PROJET MACHINE LEARNING

MASTER IEF - 2020

SUJET 2 : IMPORTANCE DES MOTS

Paul GENEVRIER – Marie-Clémence CIUPEK

Caractéristiques de la base :

On dispose d'une base de données comportant les caractéristiques de plusieurs actions entre 2007 et 2017.

Ces caractéristiques sont les suivantes :

- OP, UP, DO, CL - Prix Open, High, Low et Close
- VO – Volume
- RDMT_x - rendement futur à 5 jours ouvrés
- HISTO_x - rendement historique de la valeur de l'indice à la clôture du marché
- VOL_x - écart historique des volumes d'échange
- UP_x - rendement historique de la valeur au plus haut en intraday
- DO_x - rendement historique de la valeur au plus bas en intraday

x = Suffixes J, S et M indiquant respectivement

- J = une période journalière (entre 2 jours ouvrés consécutifs)
- S = une période hebdomadaire
- M = une période mensuelle.

On y ajoute les caractéristiques suivantes (calculés à partir des données de la base) :

- FUTUR_VO_x - variation du volume futur pour chaque ticker
- MEDIAN_VO - médiane du volume écahangé pour chaque ticker
- 75_CENT_VO – quantile à 75% du volume écahangé pour chaque ticker

Ces nouvelles caractéristiques vont nous permettre de créer des variables à expliquer en les combinant au rendement des actions. Plusieurs variables ont été créées mais nous ne les retiendrons pas toutes.

La base comporte également les mots parus au sein des actualités boursières journalières rattachées au ticker de la même ligne.

La base comporte **106542 lignes** et **280 colonnes**.

Types de données des colonnes de la base brute :

Colonne	Type
TICKER	object
annee	int64
mois	int64
jour	int64
OP	float64
UP	float64
DO	float64
CL	float64
VO	float64
RDMT_x	int64
HISTO_x	float64
VOL_x	float64
UP_x	float64
DO_x	float64
mots	int64

Type de données de la base retraitée :

Colonne	Type
TICKER	object
date	datetime64[ns]
OP	float64
UP	float64
DO	float64
CL	float64
VO	float64
RDMT_x	int64
HISTO_x	float64
VOL_x	float64
UP_x	float64
DO_x	float64
FUTUR_VO_x	float64
MEDIAN_VO	float64
75_CENT_VO	float64
mots	int64

Nous allons zoomer sur les mots les plus fréquents des actualités boursières et créer un modèle de prédiction sur les rendements mensuels.

Quelques statistiques de la base :

Dans un premier temps, nous avons choisi de calculer des statistiques pour chaque ticker :
stats_by_ticker *(DataFrame)*

- Le rendement moyen mensuel de ce ticker : *stats_by_ticker["RDM_MOYEN_M"]*
- Le mot le plus cité pour le ticker : *stats_by_ticker["MOST_FREQ_WORD"]*
- Le nombre d'apparitions de ce mot : *stats_by_ticker["NB_WORD"]*

Puis au vu de l'objectif de ce projet, nous avons préféré calculer des statistiques plus en lien avec l'étude des modèles, pour cela nous avons calculer des statistiques par mot (tous tickers confondus) :
stats_by_word *(DataFrame)*

- Le nombre d'apparition de ce mot : *stats_by_word["APPARITIONS"]*
- Le rendement moyen lorsque ce mot est cité : *stats_by_word["RDM_MOYEN_M"]*
- La fréquence qu'une hausse du rendement (future par rapport à historique) survienne lorsque le mot est cité : *stats_by_word["HAUSSE_RDMT_M"]*
- La fréquence que le volume traité du jour soit supérieur à la médiane du volume (propre au ticker concerné) lorsque le mot est cité : *stats_by_word["VO>MEDIAN"]*
- Idem mais cette fois comparé au quantile 75% : *stats_by_word["VO>QUANTIL_75"]*
- La fréquence qu'une hausse du volume traité par rapport à la veille survienne lorsque le mot est cité : *stats_by_word["V0_HISTO_J>0"]*
- La fréquence qu'une hausse future du volume traité par rapport à demain survienne lorsque le mot est cité : *stats_by_word["V0_FUTUR_J>0"]*

Nous trions ce dataframe par '*APPARITIONS*' puis par '*RDM_MOYEN_M*'.

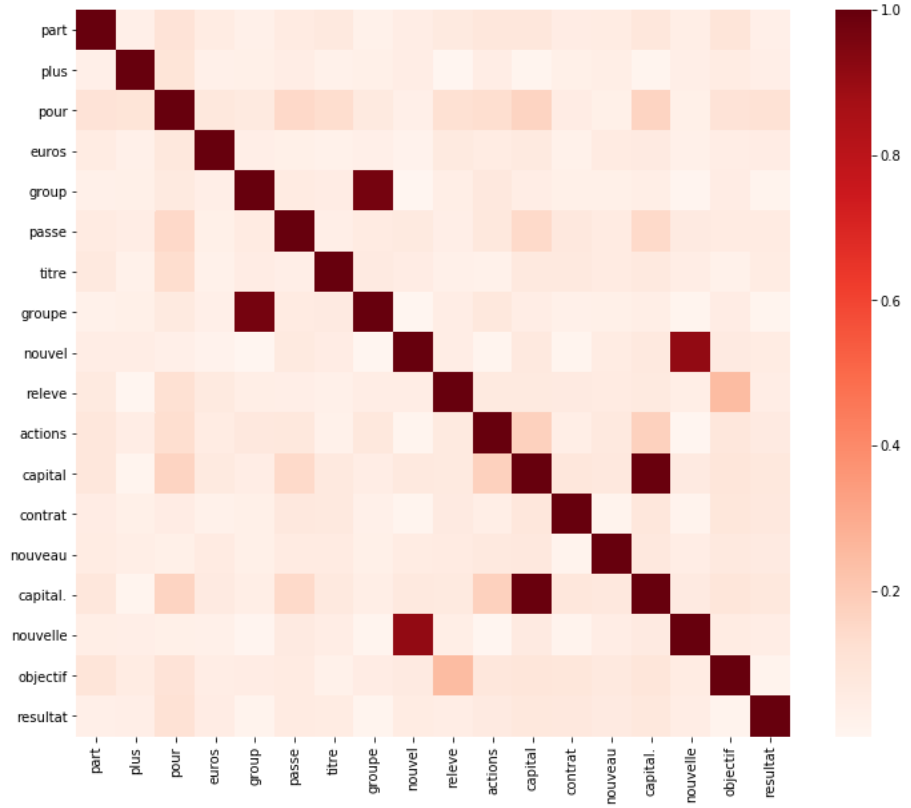
En voici les 20 premières lignes :

	APPARITIONS	RDM_MOYEN_M	HAUSSE_RDMT_M	VO>MEDIAN	VO>QUANTIL_75	VO_HISTO_J>0	VO_FUTUR_J>0
pour	1966	1.36%	48.27%	51.07%	26.81%	54.27%	46.74%
avec	1737	0.94%	47.09%	50.37%	24.99%	52.56%	44.67%
sous	1502	0.74%	52.93%	66.11%	39.55%	67.84%	33.95%
pres	994	0.85%	46.98%	64.79%	39.74%	62.78%	36.42%
dans	981	0.95%	49.44%	46.48%	20.80%	51.48%	46.28%
action	979	0.96%	48.01%	45.45%	21.65%	50.15%	49.03%
d_un	889	0.47%	47.81%	49.83%	25.08%	56.02%	48.37%
vers	807	0.73%	46.34%	72.37%	45.97%	72.00%	31.35%
achat	790	0.38%	48.23%	56.58%	31.01%	55.19%	41.01%
capital	761	1.20%	50.99%	48.75%	24.18%	50.46%	46.78%
capital.	751	1.10%	51.00%	48.20%	23.70%	50.47%	46.87%
part	716	1.19%	48.32%	52.51%	26.68%	56.01%	46.93%
objectif	636	1.47%	46.07%	73.58%	43.08%	62.58%	33.33%
actions	607	1.02%	48.76%	44.15%	20.10%	49.92%	47.78%
contrat	540	1.06%	48.70%	43.52%	17.96%	49.44%	51.48%
passe	527	1.38%	49.53%	49.53%	26.00%	57.50%	44.97%
nouveau	511	1.19%	47.16%	46.58%	21.72%	48.34%	48.14%
nouvel	494	1.15%	47.37%	46.15%	22.06%	49.19%	48.18%
group	493	1.29%	47.06%	50.51%	27.79%	51.72%	47.06%

Qq mots sur l'impact sur le volume et ou hausse rendemetn

<u>Retrait des variables trop corrélées :</u>

Matrice de corrélation des variables :



On retire les variables trop corrélées positivement ou négativement i.e. les variables dont la valeur absolue du coefficient de corrélation est supérieure à 75%. Cette filtration basée sur la corrélation

permet de retirer les mots en doublons ou étant très proches (nouveau/nouvel, capital/capital., group/groupe, …). Ces doublons risquent de fausser l'apprentissage.

A la recherche d'AUC :

Un AUC, oui, mais à quel prix ?

- Le choix de la variable à expliquer doit prendre en compte, qu'il faut que la variable ait un nombre suffisant de valeurs positive (y=1). Sinon, lorsque la majorité de la variable à expliquer est trop faible, l'optimisation sur l'AUC sera biaisée. En effet, dans le cas où seulement 1% des y sont positives, alors si le modèle prédit uniquement des valeurs négatives (y_pred=0), il aura une précision et un recall nuls, pourtant ne se trompe rarement car la grande majorité des variables à prédire sont nulles.
  - ⇨ Lors de notre projet cela nous est arrivé lors plusieurs essais de variables à expliquer, et nous obtenions des AUC entre 0.7 et 0.8

Le choix des métriques :

Dans un premier temps, nous nous sommes focalisés sur l'AUC, cependant comme expliqué au-dessus nous nous sommes rendu compte de l'importance de considérer le recall et la précision pour juger de l'efficacité du modèle.

Puisque le but du projet est de prédire un signal d'achat, nous avons souhaité favoriser la précision plutôt que le recall. Quitte à manquer des opportunités, nous préférons investir dans moins de de prévisions positives mais plus souvent bonnes.

Là encore nous avons remis en question ce choix, plusieurs modèles nous permettaient d'obtenir un AUC proche de 0.7 et une précision entre 0.65 et 0.75. Mais ces modèles donnaient un recall quasi nul (par exemple nous prédisions 5 valeurs « Vrai Positif » et 2 valeurs « Faux Positif »).
Nous ne jugeons pas optimal de laisser passer de nombreuses opportunités pour en jouer seulement 5 avec une précision de 0.7

Ainsi nos métriques d'optimisation sont l'AUC (de la courbe ROC) et le recall.

Pour l'hyperparamétrage, il faut choisir une parmi ces 2 métriques pour le *refit*. Nous avons testés les 2. Pour certains cas de notre étude prendre l'AUC comme *refit* entraîne un recall nul. On privilégiera donc le recall comme métric pour le *refit*.

Le choix de la variable à expliquer

1) Nous avons premièrement regardé la variable binaire *= if(rendements futurs mensuels > 2%)*
   Comme dit ci-dessus, nous nous focalisions seulement sur l'AUC au début, avec de nombreux hyperparamétrages sur différents hyperparamètres, l'AUC semblait avoir un maximum de 0.55.
2) Nous avons donc choisi de construire d'autres variables à expliquer et de les tester dans notre modèle à travers différents d'hyperparamétrages.
   Voici 2 autres variables binaires que nous avons regardées et analysées :
   - **Rendement futur supérieur à 2% et (Volume du jour > Quantile_Volume(0.75) ou Hausse future du volume journalier d'au moins 75%)**
     ➔ ici nous rajoutons une condition sur le volume, nous expliquons ce choix par le

fait que des actualités positives sur une entreprise provoque généralement un « rush » des investisseurs sur l'action de l'entreprise, entraînant généralement une hausse du volume traité.
Si la nouvelle apparaît au cours de la journée alors ce sera le volume de ce jour qui sera élevé (d'où la condition > au quantile(75%)), si elle apparaît après la fermeture du marché , ce sera le volume du lendemain qui sera impacté que celui de la veille (d'où hausse > 0.75).
Pour cela nous avons du calculer pour chaque ticker son quantile(75%) de volume quotidien, et à chaque ligne du ticker calculer la variation futur du volume à 1 jour.

Script : *y=filtered_data.RDMT_M.apply(lambda x : 1 if x>= 0.02 else 0)*\
((filtered_data.FUTUR_VO_J).apply(lambda x : 1 if x>0.75 else 0)+
(filtered_data.VO-filtered_data['75_CENT_VO']).apply(lambda x : 1 if x> 0 else 0)-
*((filtered_data.FUTUR_VO_J).apply(lambda x : 1 if x>0.75 else 0)**
(filtered_data.VO-filtered_data['75_CENT_VO']).apply(lambda x : 1 if x> 0 else 0)))

- **Rendement futur supérieur à 2% et (Rendement futur mensuel > Rendement historique mensuel)**
  ➔ ici nous rajoutons une condition sur la hausse du rendement, lorsque le cycle économique est propice aux actions, le rendement futur est plus une généralité sur de la dynamique de marché qu'une réaction à des actualités/nouvelles positives, ainsi on cherche les rendements positifs en hausse par rapport au mois dernier.

Script : *y=(filtered_data.RDMT_M-filtered_data.HISTO_M).apply(lambda x : 1 if x>= 0 else 0)*(filtered_data.RDMT_M.apply(lambda x : 1 if x>0.02 else 0))*

Nous avons regardé d'autres variables binaires selon les niveaux et variations du volume (Journalier, Hebdo, Mensuel), mais les niveaux d'AUC n'étaient pas satisfaisants donc nous les avons exclues pour la suite de notre étude. En voici quelques exemples :

*y=filtered_data.RDMT_M.apply(lambda x : 1 if x>= 0.02 else 0)**
(filtered_data.VO-filtered_data['MEDIAN']).apply(lambda x : 1 if x> 0 else 0)
*y=filtered_data.RDMT_M.apply(lambda x : 1 if x>= 0.02 else 0)**
(filtered_data.FUTUR_VO_M).apply(lambda x : 1 if x> 0.25 else 0)
*y=filtered_data.RDMT_M.apply(lambda x : 1 if x>= 0.02 else 0)**
(filtered_data.FUTUR_VO_S).apply(lambda x : 1 if x> 0.5 else 0)

3) Finalement après avoir effectué des hyperparamétrages pour chaque variable binaire, la plus consistante est la 3ème : **rendements futurs mensuels > 2% et rendements en hausse**.

Variable	AUC (ROC)	Recall	Precision
2	0.69	0.12	0.39
3	**0.66**	**0.31**	**0.47**

En choisissant la variable 3 on sacrifie 0.03 d'AUC pour multiplier par 3 notre recall et dans le même temps améliorer la précision.

<u>La méthode d'hyperparamétrage : Un hyperparamétrage en 2 temps</u>

1) Un modèle de départ :
   - Fonction objectif : « binary : logistic »
   - n_estimators : 200
   - learning_rate : 0.1
   ➔ on cherche à avoir une idée des paramètres, donc dans un premier temps on souhaite de la rapidité

2) Un 1er hyperparamétrage rapide : Optimisations successives sur les hyperparamètres
   a) Optimisation sur *max_depth* et *min_child_weigth*
      Si un des hyperparamètres et à la limite du range, on déplacer le range et recommencer en conservant le meilleur autre hyperparamètre
      ➔ On conserve le meilleur modèle
   b) Optimisation sur *gamma* sur le meilleur modèle obtenu en a)
      ➔ On conserve le meilleur modèle (Attention précision et/ou recall peuvent être nul, si refit='AUC')
   c) Optimisation sur *sub_sample* et *colsample_bytree* sur le meilleur modèle obtenu en b)
      ➔ On conserve le meilleur modèle
   d) Optimisation sur *sub_sample* et *colsample_bytree* sur le meilleur modèle obtenu en c)
      ➔ On conserve le meilleur modèle

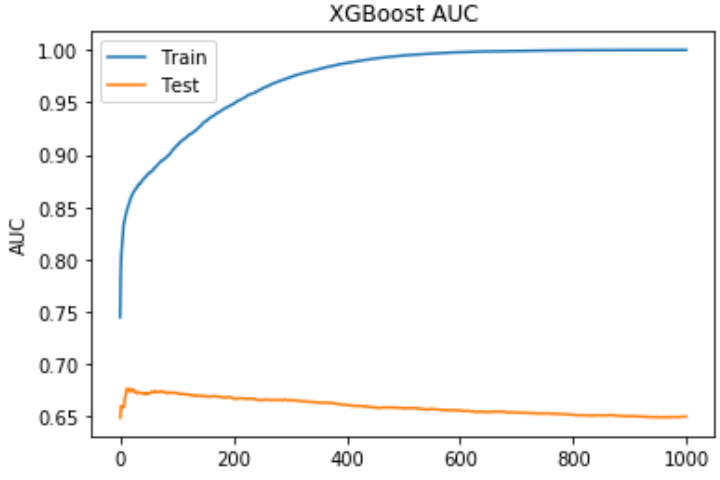
3) Un 2ème hyperparamétrage plus précis : Optimisation sur les hyperparamètres en simultanée
   A partir du modèle obtenu en d) on définit un grid élargi depuis les valeurs des hyperparamètre du modèle (d).
   On fixe :
   - n_estimators : 500
   - learning_rate : 0.02
   ➔ on cherche à être plus précis maintenant.

La prévision finale :

Nous avons appliqués cette méthode à nos 2 variables à expliquer.

Nous conservons celle pour laquelle les métriques sont les plus hautes (rappel on vise AUC = 0.75 et precision = 0.4). Comme énoncé dans la partie «Le choix de la variable à expliquer», nous conservons la variable qui est aussi conditionnée par l'augmentation du rendement par rapport à la dernière période.

Maintenant affinons notre modèle, comme on le voit graphiquement l'AUC redescend à partir d'un certain nombre d'estimations, donc en ajoutant un principe d'arrêt anticipé (*early_stopping_rounds*) sur la métrique AUC nous pouvons gagnons en rapidité.



Pour pour espérer un meilleur AUC, on peut fixer l'hyperparamètre *learning_rate=0.001* (*0.02* auparavant) sur le modèle final, plus cet hyperparamètre est faible plus le modèle est coûteux en temps, mais en ajoutant *early_stopping_rounds=100* on compense cette hausse de temps d'exécution. On gagne effectivement 0.02 d'AUC, 0.03 de précision mais on perd 0.05 de recall.

Finalement d'après notre étude, ce modèle capte plus de 25% des nos signaux d'achats. Lorsqu'il prédit une hausse de rendement par rapport au mois précédent et un rendement supérieur à 2%, il y a 50% de chance que ce soit vrai.

Importance des variables explicatives :

On y compte l'apparition des mots de la liste filtrée, le rendement historique mensuel, le volume du jour, la variation historique du volume journalier.

Pour sélectionner les variables les plus importantes nous regardons *Cover* et *Gain* des features du modèle. Puisque les *features* « mots » sont binaires, le *weight* n'est pas adapté pour interpréter l'ordre d'importance des *features*.

Evidemment, le niveau de rendement du mois précédent (*HIST_M : f15*) est largement classé 1er pour le *gain* et le *cover*, c'est logique puisqu'il fait partie de la condition de la variable à expliquer. De ces graphiques, on remarque aussi que les données sur le volume (*VO : f16 ; VOL_J : f17*) jouent un rôle non négligeable dans le modèle (2e et 4e pour le *cover* et 2e et 3e pour le *gain*), donc il est rationnel de laisser ces données de volume dans nos inputs.

Intéressons-nous aux mots en particulier, 3 mots sont communs dans le top 7 du *Gain* et du *Cover :*

- *f12 : nouveau*
- *f5 : passe*
- *f10 : capital*

Commentaires sur le projet :