Kourosh KAZEMI, Flora MULON, Jenny RAHARIMANANA

LAS VEGAS

COMMENTAIRES EN LIGNE DE 21 HOTELS DE LAS VEGAS

TABLE DE DONNÉES

Las Vegas Strip Data Set :
commentaires TripAdvisor posté en 2015 concernant
21 hôtels de Las Vegas (24 commentaires par hôtel)

TABLE DE DONNÉES

20 VARIABLES

HOTEL

Nom
Etoile(s)
Revue
Chambres
Piscine
Salle de sport
Tennis
Spa
Casino
Internet

TABLE DE DONNÉES

20 VARIABLES

HOTEL

Nom
Etoile(s)
Revue
Chambres
Piscine
Salle de sport
Tennis
Spa
Casino
Internet

CLIENT

Pays
Nombre de revues
Type
Saison du séjour
Continent
Nombre d'année de membre

TABLE DE DONNÉES

20 VARIABLES

HOTEL

Nom
Etoile(s)
Revue
Chambres
Piscine
Salle de sport
Tennis
Spa
Casino
Internet

REVUE

Nombre de votes utiles
Jour et mois de publication

CLIENT

Pays
Nombre de revues
Type
Saison du séjour
Continent
Nombre d'année de membre

TABLE DE DONNÉES

20 VARIABLES

HOTEL

Nom
Etoile(s)
Revue
Chambres
Piscine
Salle de sport
Tennis
Spa
Casino
Internet

REVUE

Nombre de votes utiles
Jour et mois de publication

RÉPONSE

Note de 1 à 5 attribuées par l'utilisateur

CLIENT

Pays
Nombre de revues
Type
Saison du séjour
Continent
Nombre d'année de membre

TABLE DE DONNÉES

20 VARIABLES

Modèle Logit Ordonné

RÉPONSE

Note de 1 à 5 attribuées par l'utilisateur

TABLE DE DONNÉES

TABLE DE DONNÉES

Estimer le modèle

TABLE DE DONNÉES

454
Echantillon d'apprentissage

50
Echantillon Test

Estimer le modèle

Valider le modèle

COMMENT EXPLIQUER LE SCORE D'UN HOTEL ?

PLAN :

Construction du modèle

Interprétation du modèle

Prédiction sur l'ensemble test

A

VARIABLES

Suppression de certaines variables non pertinentes comme :

- Jour et mois de publication du commentaire
- Nombre de votes utiles sur le commentaire
- Nombre de chambres et d'étoiles de l'hôtel
- Pays de l'utilisateur

A

A CONSTRUCTION DU MODÈLE

Résultats de la sélection des variables significatives
avec les 3 méthodes :

STEPWISE

FORWARD

BACKWARD

procédure PROC LOGISTIC de SAS

A

CONSTRUCTION DU MODÈLE

Résultats de la sélection des variables significatives
avec les 3 méthodes :

STEPWISE	FORWARD	BACKWARD
Hotelname		
Traveler		
Nbreviews		
Continent		

procédure PROC LOGISTIC de SAS

A

A CONSTRUCTION DU MODÈLE

Résultats de la sélection des variables significatives
avec les 3 méthodes :

STEPWISE	FORWARD	BACKWARD
Hotelname	Hotelname	
Traveler	Traveler	
Nbreviews	Nbreviews	
Continent	Continent	

procédure PROC LOGISTIC de SAS

A CONSTRUCTION DU MODÈLE

Résultats de la sélection des variables significatives avec les 3 méthodes :

STEPWISE	FORWARD	BACKWARD
Hotelname	Hotelname	Nbreviews
Traveler	Traveler	Traveler
Nbreviews	Nbreviews	Pool
Continent	Continent	Gym
		Casino
		Internet
		Hotelname
		Continent

procédure PROC LOGISTIC de SAS

A

MEILLEUR MODÈLE

Notre modèle est donc composé de 4 variables :
Hotelname, Traveler, Nbreviews et Continent.

Score =
constante
Hotelname
Traveler
Nbreviews
Continent

MEILLEUR MODÈLE

Notre modèle est donc composé de 4 variables :
Hotelname, Traveler, Nbreviews et Continent.

Score =
constante
Hotelname
Traveler
Nbreviews
Continent

Le modèle est globalement significatif

Quelques statistiques importantes sur le modèle :

D de Somers : 0.406
Gamma : 0.408
Tau-a : 0.271
c : 0.703

2

INTERPRÉTATION DU MODÈLE

Les modalités des variables qualitatives mises en référence sont les suivantes :

- la variable HotelName, la référence est l'hôtel Wynn Las Vegas,
- la variable Traveler (type de voyage), la référence est la modalité Solo
- la variable Continent, la référence est la modalité South America

R^2 de McFadden = 0.9164

2

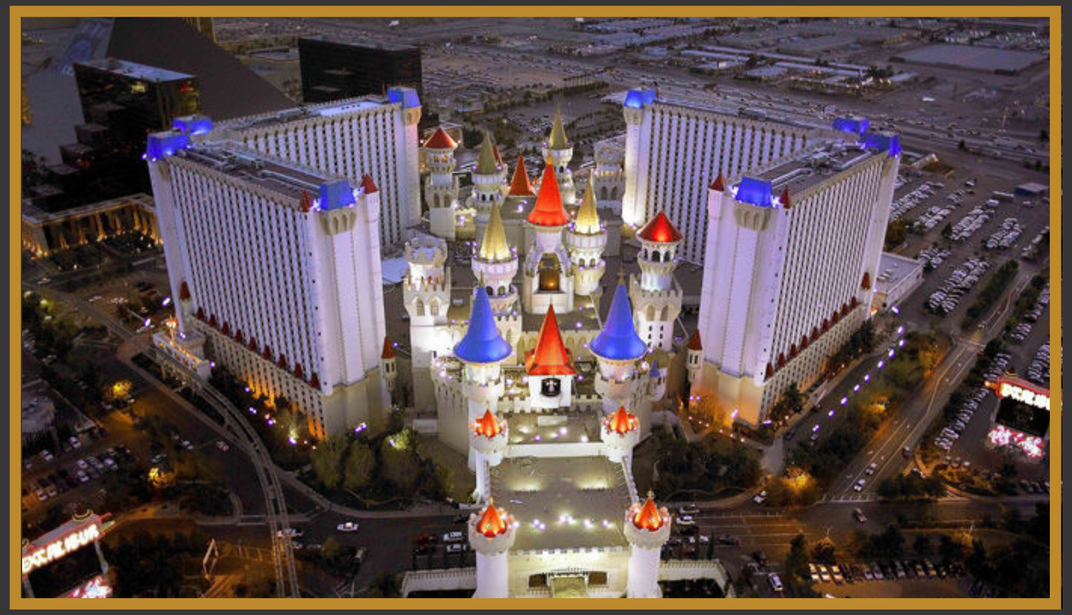
2

INTERPRÉTATION

Des signes des variables

Hotel Excalibur & Casino = 2.4482 (p-value <0.0001)

Un client ayant résidé à l'hôtel Excalibur & Casino est généralement plus satisfait qu'un client ayant résidu à l'hôtel Wynn Las Vegas.



2

2

INTERPRÉTATION

Définition des Cotes :

MODÈLE DICHOTOMIQUE SIMPLE :

$Ci = P(i=1) / 1-P(i=1)$

= probabilité de connaître l'évènement sur la probabilité de ne pas le connaître

2

2

INTERPRÉTATION

Définition des Cotes :

MODÈLE DICHOTOMIQUE SIMPLE :

$Ci = P(i=1) / 1-P(i=1)$

= probabilité de connaître l'évènement sur la probabilité de ne pas le connaître

MODÈLE MULTINOMIAL ORDONNÉ :

$Cji = P(j \leqslant m) / P(j > m)$

= probabilité que la variable y soit inférieure à une certaine modalité sur la probabilité que la variable y soit strictement supérieure à cette modalité

2

2

INTERPRÉTATION

Définition des Cotes :

MODÈLE DICHOTOMIQUE SIMPLE :

$Ci = P(i=1) / 1-P(i=1)$

= probabilité de connaître l'évènement sur la probabilité de ne pas le connaître

MODÈLE MULTINOMIAL ORDONNÉ :

$Cji = P(j \leqslant m) / P(j > m)$

= probabilité que la variable y soit inférieure à une certaine modalité sur la probabilité que la variable y soit strictement supérieure à cette modalité

Le rapport des cotes :

rapport entre la cote d'une variable sur la cote d'une autre variable

2

2

INTERPRÉTATION

Des signes des variables

RC(Hotelname Caesars Palace vs Wynn Las Vegas) = 3.846

et significatif (car l'intervalle de confiance est compris entre 1.073, et 13.784)

Un client ayant résidu à l’hôtel Caesars Palace a plus tendance a attribué une note supérieure à 1 qu’un individu ayant résidé au Wynn Las Vegas.

2

3

PRÉDICTION

Sous forme de variable dichotomique

Recodage de la variable Score

Obs.	Nbreviews	Score	Traveler	Hotelname	Continent
1	11	5	Friends	Circus Circus Hotel & Casino Las Vegas	North America
2	119	3	Business	Circus Circus Hotel & Casino Las Vegas	North America
3	36	5	Families	Circus Circus Hotel & Casino Las Vegas	North America
4	14	4	Friends	Circus Circus Hotel & Casino Las Vegas	Europe
5	5	4	Solo	Circus Circus Hotel & Casino Las Vegas	North America
6	31	3	Couples	Circus Circus Hotel & Casino Las Vegas	North America
7	45	4	Couples	Circus Circus Hotel & Casino Las Vegas	Europe
8	2	4	Families	Circus Circus Hotel & Casino Las Vegas	North America
9	24	4	Friends	Circus Circus Hotel & Casino Las Vegas	Asia
10	12	3	Families	Circus Circus Hotel & Casino Las Vegas	North America

3

PRÉDICTION

Sous forme de variable dichotomique

Recodage de la variable Score

Obs.	Nbreviews	Score	Traveler	Hotelname	Continent
1	11	1	Friends	Circus Circus Hotel & Casino Las Vegas	North America
2	119	0	Business	Circus Circus Hotel & Casino Las Vegas	North America
3	36	1	Families	Circus Circus Hotel & Casino Las Vegas	North America
4	14	1	Friends	Circus Circus Hotel & Casino Las Vegas	Europe
5	5	1	Solo	Circus Circus Hotel & Casino Las Vegas	North America
6	31	0	Couples	Circus Circus Hotel & Casino Las Vegas	North America
7	45	1	Couples	Circus Circus Hotel & Casino Las Vegas	Europe
8	2	1	Families	Circus Circus Hotel & Casino Las Vegas	North America
9	24	1	Friends	Circus Circus Hotel & Casino Las Vegas	Asia
10	12	0	Families	Circus Circus Hotel & Casino Las Vegas	North America

3

PRÉDICTION

Caractéristiques des deux échantillons :

Echantillon d'apprentissage

Moyenne : 0.4529

Ecart-type : 0.4984

Echantillon test

Moyenne : 0.44

Ecart-type : 0.4988

3

PRÉDICTION

Caractéristiques des deux échantillons :

Echantillon d'apprentissage

Moyenne : 0.4529

Ecart-type : 0.4984

Echantillon test

Moyenne : 0.44

Ecart-type : 0.4988

Test ANOVA : test d’égalité des moyennes entre l'échantillon d'Apprentissage et l'échantillon Test :

p-value du test = 0.9105

Conclusion : moyennes identiques entre les deux échantillons

3

PRÉDICTION

Caractéristiques des deux échantillons :

Echantillon d'apprentissage

Moyenne : 0.4529

Ecart-type : 0.4984

Echantillon test

Moyenne : 0.44

Ecart-type : 0.4988

Test de Levene : test d'équivalence des variances entre l'échantillons d'Apprentissage et l'échantillon Test

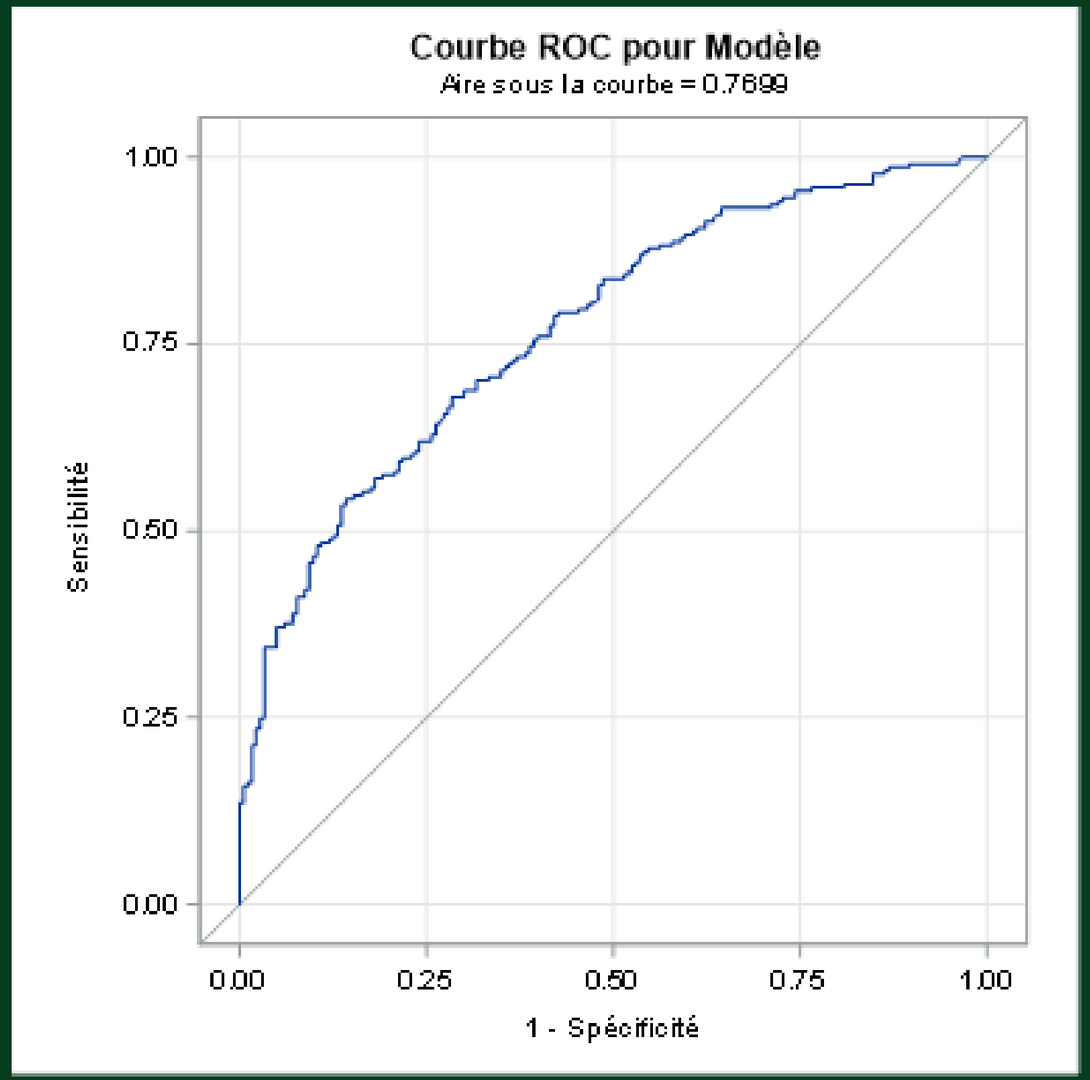
p-value du test : 0.8939

Conclusion : variances identiques entre les deux échantillons

3

PRÉDICTION

De la variable Score sur l'ensemble d'apprentissage



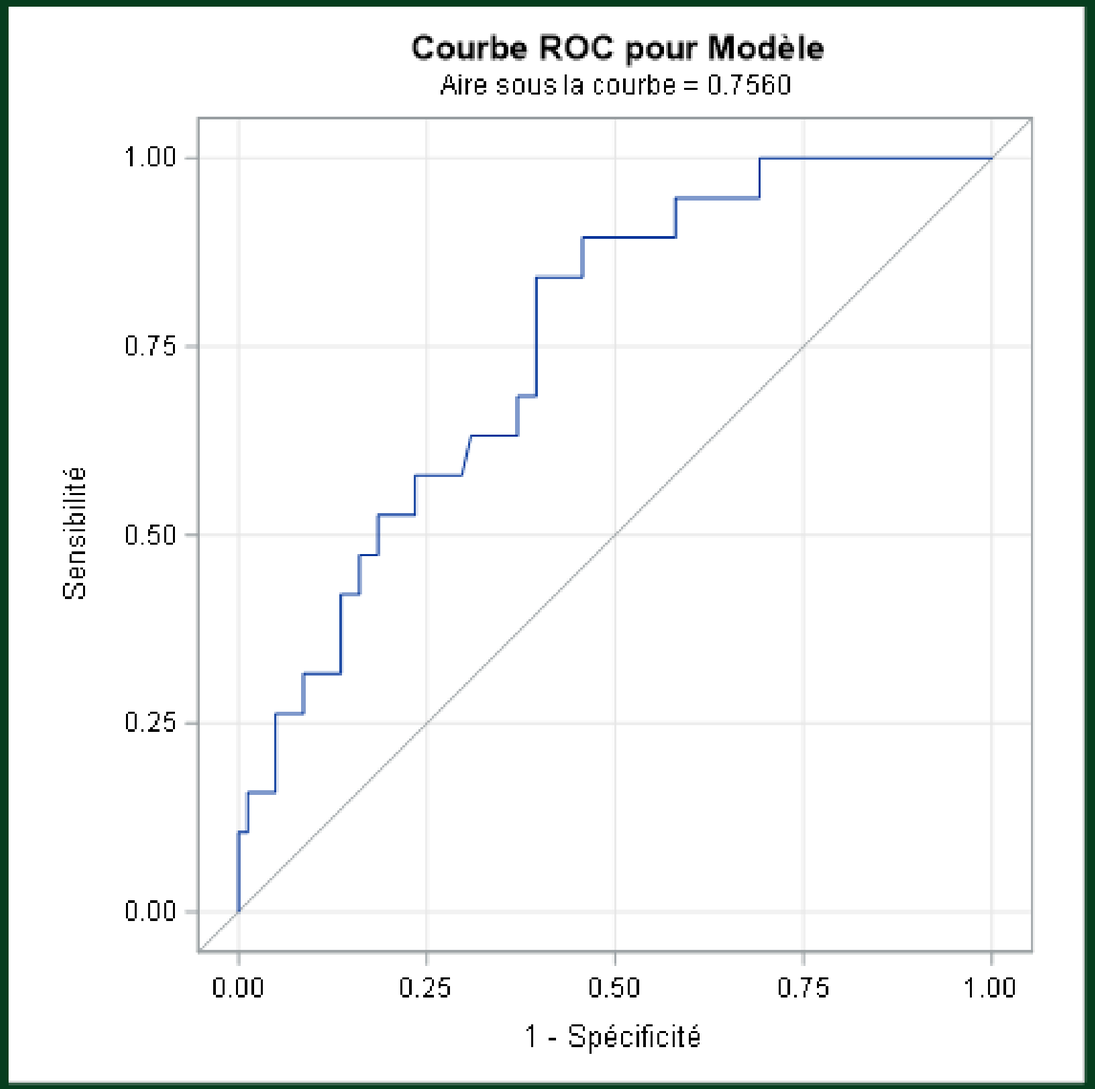
Association des probabilités prédites et des réponses observées			
Pourcentage concordant	77.0	D de Somers	0.540
Pourcentage discordant	23.0	Gamma	0.540
Pourcentage lié	0.0	Tau-a	0.268
Paires	40443	c	0.770

L'AUC est égale à la statistique c.

3

PRÉDICTION

De la variable Score sur l'ensemble TEST



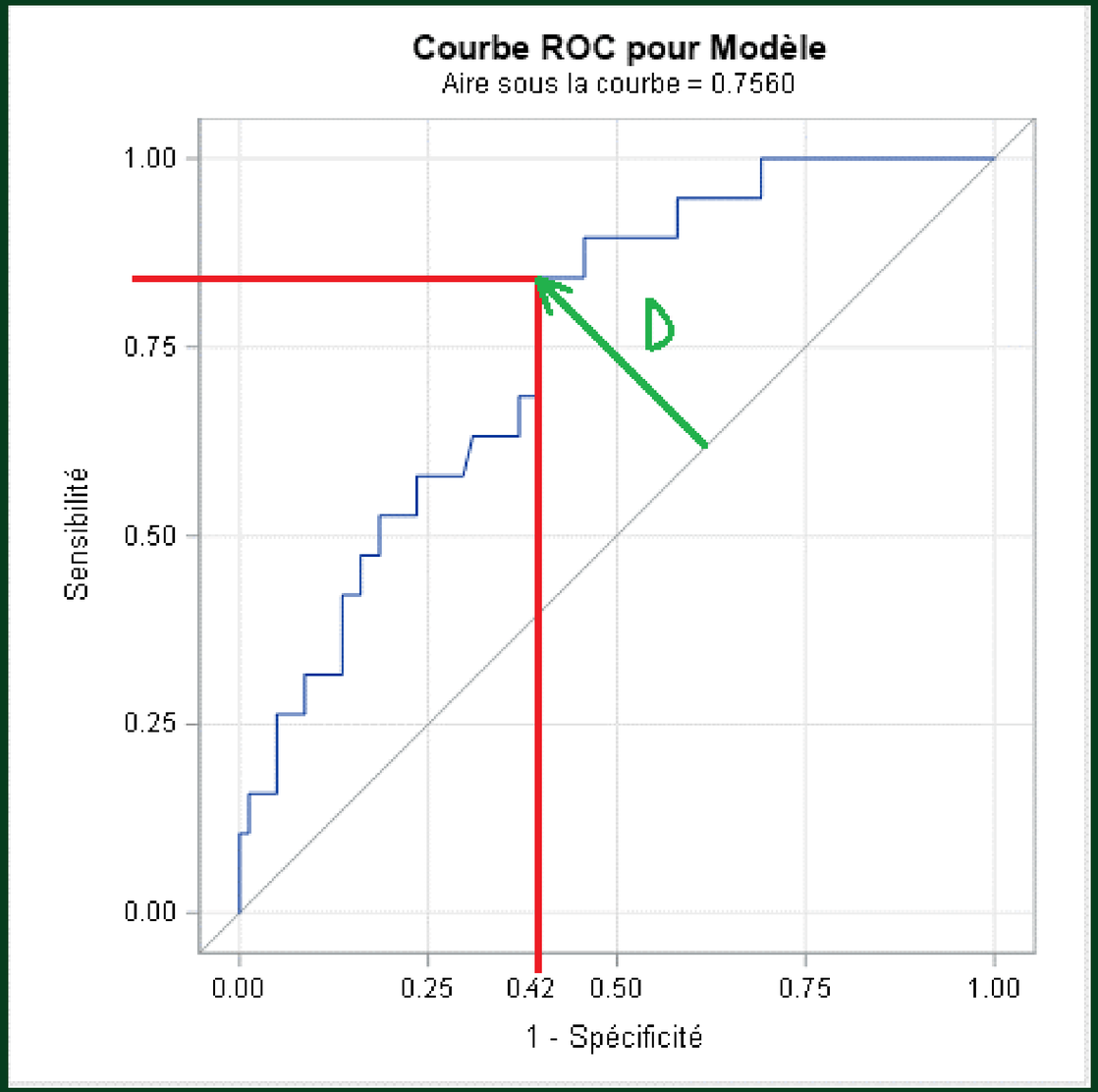
Association des probabilités prédites et des réponses observées			
Pourcentage concordant	75.1	D de Somers	0.502
Pourcentage discordant	24.8	Gamma	0.503
Pourcentage lié	0.1	Tau-a	0.250
Paires	2464	c	0.751

La différence entre les deux courbes ROC est assez faible.

3

PRÉDICTION

Déduction du Cutoff optimal du modèle



Maximisation de la distance euclidienne entre la droite à 45° et la courbe de la ROC

Obs.	_PROB_	D
1	0.21641	0.42545

Cutoff optimal = 0.42

3

PRÉDICTION

Taux de bonnes et mauvaises prédictions
avec un CUTOFF=0.42

Table de Score par PREDICTION			
Score	PREDICTION		
Fréquence Pourcentage	0	1	Total
0	37 37.00	20 20.00	57 57.00
1	0 0.00	43 43.00	43 43.00
Total	37 37.00	63 63.00	100 100.00

Sensibilité	Spécificité
1	0.6491
Faux négatifs	Faux positifs
0	0.3175

3

PRÉDICTION

Taux de bonnes et mauvaises prédictions avec un CUTOFF=0.42

Sensibilité = Proportion d’individu ayant connu l’événement et ayant était identifié comme tel par le modèle de prédiction.

Spécificité = Proportion d’individu n'ayant pas connu l’événement et ayant était identifié comme tel par le modèle de prédiction.

Sensibilité	Spécificité
1	0.6491
Faux négatifs	**Faux positifs**
0	0.3175

3

PRÉDICTION

Taux de bonnes et mauvaises prédictions avec un CUTOFF=0.42

Sensibilité = Proportion d’individu ayant connu l’événement et ayant était identifié comme tel par le modèle de prédiction.

Spécificité = Proportion d’individu n'ayant pas connu l’événement et ayant était identifié comme tel par le modèle de prédiction.

Faux positifs = Proportion d’individu identifié à tort comme ayant connu l’événement

Faux négatifs = Proportion d’individu identifié à tort comme n'ayant pas connu l’événement

Sensibilité	Spécificité
1	0.6491
Faux négatifs	**Faux positifs**
0	0.3175

3

PRÉDICTION

Taux de bonnes et mauvaises prédictions
avec un CUTOFF=0.42

Taux Mauvaises Classification	Taux Bonnes Classification
0.27970	0.72030

72% de bonnes classifications, ce qui nous permet de conclure que le modèle construit génère de bonnes prévisions.

CONCLUSION :

Même si le coefficient de détermination de notre modèle n'est pas très élevé (R^2 = 0,56), nous avons réussi à construire un modèle explicatif du score attribué à 24 hotels de Las Vegas sur le site TripAdvisor.

CONCLUSION :

Même si le coefficient de détermination de notre modèle n'est pas très élevé (R^2 = 0,56), nous avons réussi à construire un modèle explicatif du score attribué à 24 hotels de Las Vegas sur le site TripAdvisor.

On peut donc expliquer une note élevée par le nom de l'hôtel, le profil et le continent du voyageur, ainsi que le nombre de revues de l'hôtel sur le site.

CONCLUSION :

Conseil pour obtenir de meilleures notes sur TripAdvisor :

- améliorer les séjours des couples ou de ceux qui voyagent entre amis
- attirer davantage les clients avec un profil Business et familles