Équipe 2

Travail fait par

Matis Brassard-Verrier (111 182 740)
Alyson Marquis (111 183 605)
Alexis Picard (111 182 200)
Samuel Provencher (111 181 794)

Apprentissage statistique en actuariat

ACT-3114

Rapport 1

Présenté à

Marie-Pier Côté

École d'actuariat
Université Laval
27 février 2020

Table des matières

Introduction **1**

Modèle de base **2**

Ajustement des modèles **3**

Comparaison des modèles **4**

Interprétation des meilleurs modèles **5**

Conclusion **6**

Bibliographie **7**

Introduction

Dans le cadre du travail, nous allons tenter de modéliser le prix de vente des maisons dans la région de Seattle (King County, USA) en utilisant de nombreuses caractéristiques ayant une incidence sur la valeur d'une maison. La variable réponse à prédire, soit le prix de vente d'une maison, est une valeur positive évaluée en dollars américains. La modélisation de cette variable pourrait être utile pour différentes raisons dans un contexte actuariel. Comme la somme assurée d'une maison a un lien très fortement proportionnel à son prix de vente, une compagnie d'assurance pourrait être intéressée de modéliser le prix de vente de maisons dans des nouveaux développements immobiliers afin de tenter de prédire les futures soumissions d'assurance habitation et d'offrir des offres personnalisées aux acheteurs de ces nouvelles maisons. Dans un autre contexte, au niveau de la gestion des risques, certains assureurs ont un portefeuille de prêts hypothécaires ou utilisent des produits dérivés sur prêts hypothécaires pour se couvrir du risque (*hedging*). Ainsi, il pourrait être intéressant d'avoir une estimation des montants de prêts hypothécaires dans une région donnée en se basant sur le prix de vente des maisons afin de mieux gérer le risque de la compagnie. La pertinence de trouver cette variable qu'est le prix de vente des maisons devient alors fort intéressante.

Le jeu de données utilisé sera le suivant : kc_house_sales (House sales in King County, USA). Il contient de nombreuses variables explicatives qui seront analysées dans la prochaine section.

Modèle de base

Ajustement des modèles

Comparaison des modèles

Interprétation des meilleurs modèles

Conclusion

Bibliographie

1. Kaggle, harlfoxen (2017). House sales in King County, USA. Récupéré le 27 février 2020 de https://www.kaggle.com/harlfoxem/housesalesprediction.

2. Max Kuhn (2020). caret : Classification and Regression Training. R package version 6.0-85. https://CRAN.R-project.org/package=caret

3. Terry Therneau and Beth Atkinson (2019). rpart : Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart

4. Stephen Milborrow (2019). rpart.plot : Plot 'rpart' Models : An Enhanced Version of 'plot.rpart'. R package version 3.0.8. https://CRAN.R-project.org/package=rpart.plot

5. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.

6. Brandon Greenwell, Bradley Boehmke, Jay Cunningham and GBM Developers (2019). gbm : Generalized Boosted Regression Models. R package version 2.1.5. https://CRAN.R-project.org/package=gbm

7. Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.