Il pense que faut choisir une date (une fenetre) car les données commencent en 2007, au niveau de la crise des subprimes ce qui peut fausser le modèle. En effet, un mec qui va tomber en défaut en 2009 pouvait être dans un situation stable mais a pu perdre son emploi à cause de la crise. Avec la crise tout le monde peut être en défaut. Pour lui faut virer 2007 2008 2009 2010.

En général les banques travaillent sur 3 ans d'historiques.

Comme on a la chance d'avoir énormément de variables : retenir les plus parfaites

Modèle de scoring construit avec max 10 variables

1. Traitement des variables :

Suppression des variables :

- avec trop de valeurs manquantes ;
- incohérentes pour le sujet ;
- qui apparaissent après la tombé en défaut (durée et montant du défaut par exemple);

Isolement des crédits en cours :

application du modèle finale sur ces clients en guise d'exemple. On créé grille de score à partir des défauts et « full paid ».

Suppression des LIGNES qui ont des données aberrantes évidentes (exemple âge = 400 ans)

Comble les valeurs manquantes :

- imputation par rapport à d'autres variables
- notion métier
- ne pas inventer (ne pas essayer de prédire des valeurs)

2. Lien avec la cible :

- Au travers de corrélation : Pearson et V de cramer et autre ;
- Fixe un seuil de corrélation pour faire en sorte d'avoir pas mal de variable supprimées ;
- Virer une des variables parmi un groupe de variable extrêmement corrélées ou colinéaires (genre age et date de naissance ou ce genre de bail).

On a alors des variables corrélées avec la variables cibles, elles peuvent être de nature quali et quanti

3. Discrétisation :

Variables quali :

- Population des modalités : Si les variables quali ou trop de modalité (genre plus de 5) il faut regrouper certaines modalités. Au min 10% du total d'observation (full paid et défaut) pour qu'une modalité soit « valide ». (à vérifier pour toutes les quali au-dessus de 3 modalités) .
- Stabilité des modalités dans le temps : vérifier la stabilité des variables dans le temps (le nombre de personne dans les modalités est stables dans le temps). Sinon risque d'overfiting (ça c'est de moi). Si pas stable dans le temps, faut virer.

Variables quantitative : il faut discrétiser :

- 10% d'effectif minimum dans chaque modalité + stable dans le temps ;
- Truc en plus : il faut un sens, il faut une différence entre le fait d'être dans un classe ou dans l'autre. Pour tester la différences → variances intra classe.

On maintenant toutes les variables.

4. Logistique

Si après cette étape on a encore au-dessus de 10 variables :

- Régression logistique avec fonction de pénalité : stepwise lasso ou autre ;
- À partir des coeff on créer la grille de score ;
- (faire un diapo avec :
    - variable et modalité,
    - taux de défaut par modalités,
    - nombre de points de score que ça donne d'avoir telle ou telle modalité (être marié donne combien de points ce genre de bail),
    - effectif par modalité, contribution de la variable dans le modèle : la situation maritale du mec c'est combien de pourcentage de contribution dans le modèle (exemple si la contribution est de 25%, le fait d'avoir cette modalité donne 250points)) ;

5. Classe de risque :

- A partir des scores de chaque client on fait des classes de score (soit avec distribution soit à la main) ;
- Vérifier le nombre de personne par classe de score ;
- Classe de score doivent être stable dans le temps en termes de proba de défaut ;
- Calcul des probabilité de défaut par classe de score (on regarde le nombre de gens en défaut par classe qu'on divise par le total de la classe (niveau 5ème cette étape)) ;
- Faut surtout pas que les classes se croisent en terme de proba de défaut.

6. Métric à l'aide de la base test :

- AUC et tout le bordel

7. On peut appliquer sur les crédit en cours en guise d'exemple