Nguyen Kim
Court Thomas

Rapport Projet SAS

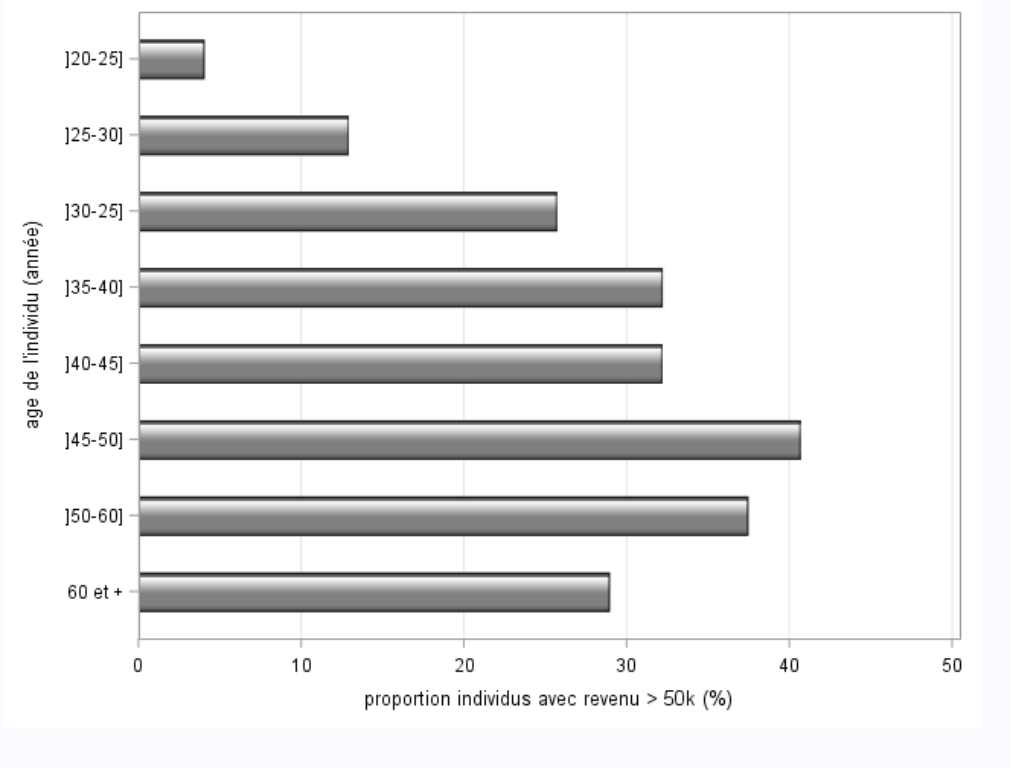
Introduction :

Le but de ce projet est d'étudier les différents modèles (et de choisir le meilleur), estimant la probabilité qu'un individu ait un revenu supérieur à 50K à partir d'un jeu de données de 1845 individus. Nous connaissons de nombreuses caractéristiques sur ces individus, à savoir : leur âge, leur nombre d'années d'études, leur situation familiale, leur profession, etc.

Le logiciel SAS nous a permis d'analyser ces données en testant des modèles statistiques. Le but étant, par exemple, de comprendre que certaines variables explicatives sont peu utiles à la résolution du problème dans les modèles.

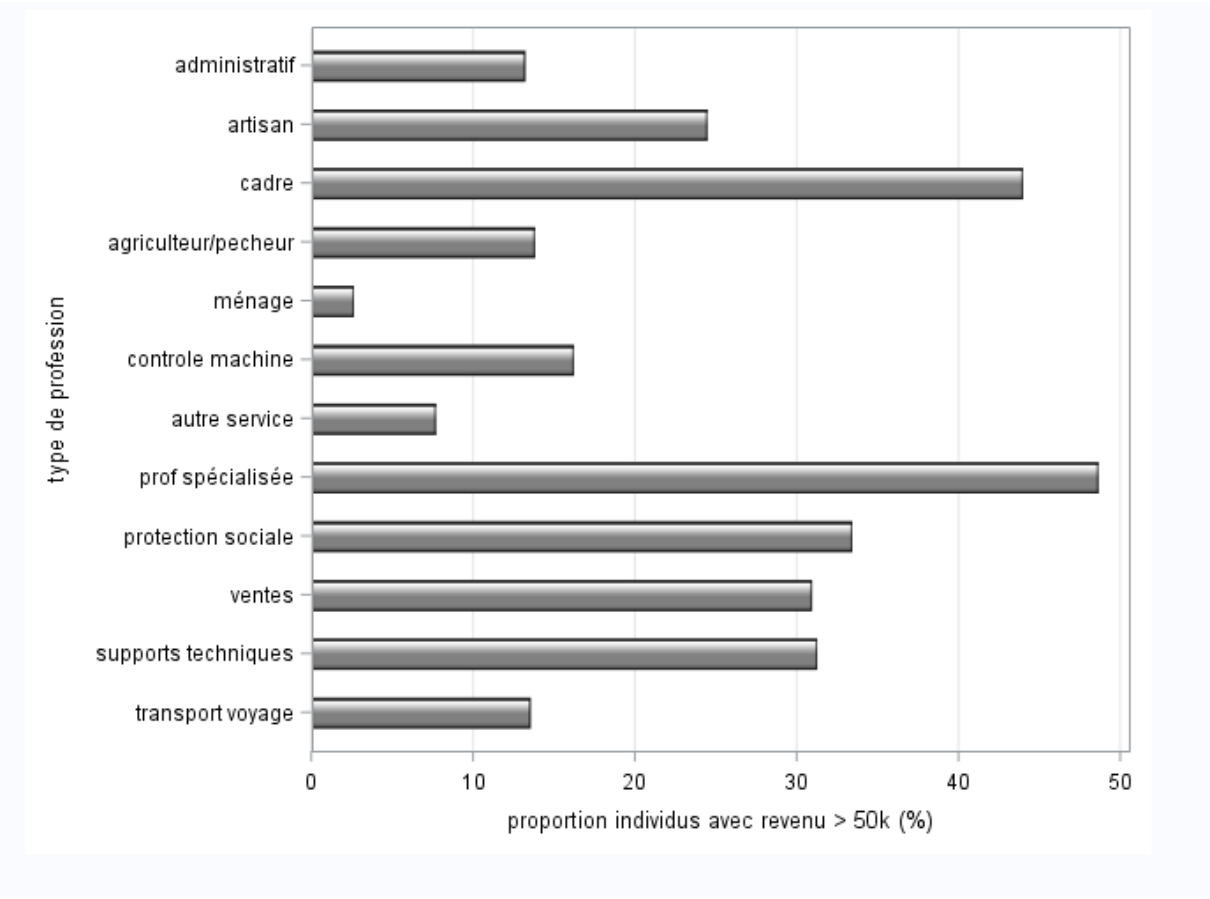
Partie Analyse :

Nous allons dans cette partie faire une approche basique de statistiques entre les différentes variables explicatives comparativement à la variable à expliquer qui est le revenu en utilisant les fonctions plot/sgplot :

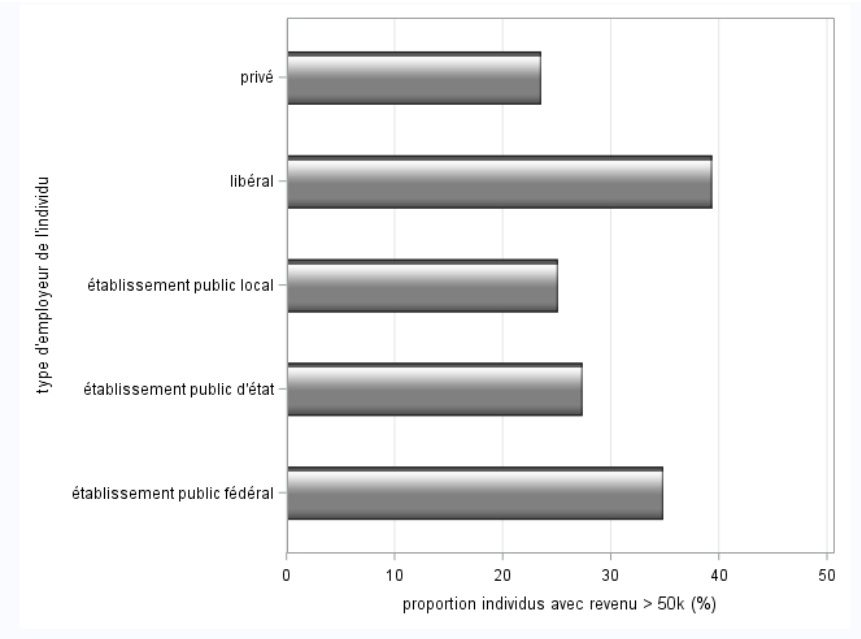


On peut voir ici que plus de 40% des personnes ayant entre 45-50 ans sont payées au-delà des 50k par an tandis que seulement 5% des 20-25 ans ont un tel revenu.

Nguyen Kim
Court Thomas



De même, ici les personnes travaillant dans des professions spécialisées ou étant cadre sont globalement plus souvent au-dessus des 50k de salaire que les ménages/autres services qui sont rarement payées plus de 50k par an.



Au contraire, dans ce graphique on voit que les types d'employeurs des individus sont plutôt homogènes dans leur répartition.

A l'aide de ces graphiques on peut s'attendre à ce que la probabilité d'avoir un salaire de plus de 50k pour une personne étant cadre et ayant plus de 40 ans soit nettement plus grande qu'un personne travaillant dans les ménages et ayant 30 ans. Ce ne sont que des déductions faites à partir des données observées mais nous calculerons les probabilités dans la suite.

Avant de calculer ces probabilités, il faut au préalable enlever des variables si elles sont inutiles afin d'améliorer le modèle saturé. Pour ce faire, nous avons utilisé la méthode de sélection automatique backward par la fonction proc logistic qui nous a permis d'obtenir les données suivantes :

Informations sur le modèle

Etape 0. Les effets suivants ont été saisis :

Intercept heurecat sexe agecat prof marital emploi origine etude invest

Statistiques d'ajustement du modèle

Critère	*Constante uniquement*	*Constante et Covariables*
AIC	2117.207	1354.527
SC	2122.728	1636.059
-2 Log L	2115.207	1252.527

R carré	0.3735	*R carré remis à l'échelle max.*	0.5474

Test de l'hypothèse nulle globale : BETA=0			
Test	*Khi-2*	*DDL*	*Pr > Khi-2*
Rapport de vrais	862.6804	50	<.0001
Score	730.5788	50	<.0001
Wald	402.5141	50	<.0001

Estimations par l'analyse du maximum de vraisemblance

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	90.2	*D de Somers*	0.804
Pourcentage discordant	9.8	*Gamma*	0.804
Pourcentage lié	0.0	*Tau-a*	0.310
Paires	655200	*c*	0.902

Analyse des effets éligibles pour la suppression			
Effet	*DDL*	*Khi-2 de Wald*	*Pr > Khi-2*
heurecat	3	23.8123	<.0001
sexe	1	1.0000	0.3173
agecat	7	26.4138	0.0004
prof	11	53.2592	<.0001
marital	4	177.9221	<.0001
emploi	4	5.2802	0.2597
origine	3	11.0641	0.0114
etude	15	50.1901	<.0001
invest	2	60.8276	<.0001

Etape 1. Effet sexe supprimé :

Etat de convergence du modèle
Critère de convergence (GCONV=1E-8) respecté.

Statistiques d'ajustement du modèle		
Critère	*Constante uniquement*	*Constante et Covariables*
AIC	2117.207	1353.526
SC	2122.728	1629.538
-2 Log L	2115.207	1253.526

R carré	0.3731	*R carré remis à l'échelle max.*	0.5469

Test de l'hypothèse nulle globale : BETA=0			
Test	*Khi-2*	*DDL*	*Pr > Khi-2*
Rapport de vrais	861.6813	49	<.0001
Score	728.4248	49	<.0001
Wald	402.2130	49	<.0001

Analyse des effets Type 3			
Effet	*DDL*	*Khi-2 de Wald*	*Pr > Khi-2*
heurecat	3	25.6402	<.0001
agecat	7	26.7411	0.0004
prof	11	52.5194	<.0001
marital	4	216.7018	<.0001
emploi	4	5.4620	0.2431
origine	3	11.4178	0.0097
etude	15	50.6199	<.0001
invest	2	60.7214	<.0001

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	90.2	*D de Somers*	0.803
Pourcentage discordant	9.8	*Gamma*	0.803
Pourcentage lié	0.0	*Tau-a*	0.309
Paires	655200	*c*	0.902

Test du Khi-2 résiduel		
Khi-2	*DDL*	*Pr > Khi-2*
1.0013	1	0.3170

Etape 2. Effet emploi supprimé :

Statistiques d'ajustement du modèle		
Critère	Constante uniquement	Constante et Covariables
AIC	2117.207	1351.069
SC	2122.728	1604.999
-2 Log L	2115.207	1259.069

R carré	0.3713	R carré remis à l'échelle max.	0.5442

Test de l'hypothèse nulle globale : BETA=0			
Test	Khi-2	DDL	Pr > Khi-2
Rapport de vrais	856.1387	45	<.0001
Score	724.9685	45	<.0001
Wald	402.9808	45	<.0001

Analyse des effets Type 3			
Effet	DDL	Khi-2 de Wald	Pr > Khi-2
heurecat	3	26.0791	<.0001
agecat	7	26.2166	0.0005
prof	11	51.4411	<.0001
marital	4	218.2637	<.0001
origine	3	12.2288	0.0066
etude	15	48.1412	<.0001
invest	2	61.4369	<.0001

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	90.1	*D de Somers*	0.802
Pourcentage discordant	9.9	*Gamma*	0.802
Pourcentage lié	0.0	*Tau-a*	0.309
Paires	655200	*c*	0.901

Test du Khi-2 résiduel		
Khi-2	*DDL*	*Pr > Khi-2*
6.4981	5	0.2607

Analyse des effets éligibles pour la suppression			
Effet	*DDL*	*Khi-2 de Wald*	*Pr > Khi-2*
heurecat	3	26.0791	<.0001
agecat	7	26.2166	0.0005
prof	11	51.4411	<.0001
marital	4	218.2637	<.0001
origine	3	12.2288	0.0066
etude	15	48.1412	<.0001
invest	2	61.4369	<.0001

Note: No (additional) effects met the 0.1 significance level for removal from the model.

Récapitulatif sur l'élimination en arrière						
Etape	*Effet supprimé*	*DDL*	*Nombre dans*	*Khi-2 de Wald*	*Pr > Khi-2*	*Libellé de variable*
1	*sexe*	1	8	1.0000	0.3173	Sexe
2	*emploi*	4	7	5.4620	0.2431	type d'employeur de l'individu

Statistique d'adéquation de la déviance et de Pearson

Critère	*Valeur*	*DDL*	*Valeur/DDL*	*Pr > Khi-2*
Ecart	1152.9184	1620	0.7117	1.0000
Pearson	1692.4994	1620	1.0448	0.1026

Nombre de profils uniques : 1666

Association des probabilités prédites et des réponses observées

Pourcentage concordant	90.1	*D de Somers*	0.802
Pourcentage discordant	9.9	*Gamma*	0.802
Pourcentage lié	0.0	*Tau-a*	0.309
Paires	655200	*c*	0.901

Partition pour les tests de Hosmer et de Lemeshow

		revenu = au dessus de 50k		*revenu = en dessous de 50k*	
Groupe	*Total*	*Observé*	*Attendu*	*Observé*	*Attendu*
1	186	1	0.64	185	185.36
2	185	2	2.59	183	182.41
3	185	2	4.93	183	180.07
4	185	8	8.81	177	176.19
5	185	9	16.27	176	168.73
6	185	41	30.32	144	154.68
7	185	57	54.19	128	130.81
8	185	85	83.72	100	101.28
9	185	124	121.35	61	63.65
10	179	151	157.16	28	21.84

Le logiciel SAS nous dit que les variables sexe et emploi sont peu significatives et peuvent être enlevées. En effet, on peut le voir aux valeurs de la colonne *Pr > Khi-2* (surlignées ci-dessus). Lorsque l'on prend en compte toutes les variables explicatives du modèle, toutes les valeurs de la colonne sont petites exceptées celles de sexe et emploi. Le logiciel enlève alors la plus grande valeur : ici Sexe (0.3173). Après élimination de la variable sexe, on remarque que la valeur *Pr > Khi-2* de variable emploi est toujours grande tandis que les autres restent petites. Le logiciel procède alors à la suppression de la variable « emploi ». Après avoir enlevé les variables explicatives sexe et emploi, on obtient le tableau suivant :

Analyse des effets Type 3

Effet	*DDL*	*Khi-2 de Wald*	*Pr > Khi-2*
agecat	7	27.4523	0.0003
prof	11	103.8730	<.0001
marital	4	220.5474	<.0001
origine	3	19.0328	0.0003
invest	2	65.2242	<.0001
heurecat	3	31.7094	<.0001
etudecat	2	12.6420	0.0018

On remarque que toutes les variables sont significatives dans ce modèle. En effet, les valeurs de la colonne de droite (PR>Khi-2) sont toutes suffisamment petites.

Nous avons finalement utilisé les coefficients de chaque donnée grâce à la fonction proc logit :

Estimations par l'analyse du maximum de vraisemblance

Paramètre		*DDL*	*Estimation*	*Erreur type*	*Khi-2 de Wald*	*Pr > Khi-2*	*Exp(Est)*	*Libellé*
Intercept		1	-1.6017	0.7235	4.9012	0.0268	0.202	Intercept: revenu=au dessus de 50k
agecat	*60 et +*	1	0.1173	0.3523	0.1109	0.7392	1.124	age de l'individu (année) 60 et +
agecat	*]20-25]*	1	-1.3944	0.4009	12.0984	0.0005	0.248	age de l'individu (année)]20-25]
agecat	*]25-30]*	1	-0.8460	0.2842	8.8608	0.0029	0.429	age de l'individu (année)]25-30]
agecat	*]30-25]*	1	-0.3065	0.2541	1.4552	0.2277	0.736	age de l'individu (année)]30-25]
agecat	*]35-40]*	1	-0.1320	0.2429	0.2950	0.5870	0.876	age de l'individu (année)]35-40]
agecat	*]40-45]*	1	-0.1212	0.2457	0.2434	0.6217	0.886	age de l'individu (année)]40-45]
agecat	*]45-50]*	1	0.2964	0.2609	1.2904	0.2560	1.345	age de l'individu (année)]45-50]
prof	*administratif*	1	-0.5062	0.3062	2.7318	0.0984	0.603	type de profession administratif
prof	*agriculteur/pecheur*	1	-1.7910	0.5062	12.5203	0.0004	0.167	type de profession agriculteur/pecheur
prof	*artisan*	1	-0.5771	0.2721	4.4994	0.0339	0.562	type de profession artisan
prof	*autre service*	1	-0.9553	0.3942	5.8721	0.0154	0.385	type de profession autre service

Estimations par l'analyse du maximum de vraisemblance

Paramètre		*DDL*	*Estimation*	*Erreur type*	*Khi-2 de Wald*	*Pr > Khi-2*	*Exp(Est)*	*Libellé*
prof	*cadre*	1	0.3713	0.2628	1.9963	0.1577	1.450	type de profession cadre
prof	*controle machine*	1	-1.0187	0.3538	8.2931	0.0040	0.361	type de profession controle machine
prof	*ménage*	1	-2.6260	0.8218	10.2107	0.0014	0.072	type de profession ménage
prof	*prof spécialisée*	1	0.9671	0.2743	12.4261	0.0004	2.630	type de profession prof spécialisée
prof	*protection sociale*	1	0.1260	0.4319	0.0851	0.7706	1.134	type de profession protection sociale
prof	*supports techniques*	1	0.3143	0.3837	0.6712	0.4127	1.369	type de profession supports techniques
prof	*transport voyage*	1	-1.2764	0.3842	11.0352	0.0009	0.279	type de profession transport voyage
marital	*célibataire*	1	-0.2837	0.5305	0.2860	0.5928	0.753	statut marital célibataire
marital	*divorcé*	1	-0.2027	0.5362	0.1430	0.7054	0.816	statut marital divorcé
marital	*marié*	1	2.2187	0.4997	19.7166	<.0001	9.195	statut marital marié
marital	*séparé*	1	0.0145	0.6773	0.0005	0.9829	1.015	statut marital séparé
origine	*amérindien*	1	-0.7970	0.3145	6.4242	0.0113	0.451	Origine ethnique amérindien
origine	*asie pacifique*	1	0.4379	0.2839	2.3800	0.1229	1.550	Origine ethnique asie pacifique
origine	*blanc*	1	0.1539	0.2316	0.4413	0.5065	1.166	Origine ethnique blanc
invest	*gains*	1	0.6292	0.3481	3.2670	0.0707	1.876	Investissements annuel gains
invest	*ni pertes ni gains*	1	-1.0667	0.2881	13.7037	0.0002	0.344	Investissements annuel ni pertes ni gains
heurecat	*1-20*	1	-1.9819	0.4363	20.6338	<.0001	0.138	Nombre d'heures travaillées/semaine 1-20
heurecat	*21-39*	1	-0.9056	0.2787	10.5587	0.0012	0.404	Nombre d'heures travaillées/semaine 21-39
heurecat	*40-45*	1	-0.6853	0.1657	17.0972	<.0001	0.504	Nombre d'heures travaillées/semaine 40-45
etudecat	*bac*	1	1.1606	0.3477	11.1395	0.0008	3.192	Nombre d'années d'études bac
etudecat	*doctorat*	1	2.1801	0.9015	5.8483	0.0156	8.847	Nombre d'années d'études doctorat

A partir de cela, nous avons pu faire une régression logistique et déterminer suivant les données d'un nouvel individu la probabilité qu'il ait un revenu supérieur à 50K. Voici un aperçu de ce que nous avons obtenu :

Obs.	id	revenu	age	agecat	emploi	prof	marital	etude	origine	sexe	invest	heure	heurecat	etudecat	revenu2	note	gain	Proba
1	84	en dessous de 50k	22	]20-25]	privé	ménage	célibataire	8	amérindien	homme	ni pertes ni gains	40	40-45	pas le bac	0	-7.3881	.	0.00062
2	121	en dessous de 50k	23	]20-25]	privé	ménage	célibataire	8	amérindien	homme	ni pertes ni gains	40	40-45	pas le bac	0	-7.3881	.	0.00062
3	475	en dessous de 50k	30	]25-30]	privé	ménage	célibataire	9	amérindien	femme	ni pertes ni gains	8	1-20	bac	0	-6.9757	.	0.00093
4	1817	en dessous de 50k	68	60 et +	privé	autre service	veuf	4	amérindien	femme	ni pertes ni gains	20	1-20	pas le bac	0	-6.7303	.	0.00119
5	34	en dessous de 50k	21	]20-25]	privé	ménage	célibataire	9	noir américain	homme	ni pertes ni gains	20	1-20	bac	0	-6.7271	.	0.00120
6	95	en dessous de 50k	23	]20-25]	privé	ménage	célibataire	10	blanc	homme	ni pertes ni gains	10	1-20	bac	0	-6.5732	.	0.00140
7	210	en dessous de 50k	25	]20-25]	privé	agriculteur/pecheur	célibataire	6	amérindien	homme	ni pertes ni gains	40	40-45	pas le bac	0	-6.5531	.	0.00142
8	136	en dessous de 50k	23	]20-25]	privé	ménage	célibataire	10	amérindien	femme	ni pertes ni gains	35	21-39	bac	0	-6.4478	.	0.00158
9	238	en dessous de 50k	25	]20-25]	privé	ménage	célibataire	8	blanc	homme	ni pertes ni gains	43	40-45	pas le bac	0	-6.4372	.	0.00160
10	1767	en dessous de 50k	62	60 et +	privé	ménage	divorcé	4	blanc	homme	ni pertes ni gains	40	40-45	pas le bac	0	-6.3562	.	0.00173
11	35	en dessous de 50k	21	]20-25]	privé	ménage	célibataire	9	amérindien	homme	ni pertes ni gains	40	40-45	bac	0	-6.2275	.	0.00197
12	47	en dessous de 50k	22	]20-25]	privé	ménage	célibataire	9	amérindien	homme	ni pertes ni gains	40	40-45	bac	0	-6.2275	.	0.00197
13	94	en dessous de 50k	23	]20-25]	privé	ménage	célibataire	10	amérindien	homme	ni pertes ni gains	40	40-45	bac	0	-6.2275	.	0.00197
14	223	en dessous de 50k	25	]20-25]	privé	ménage	célibataire	10	amérindien	homme	ni pertes ni gains	40	40-45	bac	0	-6.2275	.	0.00197
15	242	en dessous de 50k	25	]20-25]	établissement public fédéral	ménage	célibataire	9	amérindien	homme	ni pertes ni gains	40	40-45	bac	0	-6.2275	.	0.00197
16	1801	en dessous de 50k	66	60 et +	privé	transport voyage	veuf	6	blanc	femme	ni pertes ni gains	11	1-20	pas le bac	0	-6.1005	.	0.00224
17	274	en dessous de 50k	26	]25-30]	établissement public local	ménage	célibataire	9	amérindien	homme	ni pertes ni gains	30	21-39	bac	0	-5.8994	.	0.00273
18	78	en dessous de 50k	22	]20-25]	privé	agriculteur/pecheur	célibataire	8	amérindien	homme	ni pertes ni gains	50	45 et +	pas le bac	0	-5.8678	.	0.00282
19	736	en dessous de 50k	35	]30-25]	privé	agriculteur/pecheur	célibataire	8	blanc	homme	ni pertes ni gains	20	1-20	pas le bac	0	-5.8109	.	0.00299
20	1774	en dessous de 50k	63	60 et +	privé	autre service	veuf	5	blanc	femme	ni pertes ni gains	20	1-20	pas le bac	0	-5.7794	.	0.00308

1823	588	au dessus de 50k	32	]30-25]	privé	prof spécialisée	marié	15	asie pacifique	femme	gains	50	45 et +	bac	100	2.8761	.	0.94665
1824	1437	en dessous de 50k	48	]45-50]	libéral	cadre	marié	10	asie pacifique	homme	ni pertes ni gains	65	45 et +	bac	0	2.8832	.	0.94701
1825	1686	au dessus de 50k	57	]50-60]	privé	prof spécialisée	marié	13	blanc	homme	gains	50	45 et +	bac	100	2.8986	.	0.94778
1826	1709	au dessus de 50k	58	]50-60]	établissement public d'état	prof spécialisée	marié	14	blanc	homme	pertes	50	45 et +	bac	100	2.8986	.	0.94778
1827	1726	au dessus de 50k	59	]50-60]	privé	prof spécialisée	marié	14	blanc	homme	ni pertes ni gains	50	45 et +	bac	100	2.8986	.	0.94778
1828	1731	en dessous de 50k	59	]50-60]	établissement public d'état	prof spécialisée	marié	14	blanc	homme	ni pertes ni gains	50	45 et +	bac	0	2.8986	.	0.94778
1829	1375	au dessus de 50k	47	]45-50]	établissement public local	prof spécialisée	marié	10	noir américain	homme	gains	50	45 et +	bac	100	3.0411	.	0.95440
1830	926	au dessus de 50k	38	]35-40]	libéral	prof spécialisée	marié	13	asie pacifique	homme	gains	60	45 et +	bac	100	3.0506	.	0.95481
1831	871	en dessous de 50k	37	]35-40]	établissement public d'état	prof spécialisée	marié	16	blanc	homme	ni pertes ni gains	45	40-45	doctorat	0	3.1008	.	0.95693
1832	1022	au dessus de 50k	40	]35-40]	privé	prof spécialisée	marié	16	blanc	homme	ni pertes ni gains	40	40-45	doctorat	100	3.1008	.	0.95693
1833	1032	au dessus de 50k	40	]35-40]	privé	prof spécialisée	marié	16	blanc	homme	ni pertes ni gains	40	40-45	doctorat	100	3.1008	.	0.95693
1834	1317	au dessus de 50k	46	]45-50]	privé	prof spécialisée	marié	13	blanc	homme	ni pertes ni gains	50	45 et +	bac	100	3.1950	.	0.96065
1835	1349	au dessus de 50k	46	]45-50]	libéral	prof spécialisée	marié	15	blanc	homme	ni pertes ni gains	65	45 et +	bac	100	3.1950	.	0.96065
1836	1355	au dessus de 50k	46	]45-50]	privé	prof spécialisée	marié	14	blanc	homme	gains	46	45 et +	bac	100	3.1950	.	0.96065
1837	1366	au dessus de 50k	47	]45-50]	privé	prof spécialisée	marié	13	blanc	homme	gains	50	45 et +	bac	100	3.1950	.	0.96065
1838	1447	au dessus de 50k	49	]45-50]	établissement public local	prof spécialisée	marié	15	blanc	femme	pertes	60	45 et +	bac	100	3.1950	.	0.96065
1839	1523	au dessus de 50k	51	]50-60]	établissement public d'état	prof spécialisée	marié	16	blanc	homme	ni pertes ni gains	40	40-45	doctorat	100	3.2328	.	0.96205
1840	1612	au dessus de 50k	54	]50-60]	privé	prof spécialisée	marié	16	blanc	homme	gains	45	40-45	doctorat	100	3.2328	.	0.96205
1841	1342	au dessus de 50k	46	]45-50]	établissement public fédéral	prof spécialisée	marié	13	asie pacifique	homme	ni pertes ni gains	50	45 et +	bac	100	3.4790	.	0.97008
1842	1331	au dessus de 50k	46	]45-50]	établissement public d'état	prof spécialisée	marié	16	blanc	homme	gains	45	40-45	doctorat	100	3.5292	.	0.97151
1843	1276	au dessus de 50k	45	]40-45]	libéral	prof spécialisée	marié	16	blanc	homme	ni pertes ni gains	60	45 et +	doctorat	100	3.7969	.	0.97805
1844	1510	au dessus de 50k	51	]50-60]	établissement public fédéral	prof spécialisée	marié	16	blanc	homme	gains	52	45 et +	doctorat	100	3.9181	.	0.98051
1845	1600	au dessus de 50k	53	]50-60]	établissement public d'état	prof spécialisée	marié	16	blanc	homme	ni pertes ni gains	50	45 et +	doctorat	100	3.9181	.	0.98051

Par exemple, si nous ajoutons un individu de sexe masculin, amérindien, célibataire, de 20-25 ans travaillant entre 40 et 45 heures par semaine dans les ménages, dans le privé et n'ayant pas son bac, sa probabilité de gagner plus de 50 000 euros par an est de 0.00062 soit presque autant de chance que d'être frappé par la foudre !! (*En réalité la probabilité d'être frappé par la foudre est de 1 sur 250000*).

A l'inverse, un homme marié blanc, âgé de 50-60 ans, travaillant dans un établissement public d'état pour une profession spécialisée doté d'un doctorat et travaillant plus de 45 heures par semaine a une probabilité de gagner plus de 50K euros par an de 0.98051.