

# Etude préalable à l'investissement

Rapport rédigé le 15 janvier 2021 à Strasbourg par Matthieu Colombert et Rodrigue Ulua Longomo.

Client : Banque publique d'investissement (BPI)

Objet : réalisation d'un modèle prédictif d'aide à l'investissement

La BPI cherche à investir dans des startups. Elle avait besoin d'un outil qui permette de prédire quelles startups étaient les plus intéressantes.

Le jeu de données qui nous a été communiqué contenait cinq variables. Nous disposons des informations suivantes pour chaque entreprise : les dépenses de recherche et développement, les dépenses de marketing, les dépenses d'administration, la ville de l'entreprise et son profit.

Le critère principal dans une décision d'investissement nous a semblé être le profit de l'entreprise. Nous avons fait du profit le cible de notre modèle de prédiction.

I\ La ville et les dépenses d'administration : des critères peu pertinents

La mise en œuvre de la méthode de la « backward elimination » (A) nous a permis d'écarter les variables villes et dépenses d'administration de notre modèle prédictif du profit (B).

A\ Le choix de la méthode de la « backward elimination »

L'étude a débuté par un travail d'exploration des données fournies. Cette étape a mis en évidence une très forte corrélation linéaire entre les dépenses de recherche et développement et le profit et, dans une moindre mesure, les dépenses en marketing et le profit.

En revanche, dans ce travail préparatoire nous n'avons identifié aucune corrélation entre les dépenses d'administration et la localisation géographique et le profit.

Cette première étape nous a donné une idée des variables pertinentes et de celles que nous pourrions éliminer de notre modèle.

Nous avons fait le choix de la méthode de la « backward elimination » afin de retirer de notre modèle les variables les moins pertinentes et ainsi augmenter sa fiabilité. Notre client cherche à investir. Nous pouvons tolérer un certain risque dans l'élaboration de notre modèle. Nous avons choisi un seuil de significativité de 5%.

B\ L'élimination des dépenses d'administration et la ville de notre modèle

Après la phase exploratoire, nous sommes passés à l'entraînement de notre modèle. L'entraînement de notre premier modèle avec la totalité des variables indépendantes a confirmé nos hypothèses émises pendant la phase exploratoire. En effet, la variable ville a une P value de 0,788 (78,8%) soit bien au-dessus de notre seuil de significativité de 5%.

Nous avons réentraîné notre modèle sans la variable villes. Nous avons constaté que la P value de la variable administration était de 0,738. Elle aussi est nettement supérieure à notre seuil de significativité de 5%. Nous avons décidé de supprimer de notre modèle la variable dépenses d'administration. Ces résultats ont été confirmés par le logiciel Gretl.

gretl : modèle 5

Fichier Édition Tests Enregistrer Graphiques Analyse LaTeX

Modèle 5: MCO, utilisant les observations 1-50  
Variable dépendante: Profit

	coefficient	éc. type	t de Student	p. critique	
const	50122.2	6572.35	7.626	1.06e-09	***
RD	0.805715	0.0451473	17.85	2.63e-22	***
Administration	-0.0268160	0.0510288	-0.5255	0.6018	
Marketing	0.0272281	0.0164512	1.655	0.1047	
Moyenne var. dép.	112012.6	Éc. type var. dép.	40306.18		
Somme carrés résidus	3.92e+09	Éc. type régression	9232.335		
R <sup>2</sup>	0.950746	R <sup>2</sup> ajusté	0.947534		
F(3, 46)	295.9781	P. critique (F)	4.53e-30		
Log de vraisemblance	-525.3857	Critère d'Akaike	1058.771		
Critère de Schwarz	1066.420	Hannan-Quinn	1061.684		

Constante mise à part, la probabilité critique est la plus élevée pour la variable 2 (Administration)

L'élimination des critères de la localisation géographique et des dépenses d'administration de la prédiction du profit n'a pas posé de problème. Mais parfois le choix de retirer ou non une variable peut dépendre des circonstances.

## II\ L'utilisation des dépenses de recherche et développement et des dépenses de marketing dans la prédiction du profit

Nous avons choisi de conserver la variable dépenses de marketing dans notre modèle (B) malgré le fait que la « backward elimination » a montré que les dépenses de R&D étaient de loin le meilleur critère (A) de prédiction du profit.

A\ Les dépenses de R&D : la variable la plus importante dans la prédiction du profit

Après avoir éliminé les dépenses d'administration et les villes de notre modèle, nous avons poursuivi notre analyse. Nous avons à nouveau entraîné notre modèle avec seulement deux variables : les dépenses de R&D et les dépenses de marketing. Les statistiques de sortie ont mis en évidence une P value supérieure à notre seuil de significativité (6,2% pour un seuil de significativité choisi à 5%). Ces résultats ont été confirmés par le logiciel Gretl.

gretl : modèle 6

Fichier Édition Tests Enregistrer Graphiques Analyse LaTeX

Modèle 6: MCO, utilisant les observations 1-50  
Variable dépendante: Profit

	coefficient	éc. type	t de Student	p. critique	
const	46975.9	2689.93	17.46	3.50e-22	***
RD	0.796584	0.0413476	19.27	6.04e-24	***
Marketing	0.0299879	0.0155200	1.927	0.0600	*

Moyenne var. dép.	112012.6	Éc. type var. dép.	40306.18
Somme carrés résidus	3.94e+09	Éc. type régression	9160.966
R2	0.950450	R2 ajusté	0.948342
F(2, 47)	450.7713	P. critique (F)	2.16e-31
Log de vraisemblance	-525.5354	Critère d'Akaike	1057.071
Critère de Schwarz	1062.807	Hannan-Quinn	1059.255

Nous avons réentraîné notre modèle uniquement avec la variable R&D. Nous avons constaté une diminution de notre « adjusted R-squared », passé de 0,969 à 0,966. La P-value de la variable R&D tend vers 0, soit très en dessous de notre seuil de significativité. Ces résultats ont été confirmés par Gretl.

gretl : modèle 7

Fichier Édition Tests Enregistrer Graphiques Analyse LaTeX

Modèle 7: MCO, utilisant les observations 1-50  
Variable dépendante: Profit

	coefficient	éc. type	t de Student	p. critique	
const	49032.9	2537.90	19.32	2.78e-24	***
RD	0.854291	0.0293856	29.15	3.50e-32	***

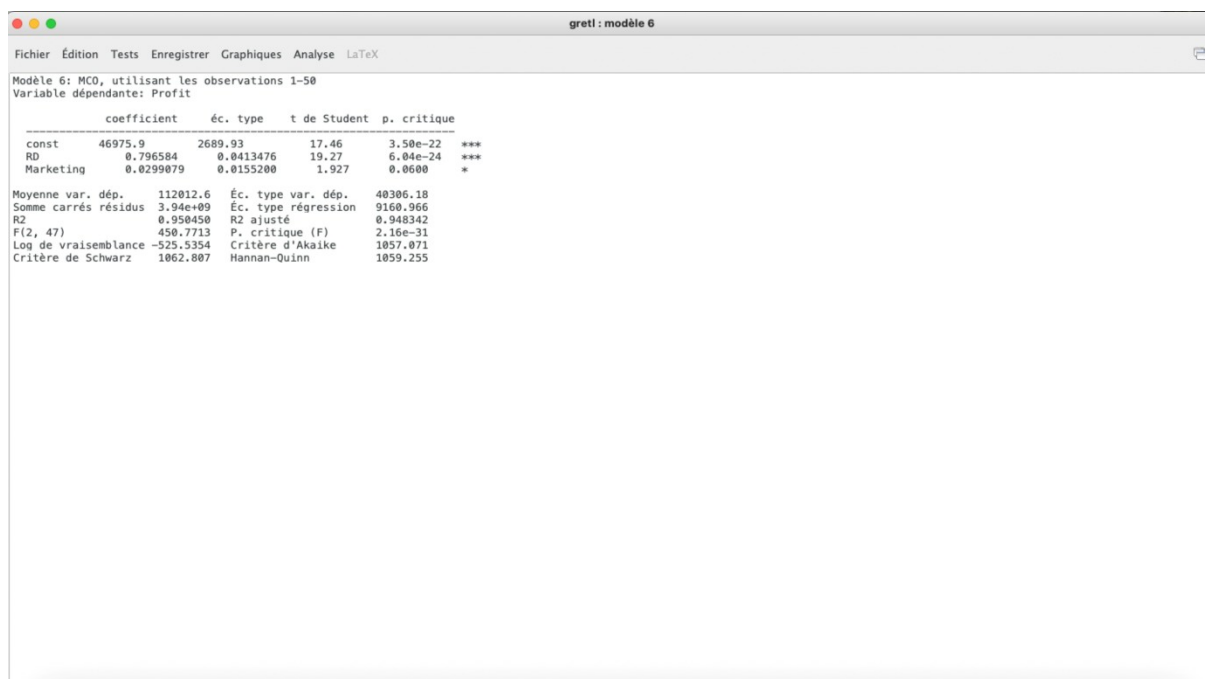
Moyenne var. dép.	112012.6	Éc. type var. dép.	40306.18
Somme carrés résidus	4.26e+09	Éc. type régression	9416.349
R2	0.946535	R2 ajusté	0.945421
F(1, 48)	849.7089	P. critique (F)	3.50e-32
Log de vraisemblance	-527.4365	Critère d'Akaike	1058.873
Critère de Schwarz	1062.697	Hannan-Quinn	1060.329

Nous avons bien entendu décidé de conserver la variable R&D, qui est effectivement le critère le plus pertinent pour prédire le profit d'une startup. Mais nous nous sommes aussi interrogés sur la pertinence d'éliminer ou non la variable Marketing de notre modèle.

## B\ Les dépenses de marketing : un critère contesté mais conservé

Comme expliqué précédemment, la variable marketing a une P-value de 6,2%, soit légèrement supérieure à notre seuil de significativité de 5%. L'intégrer à notre modèle serait un risque. Mais notre modèle gagnerait en précision. L'objet de cette étude est de réaliser des prédictions de profit de startups afin d'aider des décisions à l'investissement. Une prise de risque ne serait pas envisageable dans un domaine comme celui de la santé. En revanche, en matière de décision d'investissement, il nous a semblé plus pertinent de favoriser la précision de notre modèle au pris d'un risque plus important. Nous avons choisi de conserver la variable marketing dans notre modèle prédictif.

Notre modèle final comprend donc deux variables indépendantes : les dépenses de R&D et les dépenses de marketing. Les coefficients directs de notre modèle sont de 0,78 pour les dépenses de recherche et développement et de 0,02 pour les dépenses de marketing. Ces résultats ont été confirmés par Gretl.



	coefficient	éc. type	t de Student	p. critique	
const	46975.9	2689.93	17.46	3.50e-22	***
RD	0.796584	0.0413476	19.27	6.04e-24	***
Marketing	0.0299879	0.0155200	1.927	0.0600	*

Moyenne var. dép.	112012.6	Éc. type var. dép.	40306.18
Somme carrés résidus	3.94e+09	Éc. type régression	9160.966
R2	0.950450	R2 ajusté	0.948342
F(2, 47)	450.7713	P. critique (F)	2.16e-31
Log de vraisemblance	-525.5354	Critère d'Akaike	1057.071
Critère de Schwarz	1062.807	Hannan-Quinn	1059.255

Dans le cadre d'une décision d'investissement, nous recommandons de s'appuyer particulièrement sur les dépenses de R&D. Les dépenses de marketing sont beaucoup moins pertinentes. Elles peuvent être prises en compte mais uniquement d'une manière marginale par rapport aux dépenses de R&D.

En revanche, la localisation géographique de l'entreprise et les dépenses d'administration n'ont que peu d'impact sur la prédiction du profit. Le seul cas où les dépenses d'administration pourraient rentrer en compte dans la décision d'investissement serait la constatation d'une valeur anormale par rapport à des entreprises comparables.