des données

Le premier aspect technique est l'extraction et le nettoyage des données. Dans le contexte de l'Open Data, nous voulons récupérer des bases de données libres et si possible de bonnes qualités. Les données de l'INSEE sont imparfaites, certaines séries ne commencent pas en 2005, d'autres se terminent avant 2017. Lorsqu'une valeur est manquante, elle est remplacée par un NAN (Not A Number) dans notre tableau.

Afin d'avoir une base de donnée utilisable nous faisons au préalable les opérations suivantes :

- Enlever toutes les variables dont la proportion de NA est supérieure à 5%.
- Enlever toutes les variables dont la variance est 0.
- Remplacer NA avec le nombre le plus fréquent de cette variable.
- Eliminer les doublons
- Séparer les variables de features et les variables à prédire.
- Normaliser tous les variables de features. La formule est :

$$\hat{y_i} = \frac{y_i - \min(Y)}{\max(Y) - \min(Y)}$$

Les variables sont régularisées (placées entre 0 et 1) pour être comparables entre elles. C'est cette régularisation qui a donné les meilleures résultats au début du projet nous l'avons ensuite conservé. Après ces traitements, il nous reste 10743 features contre 16523 au début.
Nous avons séparé notre ensemble de résultats en deux, un ensemble de test et un ensemble d'apprentissage (70/30 en proportion). L'ensemble de test est la fin de la série car l'objectif est de prévoir les futures séries. Ce modèle connaîtra toujours le passé, cette méthode nous a paru la plus adapté bien qu'elle implique un certain biais.

4.2 Analyse des Données de sortie

Nous avons chosis 12 variables à prédire : les *Nombre de défaillances d'entreprises par date de jugement - Données CVS-CJO - France* dans 12 catégories :

1. Agriculture, sylviculture et pêche
2. Industrie
3. Construction
4. Commerce et réparation automobile
5. Transports et entreposage
6. Hébergement et restauration
7. Information et communication"
8. Activités financières et d'assurance
9. Activités immobilières
10. Soutien aux entreprises
11. Eseignement, santé humaine, action sociale et services aux ménages
12. Tous secteurs d'activité

Bien que les données semblent fiables, le domaine "Tous secteur d'activité" ne correspond pas à la somme de tous les autres. Il est très probable qu'une entreprise appartenant à plusieurs secteurs soit comptabilisé dans chaque catégorie et qu'il existe des entreprises n'appartenant à aucune des onze catégories.
En traçant les histogrammes des variables de sortie, on remarque qu'il est peu probable qu'elles suivent une loi normale. Un test du khi2 nous confirme ces hypothèses.

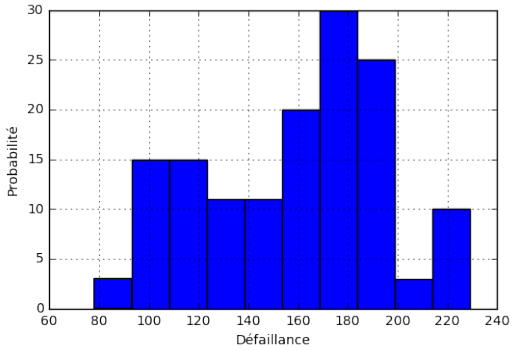

FIGURE 1 – Histogramme des Défaillances d'Activités financières et d'assurance

Enfin les variables sont peu corrélées entre elles, cependant on remarque qu'elles possèdent toutes un pic vers 2008 probablement dû à la crise, les valeurs redescenden ensuite sans se stabiliser.
On notes :

- T_e l'ensemble des mois d'entraînement.
- T_t l'ensemble des mois de test.
- Y_i le nombre de défaillance de la catégorie i
- X^t l'ensemble des variables connues au temps t

Comment déterminer l'efficacité d'un prédicteur P ?
On note l'erreur sur un ensemble T, la différence entre notre prédicteur et la réalité[5] :

$$E(P,i) = \frac{1}{|T|} \sum_{t \in T} |P(X_i^t) - Y_i^{t+2}|$$

La norme L1 semble adaptée car elle Un bon prédicteur doit avoir :

- une erreur E_e sur l'ensemble d'apprentissage faible
- une erreur E_t sur l'ensemble de test faible
- E_e et E_t assez proche (moins de 5)
- Un prédicteur est à rejeter absolument pour un label i l'approximation par la moyenne donne un meilleur résultat.

4.3 Sélections de variables

Il existe de nombreuses méthodes de sélections de variables[1]. Nous avons particulièrement regardé les méthodes par filtre. Ce sont les méthodes les plus simples à mettre en place. En effet elles consistent à associer un score à chaque feature (ou variables d'entrée) qui permet ensuite de les classer pour ne sélectionner que celles qui ont le meilleur score.

La plupart des techniques que nous avons utilisées pour calculer le score reposaient sur la corrélation linéaire entre la feature et le label. Puis nous avons mis en place une méthode pour regrouper les features en différents groupes grâce à des critères sémantiques ; pour ensuite ne sélectionner que quelques variables par groupe.

Première méthode de filtre Le nombre impressionnant de features nous force à trouver une méthode pour sélectionner un nombre réduits de variables avant même l'application de notre modèle. En effet si l'on sélectionnait toutes les variables, une interpolation grossièrement donne facilement un résultat parfait sur l'entraînement et un résultat très mauvais sur le test.
Dans un premier temps nous prenons, nous prenons les 10 variables les plus corrélées avec notre sortie ou en utilisant le gain d'information progressifs défini par :

$$I(X,Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} p(x,y) \log(\frac{p(x,y)}{p(x)p(y)}) dxdy$$

où $p(x,y)$, $p(x)$, $p(y)$ sont respectivement les densités des lois de (X,Y), X, Y.

Nous avons aussi mis en place des méthodes de filtres qui utilisent simplement la corrélation linéaire. Pour espérer obtenir de meilleurs résultats, nous avons essayé de mieux prendre en compte l'aspect temporel de nos séries. Nous avons alors partitionné l'ensemble d'apprentissage dans le temps. Nous obtenons des sous-ensembles, qui correspondent chacun à une période de temps continue incluse dans la période d'apprentissage. Pour chaque sous-ensemble, nous calculons alors la corrélation entre chaque feature et le label restreints à la période correspondante. Finalement on sélectionne les features qui se retrouvent souvent fortement corrélées avec le label. De plus nous avons essayé de donner un poids plus important aux sous-ensemble les plus récents dans l'ensemble d'apprentissage. En effet comme nous voulons prédire le futur, les parties les plus récentes de l'ensemble d'apprentissage sont susceptibles de donner de l'information qui a plus de valeur. Nous obtient comme score pour une feature X :

$$Score(X) = \sum_{i=1}^{n} iCorr(X^i, Y^i)$$

où X^i et Y^i sont la feature et le label restreint au sous-ensemble $T_i \subset T_e$

Clustering

Les méthodes par filtre sont rapides mais ont un gros défaut : les variables sélectionnées ne tiennent pas compte les unes des autres. En effet on peut sélectionner 10 variables très corrélées ce qui donne peu d'information. Pour pallier à ce problème nous effectuons un clustering des variables en les triant en plusieurs catégories selon un critère. Comme par exemple leur gain d'information ou leur corrélation.
Cependant cette méthode étant quadratique, elle est bien trop lente pour 10000 variables. Nous sélectionnons les 1000 meilleurs selon une méthode de filtre puis nous effectuons le clustering selon un norme L2 du vecteur de covariance ou de gain d'information mutuel.

Resultats Après sélection les variables semblent trop se ressembler nominalement, nous avons trop de de variables "exportation" avec des pays sûrement incohérant comme "Les îles Marshall".

4.4 Trie des noms

Nous souhaitons dans chaque catégorie (exportation, indicateur commerciale etc) quelle est la variable la plus explicative. Nous proposons de regrouper les features par leurs noms et de n'en choisir que une ou deux par groupes. Nous espérons ainsi accéder à plus d'informations avec le même nombres de variables. Pour cela nous implémentons le système suivant :

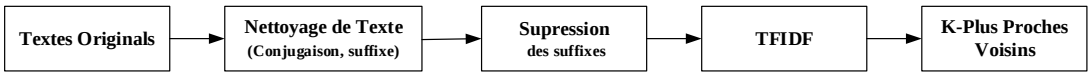

FIGURE 2 – Construction de système

Nettoyage de Texte Le but du nettoyage de texte est de séparer les noms des labels en mots importants. Pendant ce processus, on enlève les espaces, les chiffres et les ponctuations. De plus, on remplace toutes les lettres majuscules par des lettres minuscules.

Suppression des suffixes La suppression des suffixes est importante pour la classification. Pendant ce processus, on enlève les suffixes de chaque mots. C'est-à-dire que on ignore les conjugaison, les accords, etc. On ne garde que la racine de chaque mot. Après, on élimine les racines dont la longueur est inférieur à 3. On garde finalement 6026 racines. Ci-dessous on liste plusieurs racines pour vérifier l'effet du traitement.

TABLE 1 – Les racines les plus fréquents

Mot	Nombre	Mot	Nombre	Mot	Nombre
indic	13953	rev	7977	brut	5593
bas	12222	prix	7126	chiffr	5172
ser	9933	post	5968	affair	5170

TFIDF[7] Le TF-IDF(de l'anglais *Term Frequency-Inverse Document Frequency*) est une méthode statistique qui permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le score est composé de 2 termes :

— TF (Fréquence du terme) :

$$tf_{i,j} = \begin{cases} 1, & \text{if } t_i \text{appears in } d_j \\ 0, & \text{otherwise} \end{cases}$$

— IDF (Fréquence inverse de document) :

$$idf_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

, où $|D|$ est le nombre total de documents dans le corpus, et $|\{d_j : t_i \in d_j\}|$ le nombre de documents où le terme t_i apparaît dans tous les textes.

Le score final $tfidf_{i,j} = tf_{i,j} \dot{} idf_i$. Dans ce modèle, nous ne gardons que 1000 mots.

K-Moyennes K-moyennes est une méthode de clustering. Étant données n points $(x_1, x_2, \ldots, x_n)$ en dimension m, il divise l'ensemble en k groupes $S = \{S_1, S_2, \ldots, S_k\}$. $m = 1000$ ici, on utilise la distance d'euclidienne comme critère. Notre fonction d'objectif est :

$$\underset{S}{argmin} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

En variant k, on obtient la courbe suivante :

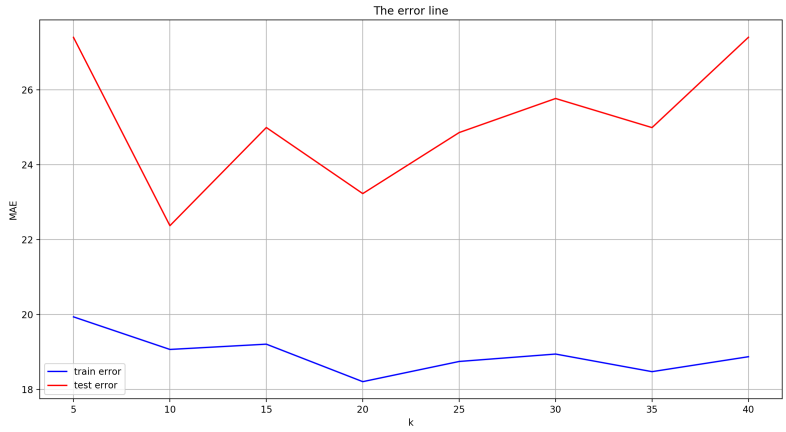

FIGURE 3 – L'évaluation de l'erreur du modèle en fonction du k

Finalement on conserve 10 groupes pour les test suivants. Pour vérifier nos résultats, on affiche plusieurs noms pour chaque groupes.

TABLE 2 – Les features choisis

Groupe	Nom
1	Indice de prix de production de l'industrie française pour le marché français - Prix de base - CPF 15.11 - Cuirs et peaux tannés et apprêtés, peaux apprêtées et teintées - Base 2010
2	Indice de chiffre d'affaires en valeur - Bibliothèques, archives, musées et autres activités culturelles (NAF rév. 2, niv. groupe poste 91.0) - Série CVS-CJO - Base 100 en 2010
3	Importations CAF de la France y c. DOM - Provenance : Espagne - Ensemble hors matériel militaire - Données estimées CVS-CJO - NAF rév. 2
4	Indice des prix à la consommation - Base 2015 - Ensemble des ménages - France métropolitaine - Nomenclature Coicop : 04.3.2.3 - Services d'entretien pour les systèmes de chauffage
5	Indice du commerce extérieur - Indice de prix des exportations - Toutes zones - CPF 10.7 - Produits de boulangerie, pâtisserie et pâtes alimentaires - Référence 100 en 2005
6	Indice de chiffre d'affaires en valeur - Base 100 en 2010 - Marché intérieur et export - Travaux de finition (NAF rév. 2, niveau groupe, poste 43.3) - Série brute
7	Indice CVS-CJO de la production industrielle (base 100 en 2010) - Construction de véhicules automobiles (NAF rév. 2, niveau groupe, poste 29.1)
8	Enquête mensuelle de conjoncture dans les services - Tendance prévue de la demande - Information et communication - Série CVS
9	Indice mensuel brut des prix d'achat des moyens de production agricole (IPAMPA) - Base 100 en 2010 - Aliments pour poulet label finition
10	Indice de chiffre d'affaires en volume - Commerce de détail en magasin non spécialisé (NAF rév. 2, niv. groupe poste 47.1) - Série CVS-CJO - Base 100 en 2010

4.5 Regression

Régression linéaire

Le méthode de régression le plus simple que nous avons utilisé est la régression linéaire. La fonction objectif est la suivante :

$$f(x, y, w) = \|y - Xw\|_2^2$$

Nous cherchons à trouver la meilleure combinaison linéaire de features (coefficients donnés par le vecteur w) pour expliquer le label. Cette méthode est assez générale et est capable de donner des résultats parfois satisfaisant même lorsqu'on a peu d'hypothèses sur les features et le label. Cependant comme nous disposons finalement d'un nombre assez réduit d'observations, le risque de sur-apprentissage est très important. Il nous est donc apparu très vite qu'il serait nécessaire d'utiliser des paramètres de régularisation pour gérer le sur-apprentissage.

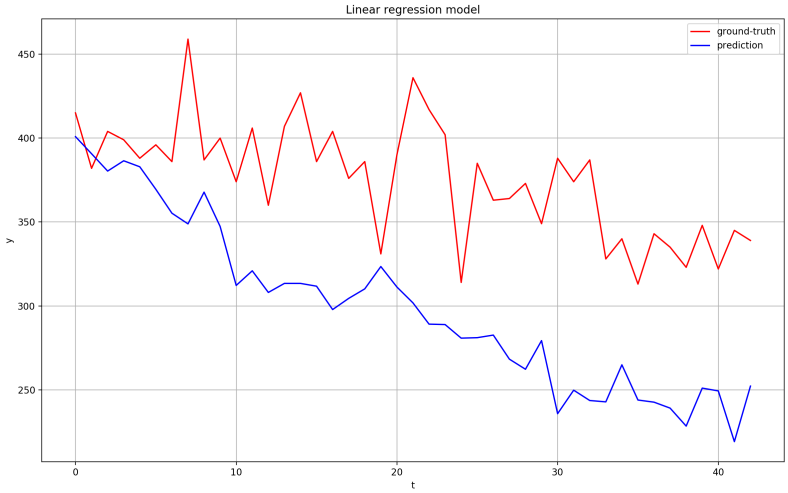

FIGURE 4 – Régression linéaire | Erreur associée = 34.00

La comparaison entre la prédiction et les valeurs de test montrent ici l'échec de la simple régression linéaire a bien expliquer le label.

Lasso[4]

La méthode Lasso nous a semblé une méthode intéressante pour ajouter de la régularisation dans notre modèle. La fonction objectif du Lasso est :

$$f(x, y, w) = \|y - Xw\|_2^2 + \lambda_1 \|w\|_1$$

Le terme $\|w\|_1$ permet de contrôler la norme 1 du vecteur de paramètres. En effet le sur-apprentissage va se caractériser par l'apparition de coefficients très élevés pour forcer l'adaptation du modèle sur l'ensemble d'apprentissage. L'intérêt du Lasso est d'utiliser la norme L1. En effet la minimisation aura pour conséquence de mettre à 0 les coefficients de certaines features pour contrôler la norme de w. Ainsi le Lasso permet de régulariser en complétant la sélection de variables.

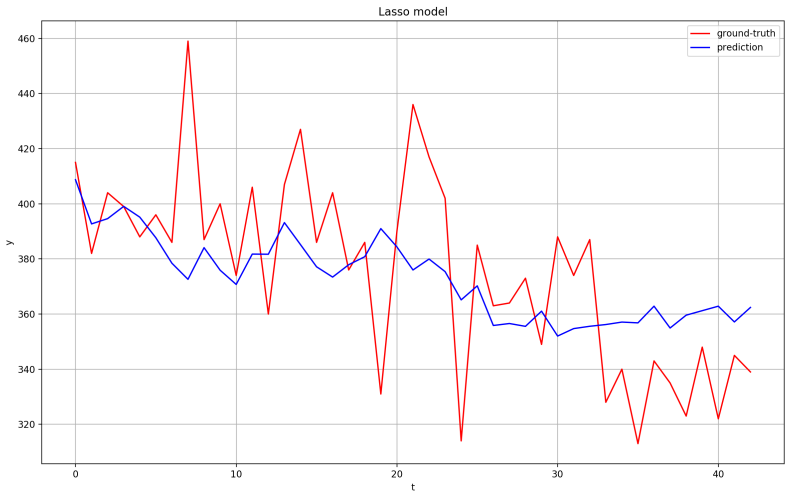

FIGURE 5 – Lasso | Erreur associée = 22.274

Ici on remarque que le modèle semble capable de suivre la tendance générale du label. Cependant certaines variations importantes et rapides dans le temps ne semblent pas pouvoir être prévues.

Elastic Net[8]

Elastic Net est un modèle efficace de régression linéaire. Il combine la perte de la norme L1(LASSO) et celle de la norme L2(RIDGE). La fonction de l'objectif est comme suit :

$$f(x, y, w) = \frac{1}{2n}\|y - Xw\|_2^2 + \lambda_1\|w\|_1 + \lambda_2\|w\|^2$$

Le terme L1 génère un modèle spacial. Le terme L2 stabilise le chemin de la régularisation de la terme L1.

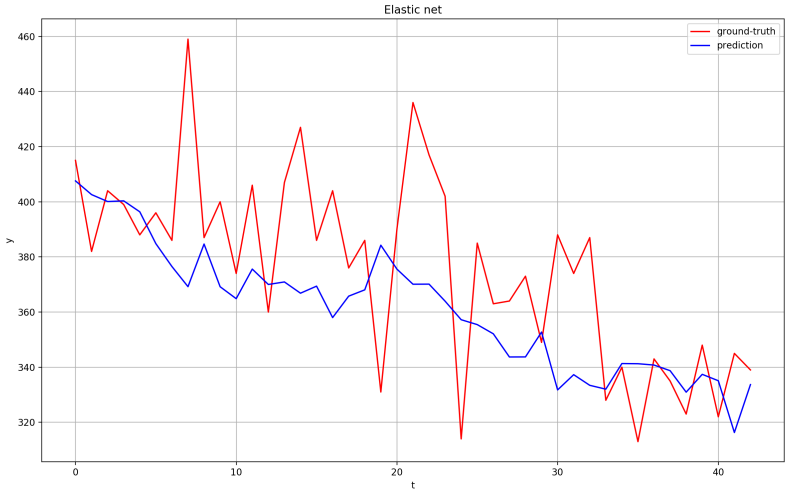

FIGURE 6 – Elastic Net | Erreur associée = 23.499

Les résultats donnés par le modèle Elastic Net sont assez proches de ceux donnés par la méthode du Lasso. Il n'y a pas de différences significatives. Ce qui montre que le Lasso est plutôt fiable malgré ce grand nombre de variables.

4.6 Time Series Model

Le grand nombre de variable étant plus un handicap qu'un avantage nous nous sommes tourné vers des méthodes sans sélection de variable qui utilise le critère temporelle de la série. Nous avons mis en place un *Time series model*[2][3], un modèle qui étudie et prédit l'évolution d'une série temporelle à partir de ses antécédents. L'intérêt est que l'on n'a plus besoin de choisir les features et qu'on fait la prédiction en considérant simplement le label. C'est un modèle auto-régressif.

Une série temporelle est un ensemble de points collectés en intervalle du temps constant. *Times Series Model* se différencie de la régression linéaire par les points suivants :

- Il dépend du temps. Donc l'hypothèse principal est que les différentes observations du label ne sont plus indépendantes.
- La plupart des séries temporelles ont une tendance liée avec les saisons.

Hypothèse Le modèle a une hypothèse : la série doit être stationnaire. Donc avant utiliser le modèle, on vérifie les propriétés suivants :

- La moyenne constante
- la variance constante
- un auto-covariance qui ne dépend pas du temps

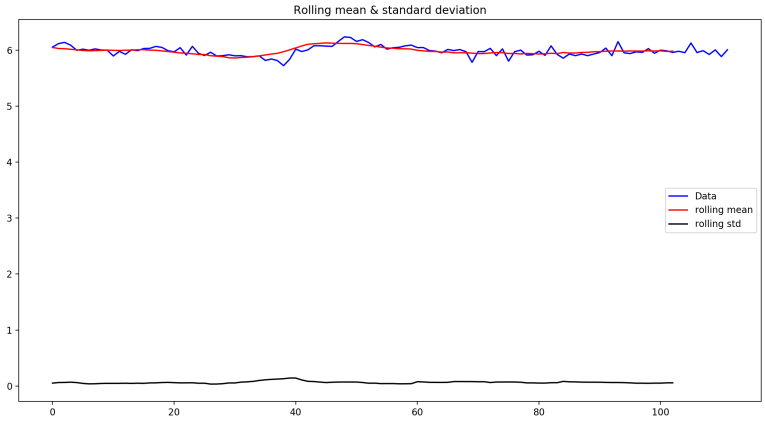

FIGURE 7 – Les courbes

On calcule la moyenne glissante et la variance glissante sur notre ensemble d'apprentissage. On observe bien que la moyenne et la variance de la variable cible sont constante dans le temps. Pour vérifier la dernière condition portant sur l'auto-covariance, on utilise le test de Dickey-Fuller, et il nous donne le résultat suivant :

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.478125
p-value                          0.008567
#Lags Used                       1.000000
Number of Observations Used    101.000000
Critical Value (5%)             -2.890611
Critical Value (1%)             -3.496818
Critical Value (10%)            -2.582277
dtype: float64
```

FIGURE 8 – Le teste

La p-valeur du test statistique est plus petite que 5%, la valeur critique.On peut donc dire, avec un niveau de confiance de 95%, que la série est temporelle.

Modèle - ARMA Le ARMA(de l'anglais *autoregressive moving average model*) est le principal modèle de séries temporelles. Dans cette partie, on travaille sur le logarithme de la variable cible.

Le modèle d'ordres (p, q) se sépare en deux parties :

— AR(p) - le modèle autorégressif

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t$$

, où $\phi_1, \phi_2, \ldots, \phi_p$ sont les paramètres du modèle, c une constante et ϵ_t un bruit blanc.

— MA(q) - le modèle moyen mobile

$$X_t = \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

, où $\theta_1, \theta_2, \ldots, \theta_q$ sont les paramètres du modèle et $\epsilon_t, \epsilon_{t-1}, \ldots$ sont des termes d'erreur.

Donc le modèle $ARMA(p, q)$ est

$$X_t = \epsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

Pour choisir les deux paramètres p et q, on trace deux courbes.

— ACF(La fonction d'autocorrélation) : le premier point qui traverse l'intervalle de confiance en haut sera le valeur de q.

— PACF(La fonction d'autocorrélation partielle) : le premier point qui traverse l'intervalle de confiance en haut sera le valeur de p.

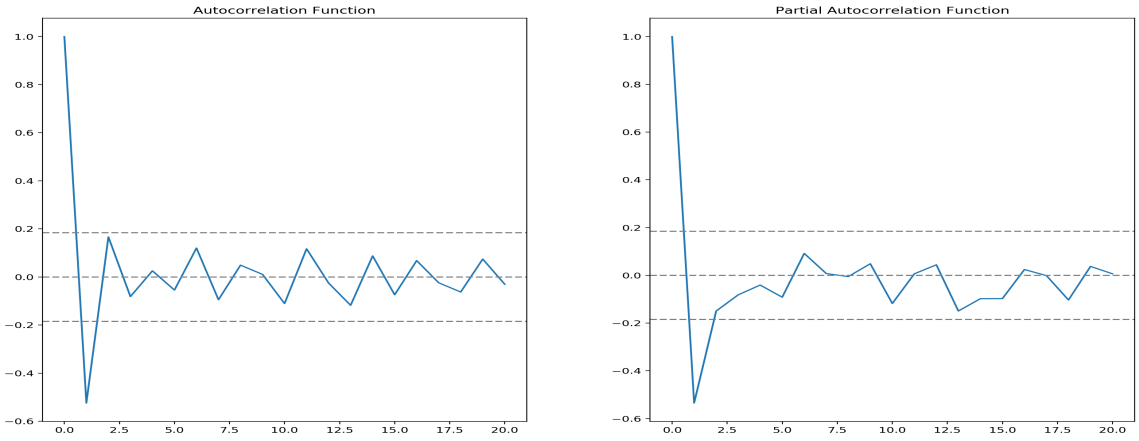

FIGURE 9 – Les courbes

Ainsi nous choisissons $p = 2$ et $q = 2$ comme les paramètres du modèle.

Résultat Nous utilise chaque deux mois de données pour prédire le mois suivant. C'est à dire pour prédire X_t, nous utilisons seulement $X_{t-1} et X_{t-2}$. Nous parcourons les valeurs 143 mois et nous obtenons l'erreur total $\epsilon = 22.776$

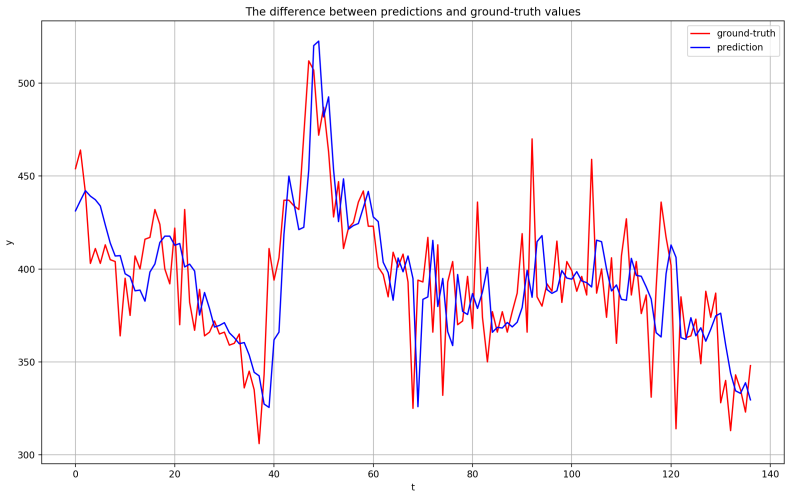

FIGURE 10 – Les résultats du time series model

4.7 Embedding

Nous avons créé un ensemble de prédicteurs plus ou moins efficaces.

- Time Series Model
- Différentes régression après un filtre sur les cluster de noms

La méthode du Embedding consiste à les combiner pour créer un meilleur prédicteur. Nous avons utilisé une méthode "Voting", il s'agit de faire une moyenne entre les différents prédicteurs. Cette moyenne n'est pas si anodine que cela. Il s'agit en réalité d'une extension du modèle de vote dans le cas de classification. Dans un modèle de classification, les prédicteurs renvoie 0 ou 1 si la sortie appartient à une certaine classe. Dans le modèle voting on fait voter chacun de nos prédicteurs et la majorité l'emporte. Ici de la même manière, on fait voter des prédicteurs sur la valeur.

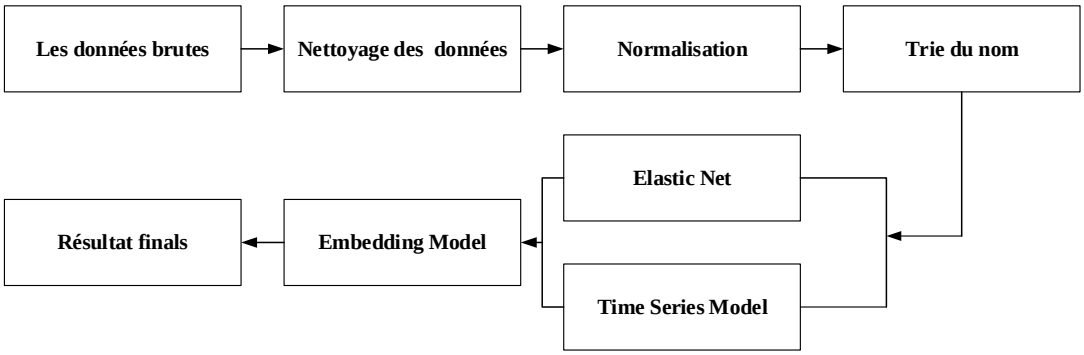

FIGURE 11 – Construction du modele finale

On obtient à la fin un modèle qui nous donne l'erreur $l = 20.384$ sur l'ensemble de validation ce qui reste notre meilleur résultat avec cependant un peu de sur-apprentissage (une erreur moyenne de 18 sur l'ensemble d'entraînement)

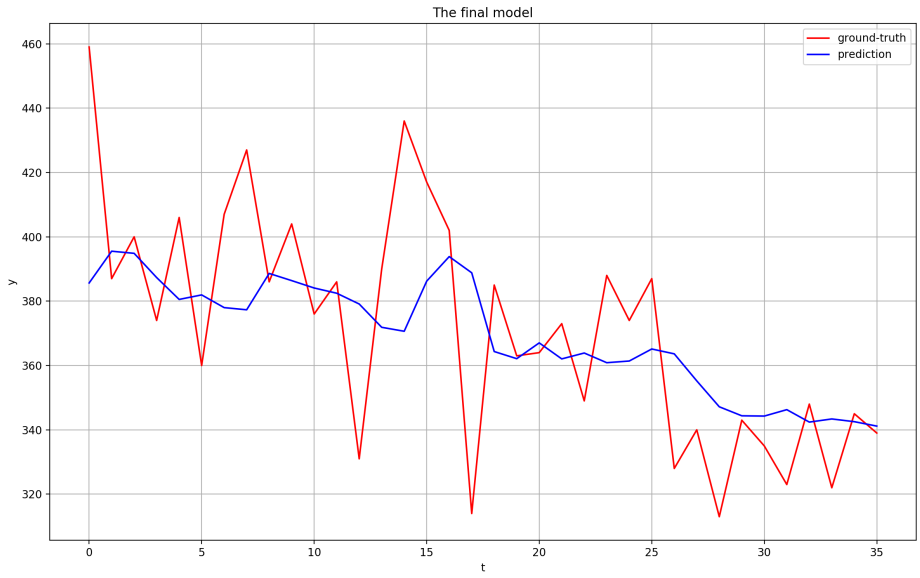

FIGURE 12 – Résultat final

5 Conclusion

Notre professeur encadrant, M. Laurier, a eu la gentillesse de parcourir l'ensemble de la base de données afin de sélectionner, à la main, les variables qu'il jugeaient susceptibles d'expliquer au mieux les défaillances d'entreprises. Ainsi, cette liste de variable, de part le moyen avec lequel elle a été construite, peut être considérée comme une sélection optimale vers laquelle notre algorithme pourrait tendre. M.Laurier a notamment insisté sur l'importance d'une variable : l'indice de la production industrielle dans le domaine de la construction de véhicules automobiles. Cette variable est choisie 4 fois sur 11 par les méthodes de sélection de Clustering de noms. Cela laisse présupposer que les méthodes de sélection implémentées sont efficaces.

Pour quantifier la pertinence de nos résultats, une manière naturelle de faire est de comparer les erreurs des différentes méthodes au prédicateur constant égal à la moyenne de l'ensemble d'entraînement. Pour les défaillances industrielle, l'erreur est 40.95. En revanche, grâce aux méthodes que nous avons implémentées, nous arrivons à obtenir une erreur de 20. Ce qui est assez satisfaisant, car grâce au machine learning nous arrivons à améliorer de moitié l'erreur initiale.

Cependant, remarquons que nous obtenons des erreurs voisines pour les différentes méthodes utilisées. Cela nous laisse penser que quelques facteurs exogènes rendent difficile un bon apprentissage. Tout d'abord, le fait de n'avoir que peu d'observations réduit grandement notre marge de manoeuvre. Ensuite, modéliser de façon linéaire est une simplification naturelle mais qui peut rapidement montrer ses limites. De plus, nous ne sommes même pas sûrs que les variables dont on dispose suffise à expliquer le nombre de défaillances d'entreprises.

Bibliographie

[1] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. Journal of machine learning research, 3(Mar) :1157–1182, 2003.

[2] James Douglas Hamilton. Time series analysis, volume 2. Princeton university press Princeton, 1994.

[3] William Wu-Shyong Wei. Time series analysis. Addison-Wesley publ Reading, 1994.

[4] Wikipedia. Lasso (statistics) — wikipedia, the free encyclopedia, 2017. [Online ; accessed 20-May-2017].

[5] Wikipedia. Mean absolute error — wikipedia, the free encyclopedia, 2017. [Online ; accessed 20-May-2017].

[6] Wikipedia. Mutual information — wikipedia, the free encyclopedia, 2017. [Online ; accessed 20-May-2017].

[7] Wikipedia. Tf–idf — wikipedia, the free encyclopedia, 2017. [Online ; accessed 20-May-2017].

[8] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 67(2) :301–320, 2005.