2024

TP01 - Rapport

62-62 DATA MINING

HÜNI VICTOR

Contents

Figure

Tableau

Introduction

L'objectif de ce rapport sera de présenté l'analyse réalisé sur le jeu de donnée « HR-prediction-all.csv ». Dans la peau d'un consultant engagé par les ressources humaines d'une entreprise notre rapport doit présenter à la direction les raisons pour lesquels certains employés démissionnent.

Grâce à cet analyse, l'entreprise souhaite mettre en avant des indicateurs pertinents afin de prévenir les départs avant qu'ils ne surviennent.

Résumé

Découvertes principales

Le datasets contient 10k employés avec chacun 9 attributs exploitable (3 qualitatifs et 6 quantitatifs).

La probabilité pour un employé de quitter l'entreprise dans ce dataset est de

$$P(left = 1) = 23{,}81\%$$

Les variables déterminées comme indicatives d'un potentiels départ sont :

- Le niveau de salaire
- Le temps passé dans l'entreprise
- Le niveau de satisfaction
- Le nombre d'heure de travail moyenne par mois

En effet, la probabilité de quitter l'entreprise est très élevé chez les personnes ayant des bas salaires avec $\mathrm{P}(\mathrm{left} = 1 \,|\mathrm{salary} = \mathrm{low}) = 30\%$

De plus, la probabilité de quitter l'entreprise augmente plus la personne reste longtemps dans l'entreprise passant de $P(left = 1|time\_spend\_company = 2) = 1.66\%$ les 2 première année à $P(left = 1|time\_spend\_company \leq 4) = 36.62\%$ à partir de 4 ans d'ancienneté.

Le niveau satisfaction semble également être un bon indicateur du potentiel départ des employés. En effet, $P(left = 1|satisfaction_level < 0{,}5) = 55{,}57\%$ contre seulement $P(left = 1|satisfaction\_level >= 0{,}5) = 9{,}92\%$

Par ailleurs, les personnes travaillant peu entre 130 et 160 heures par mois ont une probabilité de quitté l'entreprise de $P(left = 1|130 <= average\_montly\_hours < 160) = 39{,}09\%$. De même les personnes travaillantes beaucoup entre 240 et 270h ont eu une probabilité de quitté l'entreprise de $P(left = 1|130 <= average\_montly\_hours < 160) = 28{,}38\%$.Ainsi les personnes travaillant trop ou pas assez semblent être des candidats pour une future démission.

Enfin l'analyse croisé de ces deux attributs quantitatifs (satisfaction_level et average_montly_average) nous montre que des employés avec des heures équilibrés) entre 150 et 250h) et un haut niveau de satisfaction sont plus enclin à rester dans la compagnie. À contrario, les employés travaillant trop (>240h) ou pas assez (<160) ET un niveau de satisfaction bas (<0,5) ont un risque très élevé de départ.

Analyse préliminaire

Ce type d'analyse débute toujours avec un descriptif détaillé des données et de leur qualité. Dans notre cas, voici comment se présente le jeu de donnée « HR-prediction-all.csv » :

- 10 000 instances (lignes) qui représentent 10 000 employés
- 11 attributs (colonnes) qui représentent les 11 attributs à notre disposition pour effectuer notre analyse.

Variable cible

Sur ces 11 attributs une est notre variable cible, celle que nous chercherons à expliquer à travers notre analyse. Ainsi, la variable cible, ici, est la variable qualitative nommé *« left »* qui prend deux valeurs :

- 0 signifie que la personne est restée dans l'entreprise
- 1 signifie que la personne est partie de l'entreprise

$$P(left = 0) = 76{,}19\%$$

$$P(left = 1) = 23{,}81\%$$

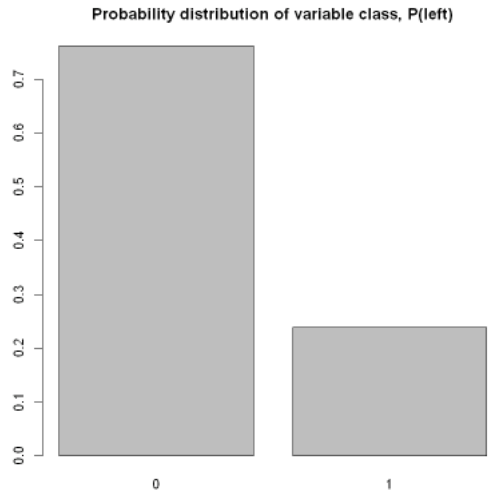

Figure 1 : Distribution de la probabilité P(left)

Cela signifie que près d'¼ des employés du jeux de données ont décidé de quitter l'entreprise. Nous allons tenter d'expliquer ce phénomène en analysant les variables explicatives.

Le mode de la variable cible est donc 0.

Variables explicatives

Le reste des attributs sont détaillé dans le tableau 1 à la page suivante. En résumé, le jeu de données comporte :

- 3 variables quantitatives

- 6 variables qualitatives

Variable exclue

La variable exclue est la première colonne *« Id »*. Bien que numérique, c'est une variable d'identification et elle n'est ni quantitatif ni qualitatif et ne nous permettra pas d'expliquer les valeurs de la variable cible.

Qualité des données

Notre analyse met en évidence qu'aucune instance du jeu de données ne possède de données manquantes et aucune variable ne contient des valeurs non définies tels que NA. Enfin, aucune instance ne semble également être dupliquées.

Tableau 1 : Classifications des variables

Nom de Variable	**Type**	**Valeurs**	**Raisons/Commentaire**
satisfaction_level	Variable quantitative	Une note comprise entre 0 et 1	Variable continue avec Min. = 0.090 1st Qu. = 0.440 Médiane =0.640 Moyenne = 0.614 3rd Qu. = 0.820 Max. = 1.000
last_evaluation	Variable quantitative	Une note comprise entre 0 et 1	Variable continue avec Minimum = 0.3600 1st Qu. = 0.5600 Médiane : 0.7200 Moyenne = 0.7176 3rd Qu. = 0.8700 Max. = 1.0000
number_project	Variable qualitative	6 valeurs possible représentant le nombre de projets : 2, 3, 4, 5, 6, 7	Variable discrète. Bien que les valeurs soient bien des valeurs numériques, l'exploitation de cet attribut sera plus logique dans une analyse qualitative que quantitative du fait du nombre faible de valeurs possible
average_montly_hours	Variable quantitative	Valeurs comprises entre 96 et 310 heures / mois en moyenne.	Variable continue avec Min. = 96.0 1st Qu. = 156.0 Médiane :199.0 Moyenne = 200.7 3rd Qu. = 245.0 Max. = 310.0
time_spend_company	Variable qualitative	8 valeurs possibles représentant le nombre d'année : 2, 3, 4, 5, 6, 7, 8, 10	Variable discrète. Bien que les valeurs soient bien des valeurs numériques, l'exploitation de cet attribut sera plus logique dans une analyse

			qualitative que quantitative du fait du nombre faible de valeurs possible
work_accident	Variable qualitative	• 0 signifie que la personne n'a pas eu d'accident • 1 signifie que la personne a eu un accident du travail	
promotion_last_5years	Variable qualitative	• 0 signifie que la personne n'a pas eu de promotion dans les 5 dernières années • 1 signifie que la personne a eu une promotion dans les 5 dernières années	
department	Variable qualitative	10 valeurs possibles : • Accounting • Hr • It • Management • Marketing, • Management, • product_mng, • RandD, Sales, Support, • technical	
salary	Variable qualitative	High Medium low	

Attributs qualitatifs

Vous trouverez ci-dessous la sélection des diagrammes pour toutes les variables qualitatives désignés comme pertinente. L'objectifs ici est de trouver les variables avec le plus d'impact sur la variable cible. C'est pourquoi pour chacune variable vous avez :

- À gauche, le diagramme de la probabilité $P(x)$ représentant la distribution de la probabilité de la variable permettant d'évaluer l'importance de la variable dans la population. **Nous souhaitons une variable distribuée équitablement à travers l'ensemble des valeurs possibles afin d'éviter une analyse biaisée**
- À droite, la probabilité conditionnelle $P(left|x)$ permettant de mettre en évidence l'impact de la variable sur la variable cible. **Nous recherchons une variable avec des différences notables dans ses probabilités conditionnelles indiquant que la variable pourrait être un indicateur pertinent**

Attributs qualitatifs avec impact un notable

Niveau de salaire

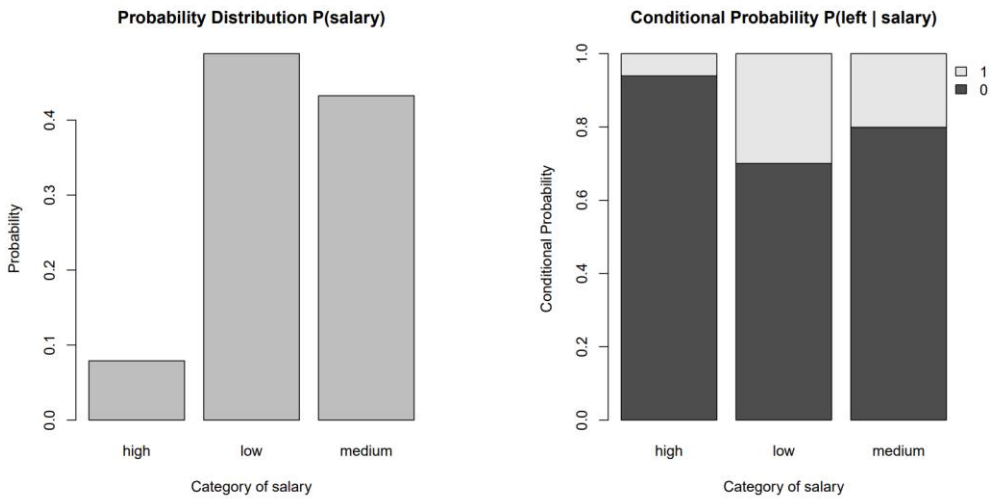

Figure 2 : Analyse distribution de la variable "salary"

Les employés à bas salaire ont un taux de départ de

$$\mathrm{P}(\mathrm{left} = 1 \,|\mathrm{salary} = \mathrm{low}) = 30\%$$

Ce taux est très significatif en comparaison des 20% et 6% des salaires medium et haut. Les écarts entre ces taux de départs sont donc respectivement. 10% et 24%.De plus la population des salaires bas et médium est presque équivalente environ 45% chacun.

Ainsi, cette variable semble être un bon candidat. En effet, nous pouvons confirmer que le niveau de salaire est un indicateur important du taux de départ dans l'entreprise. **Plus le salaire est bas plus les départs sont fréquent.**

Temps passés au sein de l'entreprise

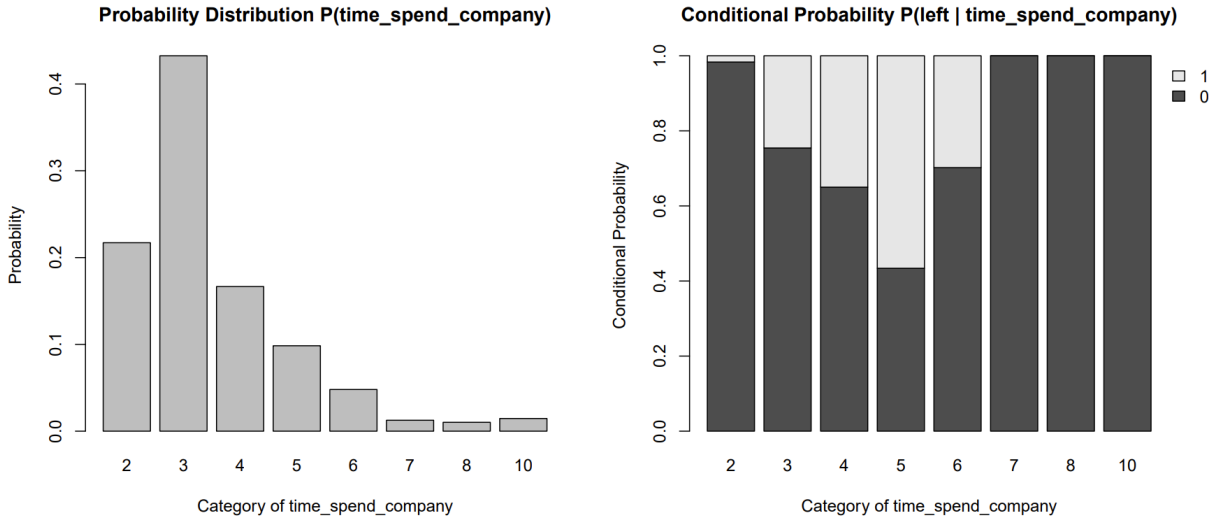

Figure 3 : Analyse distribution de la variable "time_spend_company"

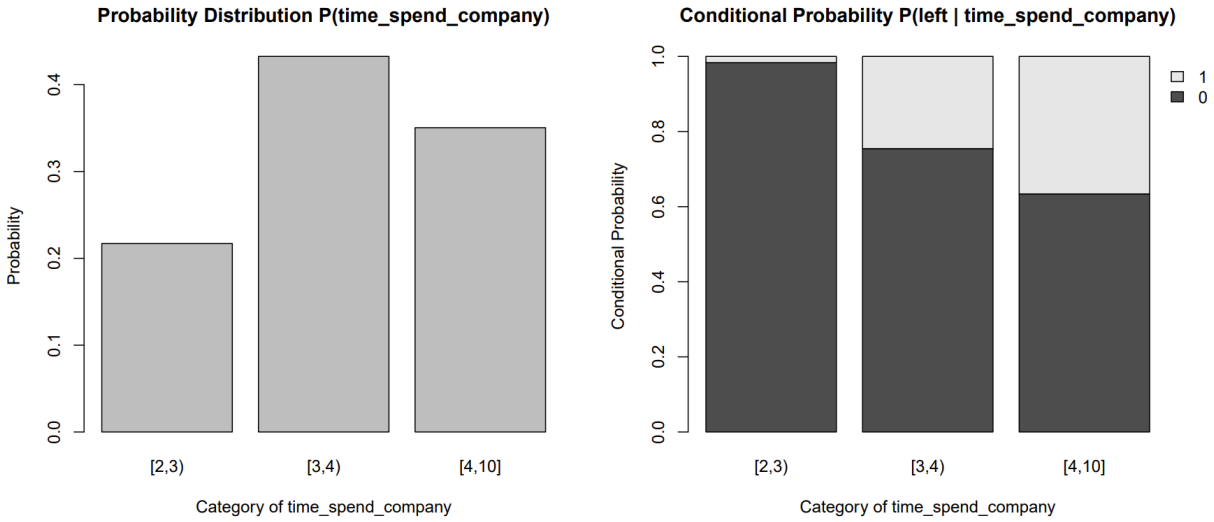

Figure 4 : Analyse distribution de la variable "time_spend_company" (grouped)

Cette variable présente une asymétrie en faveur des faibles valeurs. Autrement dit, la majorité des employés du jeux de données ont passée entre 2 et 4 ans dans l'entreprise. Il y a peu d'employés étant resté 6 ans et plus. Nous pourrions conclure qu'à priori, le jeu de données est peut-être biaisé en faveur des employés récemment arrivé.

Cependant en groupant les observations dans 3 groupes distinct et en analysant les probabilités conditionnelles de cette variable, nous pouvons diminuer ce biais avec une répartition plus

égale. De plus, nous pouvons clairement observer une corrélation positive entre la probabilité de départ et le nombre d'année d'expérience.

En effet,

[2, 3) - Le taux de départs des employés présent dans l'entreprise depuis seulement 2 ans est

$$P(left = 1 \,|time\ spend\ company = 2) = 1.66\%$$

[3, 4) – Puis le taux de départs des employés avec 3 ans d'expérience est

$$P(left = 1 \,|time\ spend\ company = 3) = 24.55\%$$

[4, 10] – Enfin, les employés avec le plus d'ancienneté ont un taux de départ de

$$P(left = 1 \mid 4 \leq time\ spend\ company \leq 10) = 36.62\%$$

Ainsi, à partir de 4 ans d'expérience on observe un risque accru de départ de l'entreprise, ce qui après cette analyse semble être un bon prédicteur pour notre variable cible.

Attributs qualitatifs avec un impact modéré/faible

Work Accident

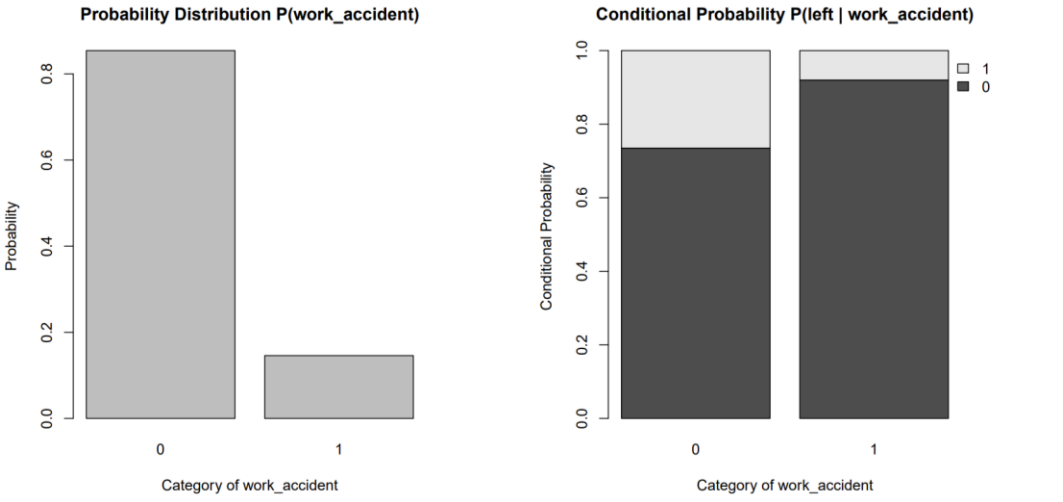

Figure 5 : Analyse distribution de la variable "work_accident"

Les employés qui n'ont pas eu d'accident ont un taux de départ

$$\mathrm{P}(\mathrm{left}\ =\ 1 \mid \mathrm{work\_accident}\ =\ 0) = 26.5\%$$

Alors que ceux qui ont eu un accident ont un taux de départ bien plus bas :

$$\mathrm{P}(\mathrm{left}\ =\ 1 \mid \mathrm{work\_accident}\ =\ 1) = 8\%$$

Ainsi la différence de taux départ entre ceux qui ont eu un accident et ceux qui n'en n'ont pas eu est d'environ 18%. Ce qui est un écart tout même important comparé aux autres variables. Cela suggère dans un premier temps que ne pas avoir d'accident est corrélé avec un taux de départ plus élevé.

Cependant la majorité des employés (85.41%) n'ont jamais eu d'accident. Ceci rend cet attribut moins significatif que d'autres pour expliquer les départs car la distribution de la variable en elle-même est très déséquilibré. Les employés n'ayant jamais eu d'accident de travail sont sur représenté.

Promotion dans les 5 dernières années

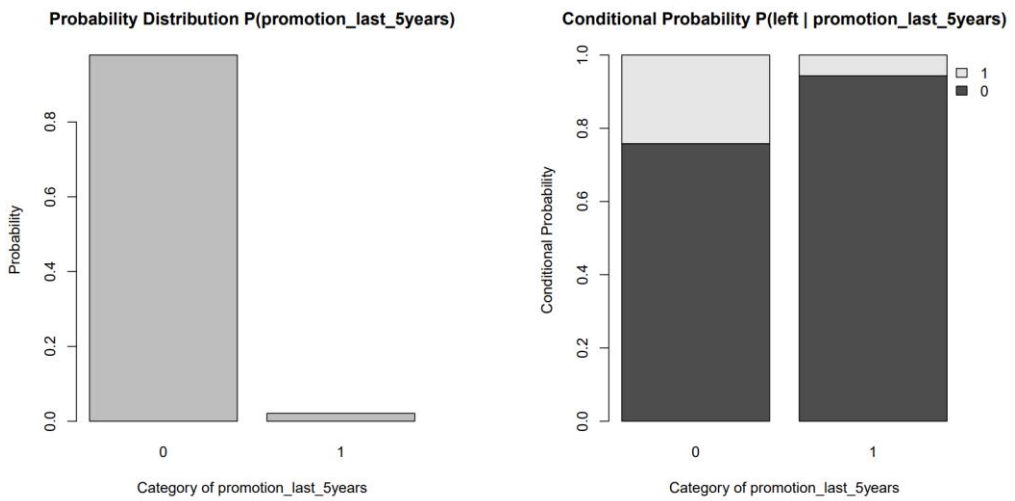

Figure 6 : Analyse distribution de la variable "promotion_last_5years"

Les employés qui n'ont pas reçu de promotion depuis 5 ans ont un taux de départ de près d'¼

$$\mathrm{P}(\text{left} = 1 \mid \text{promotion\_last\_years\_5years} = 0) = 24.2\%$$

Alors que ceux ayant reçu une promotion ont un taux de départ de seulement

$$\mathrm{P}(\text{left} = 1 \mid \text{promotion\_last\_years\_5years} = 1) = 5.7\%$$

L'écart entre ces taux de départ est donc de prêts de 20%. Ceci indique donc qu'un manque de promotion durant 5 ans est significativement corrélé aux départs des employé.

Cependant 97.8% des employés du jeux de données sont des employés n'ayant pas reçu de promotions depuis 5 ans. Ainsi ces employés sont peut-être sur représentés et donc cette variable peut fausser la prédiction de départ d'un employé.

Département

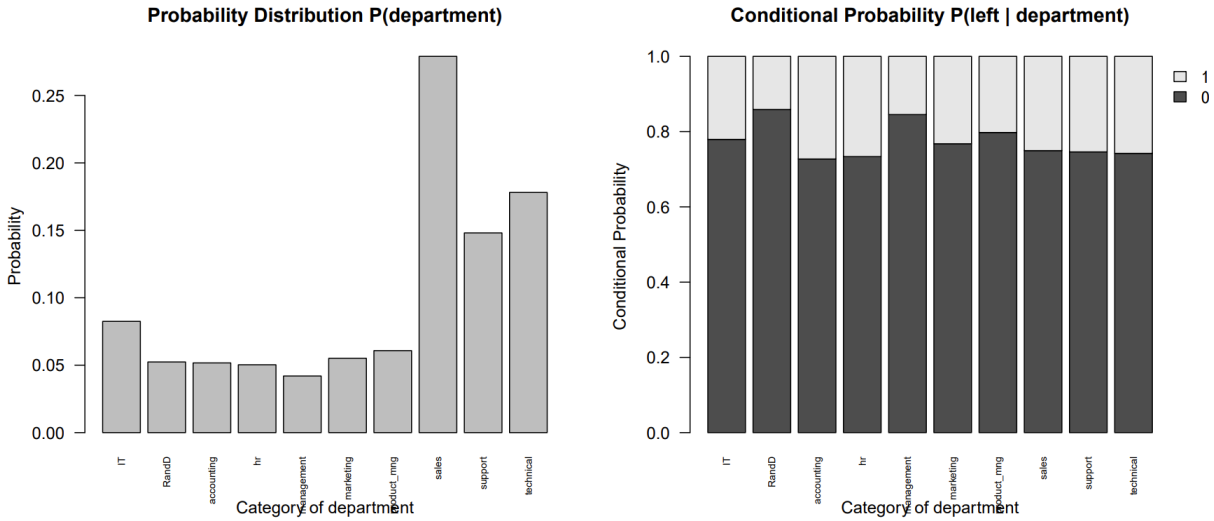

Figure 7 : Analyse distribution de la variable "department"

La population des départements est distribuée de manière relativement équitable, ce qui assure une représentation juste pour chacun des groupes. Le taux de départ dans le différent département varie de très peu d'un département à l'autre avec un minium pour

$$\mathrm{P}(\mathrm{left}\ =\ 1\ |\mathrm{deparment} = \mathrm{RandD}) = 14\%$$

Et un maximum pour

$$\mathrm{P}(\mathrm{left}\ =\ 1\ |\mathrm{department} = \mathrm{Accounting}) = 27\%$$

Cette plage de 13% indique une certaine variation entre les départements mais les différences ne sont pas si significatives. Bien que **les probabilités conditionnelles suggèrent de légères disparités dans les taux de rétention** entre les départements**, l'impact général sur la variable cible est relativement faible.**

Notions Théoriques

De la distribution jointe à la distribution conditionnelle

Prenons comme exemple la variable *« salary »*

Probabilités jointes

Tout d'abord, calculons la probabilités jointes des deux variable $P(left, salary)$ ou left est la variable cible. Cette dernières nous donne un tableau dont la somme de toutes les valeurs égales à un.

Tableau 2 : Probabilité jointes P(salary, left)

	$\boldsymbol{Left = 0}$	$\boldsymbol{Left = 1}$
$\boldsymbol{Salary = High}$	7,42%	0,48%
$\boldsymbol{Salary = Medium}$	34,46%	8,69%
$\boldsymbol{Salary = Low}$	34,21%	14,64%

Dans le cas, du niveau de salaire, la probabilité qu'une personne ait eu un salaire low et soit parti de l'entreprise est de $14{,}64\%$. À contrario, la probabilité qu'une personne ait eu un salaire low ou $medium$ et soit resté dans l'entreprise est d'environ 34% pour chacune des catégories de salaire,

Probabilités marginales

Ensuite, calculons les probabilités marginales de chacune des variables soit $P(left)$ et $P(salary)$. Si l'on observe le tableau des probabilité conjointes, on remarque que ces dernières peuvent être obtenu en sommant les valeurs de la bonne façon.

- Pour la variable cible $left$ (en colonne dans le tableau des probabilités conjointes), si on somme l'ensemble des valeurs par colonnes on obtiendra
  - $P(left = 0) = 76.19\%$
  - $P(left = 1) = 23{,}81\%$
- Pour la variable explicative $salary$ (en ligne dans le tableau des probabilités conjointes), si on somme l'ensemble des valeurs par lignes on obtiendra
  - $P(salary = high) 7{,}9\%$
  - $P(salary = medium) = 43{,}25\%$
  - $P(salary = low) = 48{,}85\%$

Enfin on calcule les distributions conditionnelles en utilisant les deux étapes précédentes.

Probabilités conditionnelles $P(left|salary)$.

Si on divise chaque cellule de la probabilités conjointes par la probabilité marginale $P(left = 1)$ ou $P(left = 0)$ en fonction de la colonne dans laquelle elle se trouve alors on obtient la probabilité conditionnelle $P(left|salary)$. Cette probabilité représente la probabilité d'un départ d'un employé sachant sont niveau de salaire. Ainsi $P(left = 1|salary = high) = 6{,}07\%$

Voici le détail des calculs associés :

Pour $salary = high$

- $P(\text{left} = 0 \mid \text{salary} = \text{high}) = \frac{P(\text{left=0 and salary=high})}{P(\text{salary=high})} = \frac{0.0742}{0.0790} = 0.9392$
- $P(\text{left} = 1 \mid \text{salary} = \text{high}) = \frac{0.0048}{0.0790} = 0.0608$

Pour $salary = medium$

- $P(\text{left} = 0 \mid \text{salary} = \text{medium}) = \frac{0.3456}{0.4325} = 0.7991$
- $P(\text{left} = 1 \mid \text{salary} = \text{medium}) = \frac{0.0869}{0.4325} = 0.2009$

Pour $salary = low$

- $P(\text{left} = 0 \mid \text{salary} = \text{low}) = \frac{0.3421}{0.4885} = 0.7003$
- $P(\text{left} = 1 \mid \text{salary} = \text{low}) = \frac{0.1464}{0.4885} = 0.2997$

Tableau 3 : Probabilité conditionnelle P(left|salary)

	$\boldsymbol{Left = 0}$	$\boldsymbol{Left = 1}$
$\boldsymbol{Salary = High}$	93,92%	6,08%
$\boldsymbol{Salary = Medium}$	79,91%	20,09%
$\boldsymbol{Salary = Low}$	70,03%	29,96%

Probabilités conditionnelle $P(salary|left)$

Si on divise chaque colonne de la probabilités conjointes par la probabilité marginale $P(salary)$ alors on obtient la probabilité conditionnelle $P(salary|left)$. Cette probabilité représente la probabilité des différents niveau salaire sachant s'il est parti ou non de l'entreprise. Ainsi $P(salary = high|left = 1) = 2{,}01\%$

Pour $left = 0$

- $P(\text{salary} = \text{high} \mid \text{left} = 0) = \frac{P(\text{left=0 and salary=high})}{P(\text{left=0})} = \frac{0.0742}{0.7619} = 0.0974$
- $P(\text{salary} = \text{medium} \mid \text{left} = 0) = \frac{0.3456}{0.7619} = 0.4536$
- $P(\text{salary} = \text{low} \mid \text{left} = 0) = \frac{0.3421}{0.7619} = 0.4489$

Pour $left = 1$

- $P(\text{salary} = \text{high} \mid \text{left} = 1) = \frac{P(\text{left=1 and salary=high})}{P(\text{left=1})} = \frac{0.0048}{0.2381} = 0.0202$
- $P(\text{salary} = \text{medium} \mid \text{left} = 1) = \frac{0.0869}{0.2381} = 0.3649$
- $P(\text{salary} = \text{low} \mid \text{left} = 1) = \frac{0.1464}{0.2381} = 0.6149$

Tableau 4 : Probabilités conditionnelle P(salary|left)

	$\boldsymbol{Left = 0}$	$\boldsymbol{Left = 1}$
$\boldsymbol{Salary = High}$	9,74%	2,02%
$\boldsymbol{Salary = Medium}$	45,36%	36,49%
$\boldsymbol{Salary = Low}$	44,90%	61,49%

Théorème de Bayes

Le théorème de Bayes s'exprime :

$$P(Y \mid X) = \frac{P(X \mid Y) \cdot P(Y)}{P(X)}$$

Où X est la variable cible (left) et Y est la variable explicative (salary).

Prenons par exemple la probabilité d'un employé ayant un salaire élevé ($Y = high$) étant donné qu'il a quitté l'entreprise ($X = 1$)

- $P(\text{salary} = \text{high} \mid \text{left} = 1) = 0.0202$
- $P(\text{left} = 1 \mid \text{salary} = \text{high}) = 0.0608$
- $P(\text{left} = 1) = 0.2381$
- $P(\text{salary} = \text{high}) = 0.079$

Appliquons le théorème de Bayes

- $P(\text{salary} = \text{high} \mid \text{left} = 1) = \frac{P(\text{left} = 1 \mid \text{salary} = \text{high}) \times P(\text{salary=high})}{P(\text{left=1})} = \frac{0.0608 \times 0.079}{0.2381}$
- $P(\text{salary} = \text{high} \mid \text{left} = 1) = 0{,}0202$

Attributs quantitatifs

Les trois variables désignées comme Quantitatives et donc analysé comme sont donc :

- satisfaction_level
- average_montly_hours
- last_evaluation

Voici le résultat de l'analyse exploratoire. Pour chacune des variables vous trouverez la valeur suivante :

- Mean : Moyenne globale de l'attribut
- Variance : Variance globale de l'attribut
- Std_Dev : Écart Type globale de l'attribut
- Mean_Stay : Moyenne de l'attribut pour les attributs avec left=0
- Std_Dev_Stay : Écart Type de l'attribut pour les instances avec left = 0
- Mean_Left : Moyenne de l'attribut pour les attributs avec left=1
- Std_Dev_Left : Écart Type de l'attribut pour les instances avec left = 1
- Importance_Score : Score permettant de discriminer les attributs les plus importants. Plus le score est important, plus est un bon prédicteur de la variable cible.

$$\frac{|\ \mu\_\{attribute|left = 0\} - \mu\_\{attribute|left = 1\}\ |}{\sigma_{attribute}}$$

Tableau 5 : Résultat d'analyse exploratoire des attributs quantitatifs

	Mean	Variance	Std_Dev	Mean_Stay	Std_Dev_Stay	Mean_Left	Std_Dev_Left	Importance_Score
satisfaction_level	0.61	0.06	0.25	0.67	0.22	0.44	0.26	0.92
average_montly_hours	200.69	2484.2	49.84	198.77	45.7	206.83	60.85	0.16
last_evaluation	0.72	0.03	0.17	0.72	0.16	0.72	0.2	0

Attributs quantitatifs avec impact un notable

Niveau de satisfaction

Le niveau de satisfaction prend des valeurs de 0.1 à 1 à travers les 10000 instances. La Moyenne du niveau de satisfaction culmine à 0,61 et l'écart type à 0,25 ce qui indique une variation plutôt modérée entre les employés.

Le premier histogramme montre une distribution plutôt égale légèrement asymétrique en faveur des valeurs au-dessus de 0,35. La moyenne est presque centrée. 68% des niveaux satisfaction se trouve entre 0,35 et 0,85.

Le second histogramme montre les distributions conditionnelles du niveau de satisfaction des employés étant resté dans l'entreprise ($left = 0$) versus ceux ayant quitté l'entreprise ($left = 1$).

Les distributions conditionnelles montrent clairement :

- Un employé avec un niveau de satisfaction bas à très bas ($satisfaction\_level < 0{,}5$) a plus de chance de quitté l'entreprise. Ceci est représenté par la prédominance des barres rouges dans cette zone.
- À contrario, un niveau de satisfaction élevé à très élevé ($satisfaction\_level \geq 0{,}5$) est plus souvent associé avec des employés qui semblent resté dans l'entreprise.

Ainsi, grâce au distribution conditionnelle, nous pouvons d'ores et déjà confirmé que le $satisfaction\_level$ est un bon prédicteur du taux de départ des employés. En effet, plus un employé est insatisfait, plus il aura de chance de quitter l'entreprise.

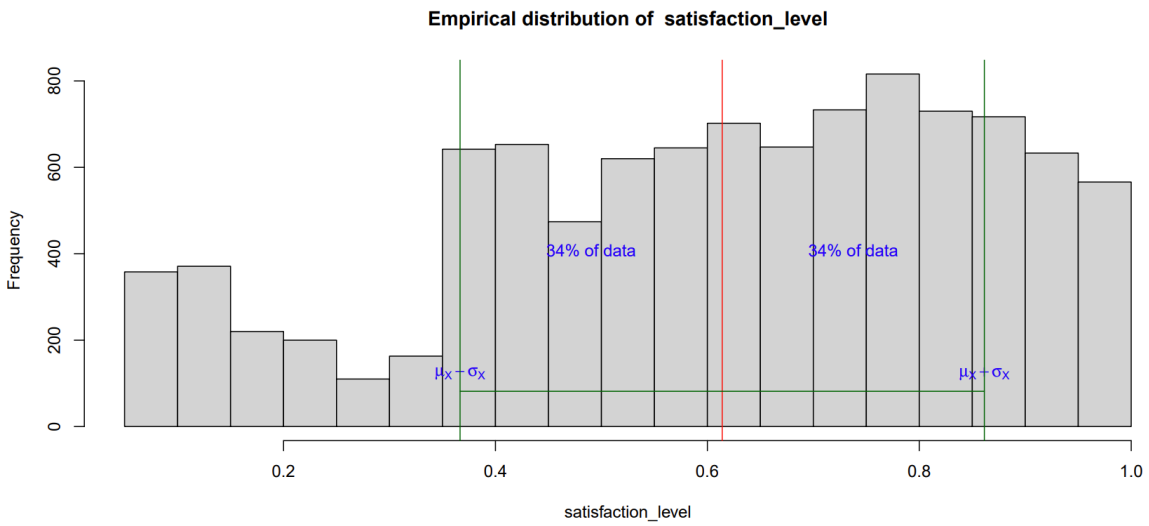

Figure 8 : Distribution empirique de la variable "satisafaction_level"

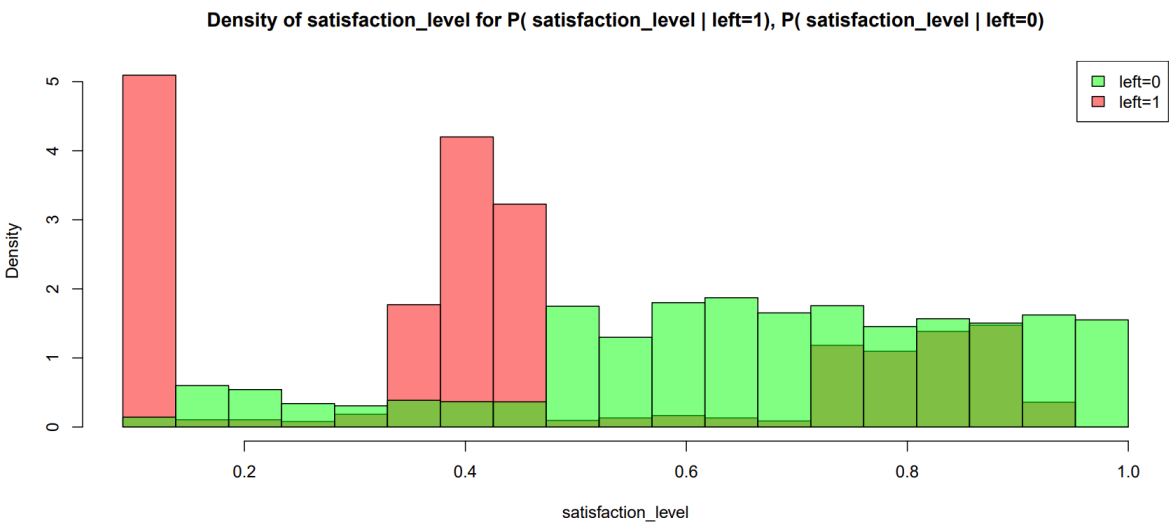

Figure 9 : Densité des probabilités conditionnelles P(satisfaction_level|left)

Nombre d'heures moyennes travaillés par mois

La distribution d'$average\_monthly\_hours$ semble démontré deux grandes tendances avec deux pics distincts :

- Entre 130 et 160 heures
- Entre 240 et 270 heures

La moyenne des heures travaillés est à 200,69h pour un écart type de 49,84h. 68% des données recueilli sont comprise entre 150h et 250h avec les deux pics aux extrémités comme discuté précédemment.

Dans la seconde visualisation, nous pouvons voir les distributions conditionnelles des average_montly_hours des personnes ayant quitté l'entreprise $P(average\_monthly\_hours|left = 1)$ et de celle étant resté $P(average\_monthly\_hours|left = 0)$.

Nous observons également deux pics dans le nombre de personne ayant quitté l'entreprise plus ou moins autour de mêmes valeurs que les deux clusters discutés précédemment :

- Entre 130 et 160 heures
- Entre 240 et 270 heures

Cela suggère donc que l'employé avec peu d'heure travaillé par mois et l'employé avec au contraire beaucoup d'heure par mois sont tous les deux de potentiel démissionnaire.

Les employés avec peu d'heure de travail (<150) sont synonyme d'insatisfaction ou de sous-utilisation des compétences alors que les personnes avec beaucoup d'heure peuvent signifier des personnes surchargées.

Enfin les personnes travaillantes autour des 200h par mois semblent être les personnes les plus enclin à rester dans l'entreprise.

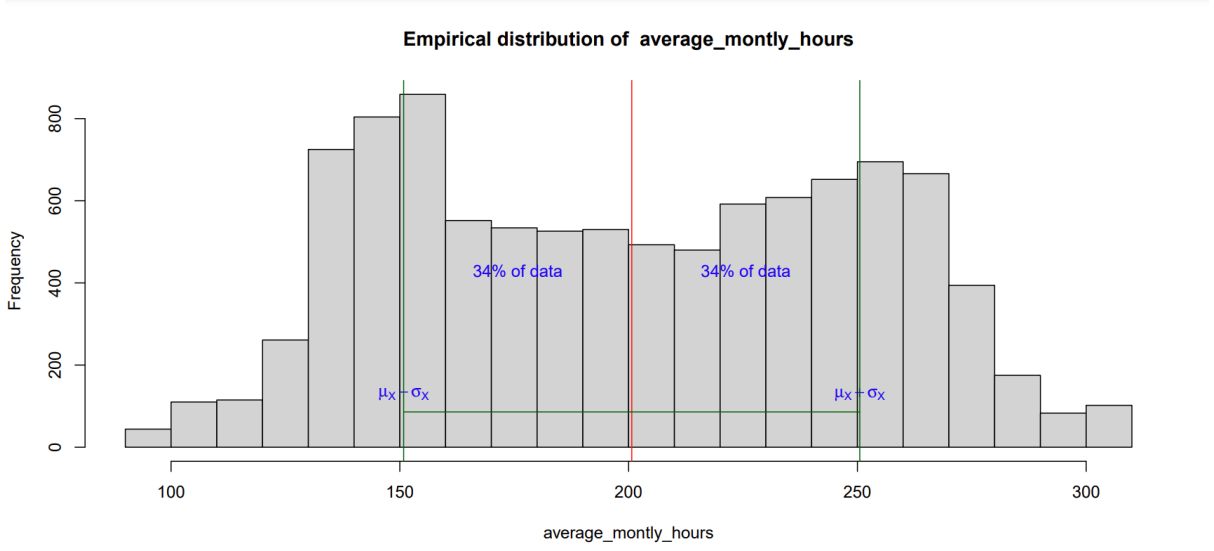

Figure 10 : Distribution empirique de la variable "average_monthly_hours"

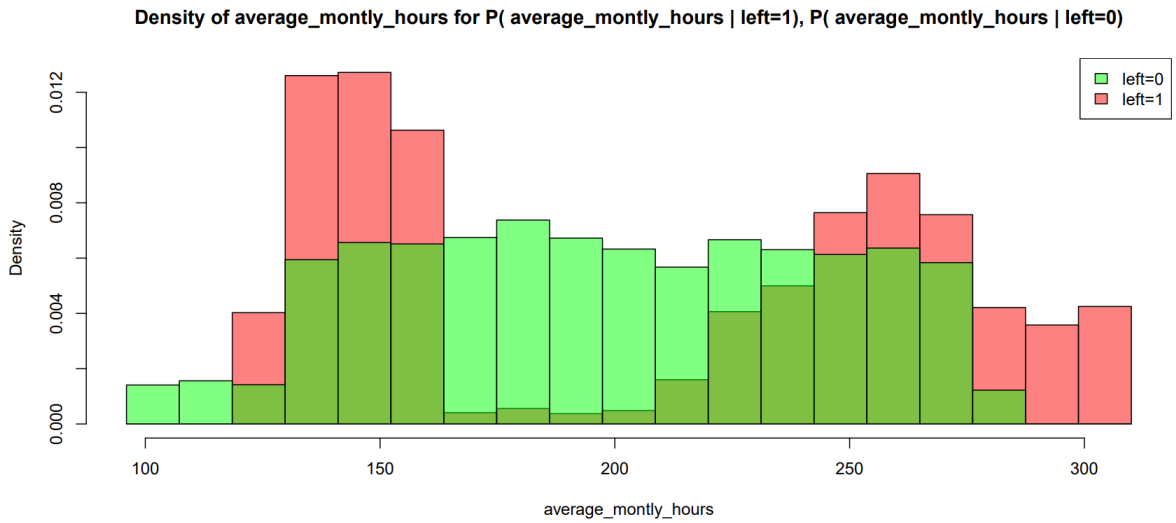

Figure 11 : Densité des probabilités conditionnelles P(average_monthly_hours|left)

Niveau de satisfaction vs Heures mensuelles

Tout d'abord, l'aspect régulier des points de mesures semblent indiquer un aspect synthétique des données. Les niveaux de satisfaction varient très régulièrement ce qui peux indiquer que les données ont été récupéré avec un « slider » cranté entre 0 et 1 et un step défini de 0,1. Les heures semblent très clustérisées laissant penser que les propositions faites dans le questionnaire étaient peut-être limité également.

Analysons désormais les différents clusters :

Niveau de satisfaction élevé et Heures moyenne

La majorité des employés étant resté dans l'entreprise semble avoir un haut taux de satisfaction $satisfaction\_level \geq 0{,}5$ et un nombre d'heures moyenne travaillé par mois entre 150h et 250h. Ce groupe est très visible en haut à droite du nuage de point.

Niveau de satisfaction très bas et heures moyennes basses OU élevés

Nous avons déterminé plus haut que les employés avec un taux de satisfaction bas $satisfaction\_level \leq 0{,}5$ tendent à partir de l'entreprise. De plus nous avons également établi que les employés avec peu d'heures de travail (<150h / mois) ou à contrario trop d'heure de travail (>250h) étaient également de bon candidat pour partir de l'entreprise.

Sur le nuage de points, nous pouvons bel et bien observer 2 clusters représentant ces personnes distinctement. Le premier cluster est en haut à gauche. Ce dernier représente les personnes avec un niveau de satisfaction très bas et beaucoup d'heure de travail. Le second cluster se trouve au milieu autour du niveau de satisfaction 0,4 et des heures aux alentours de 150h.

Niveau de satisfaction élevé et heures moyennes élevés

Le 3ème et dernier cluster représentant des employés sur le départ se trouve en haut à droite. Ces personnes ont un niveau de satisfaction plutôt élevé entre 0,7 et 0,9 mais des heures de travails importantes également aux alentours de 250h. Ainsi, certaine personne a beau être satisfaite de leur environnement de travail, la charge de travail trop importante fini tout de même par les pousser vers la sortie.

Le nuage de point vient donc corroborer les constats réalisés dans les analyses des distributions des deux attributs quantitatifs les plus significatifs. L'analyse croisé des deux nous permet d'affirmer que les employés prêts à partir ont :

- $satisfaction\_level \leq 0{,}1$ ET $average\_monthly\_hours > 250h$
- $satisfaction\_level$ aux alentours de 0,4 ET $average\_monthly\_hours$ atour de 150
- $0{,}7 < satisfaction\_level < 0{,}9$ ET $220 < average\_monthly\_hours < 270h$

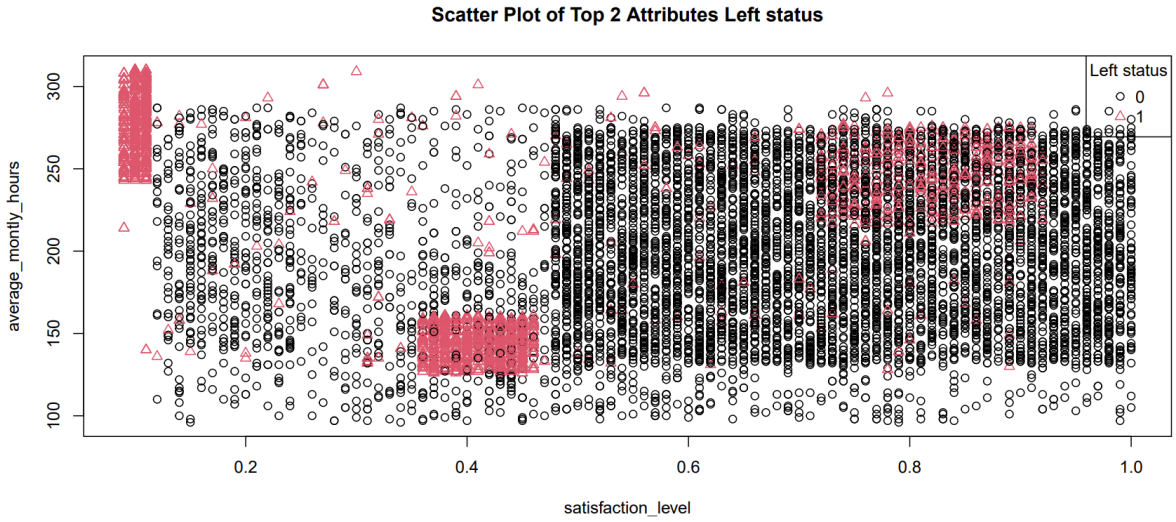

Figure 12 : Nuage de points "average_monthly_hours" et "satisfaaction_level"

Notions théoriques

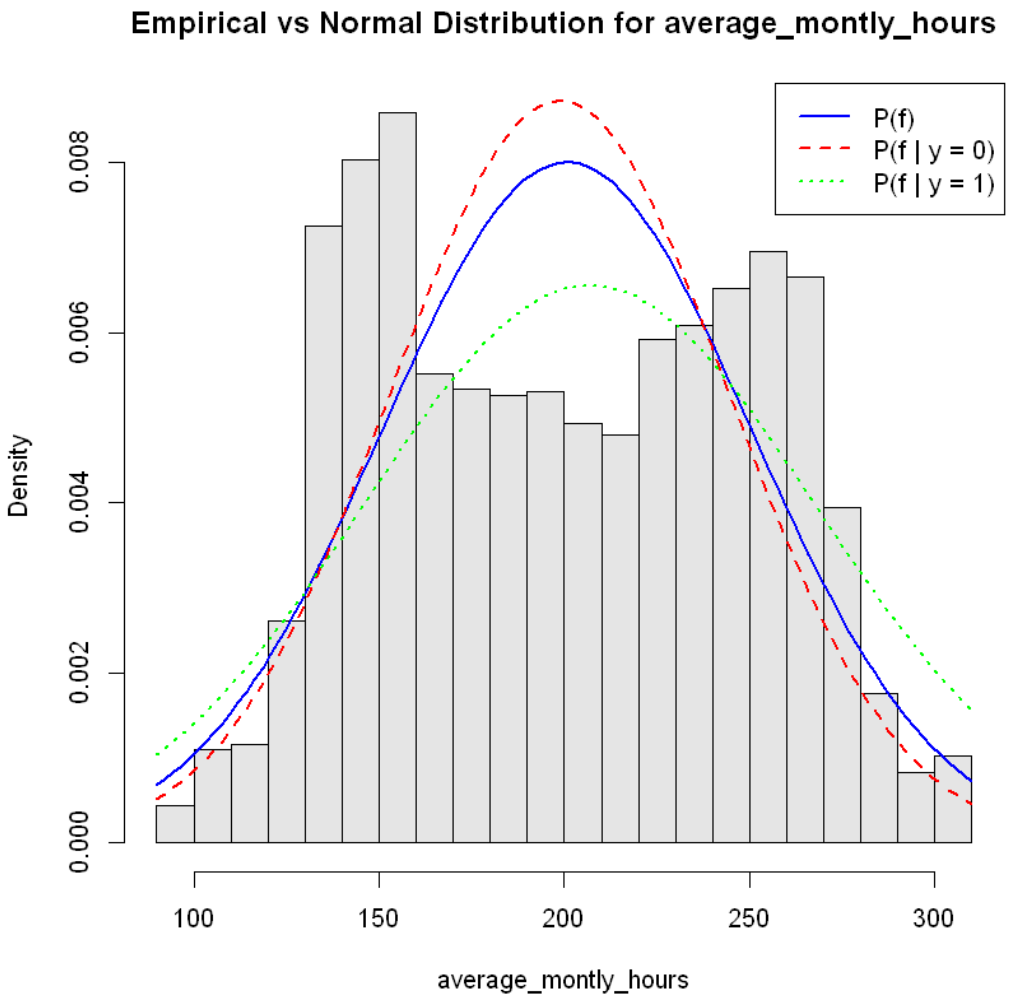

Figure 13 : Distribution empirique VS distribution normale de la variable "average_monthly_hours"

Vous trouverez ci-dessus une représentation graphique de la variable average_montly_hours. Cette représentation contient :

Tableau 6 : Analyses des valeurs théorique VS empirique pour "average_monthly_hours"

Probabilité représenté	Type	Représentation	Moyenne	Écart Type
$\boldsymbol{P(average\_monthly\_hours)}$	Empirique	Histogramme	200.69	49.84
$\boldsymbol{P(average\_monthly\_hours)}$	Théorique	Courbe bleu	200.69	49.84
$\boldsymbol{P(avg\_monthly\_hours \mid left = 0)}$	Théorique	Courbe rouge	198.77	45.70
$\boldsymbol{P(avg\_monthly\_hours \mid left = 1)}$	Théorique	Courbe verte	206.83	60.85

Analyse comparative

Est-ce la variable est normalement distribuée ?

Si l'on compare l'histogramme (valeur empirique) avec la courbe bleue, nous pouvons constater :

- À l'extrême, la valeur empirique semble suivre une loi normale
- Puis aux alentours de 150h et 250h, deux clusters semble se former avec un nombre important d'instances et une densité proche 0,008 et 0,006 les deux maximums de la variable.
- Puis autour de la moyenne (200.69), les valeurs deviennent moins courantes que celles dans les deux clusters avec une densité < 0,006

Grâce à ces observations, nous pouvons conclure que les valeurs réelles semblent suivre la loi normale théorique uniquement aux extrême et divergent complètement pour les autres valeurs.

Quid des deux autres distributions normales ?

Pour courbe rouge, la moyenne (198,77) est légèrement inférieure à celle utilisé dans la distribution globale (200,69). Les gens qui sont restés dans l'entreprises travaillent donc légèrement moins en moyenne. L'écart type (45,70) est également légèrement inférieure à celui utilisé pour la distribution globale (49,84). Ainsi les personnes étant resté dans l'entreprises semblent avoir moins de variations dans le nombre de travail et plus groupé autour de la moyenne.

Pour la courbe verte, la moyenne est légèrement supérieure (206,83). Ceci nous indique que les personnes qui démissionne semble travailler globalement un peu plus que la moyenne générale. Quant é l'écart type, il est cette fois beaucoup plus élevé (60,85) ce qui indique une grande variabilité des heures de travail dans ce groupe. Des gens de cette population semblent plus enclin à travailler de manière plus excessive ou parfois moins engagés que l'employé moyen.