CUNEF
COLEGIO UNIVERSITARIO DE ESTUDIOS FINANCIEROS

FINANCIAL [THINKING]

Sistema de Reservas de Hotel
Informe Predicción 2:
Modelo de Regresión

Mayra Goicochea Neyra

CONTENIDO

ABSTRACT

Las cancelaciones de reservas en la industria hotelera no solo generan pérdida de ingresos y afectan las decisiones de asignación de precios e inventario, sino que también, en situaciones de sobreventa, tienen el potencial de afectar la reputación social del hotel.

Con el dataset de un hotel tipo resort de la ciudad de Algarve en Portugal, se aborda este inconveniente como un problema de clasificación en el ámbito de Data Science.

Un modelo de regresión logística resuelve este problema mediante la estimación de probabilidades de que la variable objetivo resulte determinada clase (en este caso, que la reserva sea una cancelación)

En el presente informe, se resumen, la construcción de modelos predictivos de regresión logística, su validación ante un subconjunto de prueba y la selección de uno de ellos eficiente para el caso del hotel de Algarve. Para la realización de estas actividades, se utilizaron, además de modelos de regresión logística, técnicas de regularización (Ridge, Elastic Net, Lasso), métodos de selección Stepwise y Cross Validation para realizar una comparación fina.

INTRODUCCIÓN

Vender la habitación *correcta* al cliente *correcto* en el momento *correcto* por el precio *correcto* es el desafío en la industria hotelera. Con las estrategias de gestión de ingresos (RM), los hoteles intentan optimizar sus ingresos con, por ejemplo, precios dinámicos y asignación (Talluri y van Ryzin [1]). La forma clásica de RM en la industria hotelera es vender un número fijo (la capacidad) de habitaciones, que son perecederas en un plazo fijo (el horizonte de reserva). Basado en reservas históricas, información de mercado, información de huéspedes y más información disponible, los hoteles eligen controles óptimos en forma de precios dinámicos y asignación de capacidad para maximizar sus ingresos. Estos controles son la configuración de precios y las disponibilidades para varios tipos de habitaciones.

En el mundo de hoy, los hoteles ofrecen tarifas reembolsables, sin depósito y no reembolsables a los huéspedes. Recientemente, hay un mayor interés en las tarifas reembolsables y sin depósito, donde los huéspedes aún tienen la posibilidad de cancelar (inclusive hasta último minuto). Mientras que los huéspedes valoran la flexibilidad, los hoteles se enfrentan al riesgo de habitaciones vacías y, por lo tanto, a la pérdida de ingresos, lo cual es un problema para la industria.

Las altas tasas de cancelación pueden conducir a la consiguiente pérdida de ingresos debido a habitaciones vacías. Con las cancelaciones de último minuto y los "no-shows", la asignación de capacidad ya no es óptima porque los hoteles no logran atraer invitados con tan poca antelación.

Según el sitio web HotelManagement.net, sus estudios revelaron que la tasa de cancelaciones ha incrementado significativamente, se tenía una tasa promedio de 32.9% en el 2014 y en el 2017, ha alcanzado a ser de 41.3% [1]

En este informe, se presenta un prototipo basado en técnicas de construcción de modelos predictivos, por modelos predictivos solo se basaron en los de regresión (exclusivamente regresión logística). Mediante herramientas como regularización, Selección de Variables Stepwise y Cross Validation, se ajustaron los modelos para obtener finalmente un modelo eficiente a un nivel de 0.834% de accuracy.

OBJETIVO

El prototipo busca resolver la problemática de cancelación en las reservas, mediante la predicción de cuan probable es que una reserva sea cancelada.

PARTICION DE LAS MUESTRAS DE ENTRENAMIENTO Y PRUEBAS

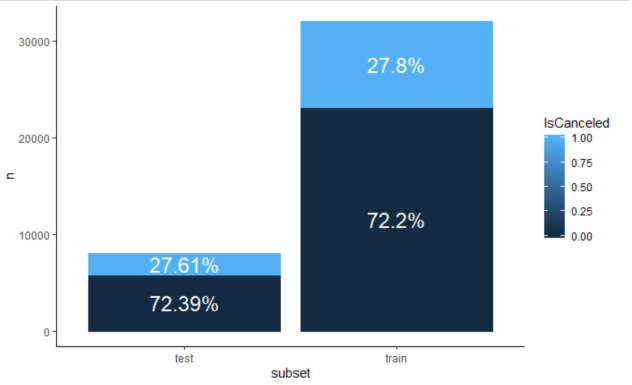
El conjunto de datos contiene 40060 observaciones de reservas, seleccionadas según

su fecha de arribo, en el periodo del 1 de julio del 2015 al 31 de agosto del 2017.

Se repartieron en dos subconjuntos (con una proporción de 80%-20%) para realizar los siguientes objetivos (Ver Fig.1):

- Entrenamiento (Train): se utilizó para la construcción de los modelos. Contiene 32048 registros (con un 27.8% de reservas canceladas y un 72.2% de no canceladas)
- Pruebas (Test): se utilizó para la evaluación de los modelos. Contiene 8012 registros (con un 27.61% de reservas canceladas y un 72.39% de no canceladas)

Fig: 1.



MODELAMIENTO

La regresión logística es un tipo de modelo de regresión donde la variable dependiente es categórica (recordemos que la variable objetivo es binaria '1' y '0', y representa si la reserva se cancela sí / no). Este modelo estima la probabilidad de la respuesta de la variable objetivo, y aquí recae su ventaja, porque al tener probabilidades se puede determinar un borde (que llamaremos threshold) y este umbral determinara el valor de la variable (0/1). Este umbral debe considerar la política del hotel (si desea colocar más peso a los casos de cancelaciones no identificadas o cancelaciones falsas).

El primer modelo incluye todas las variables (a excepción de la variable ArrivalDateYear porque es un dato que no será replicable para las nuevas reservas), se obtuvo un AIC de 340498 (muy alto). El coeficiente estimado para la intersección es el valor esperado del logaritmo de odds de que una reserva sea cancelada (teniendo todas las otras variables en 0). Los odds son muy bajos $e^{(-3.992*e+12)}$, por lo que la probabilidad de cancelar una reserva es: $p = \frac{e^{(-3.992*e+12)}}{2a1+e^{(-3.992*e+12)}}$

Acorde al modelo, el logaritmo de los odds de que una reserva sea cancelada esta positivamente relacionada al PreviousCancellations (con un coeficiente de 3.3487), mientras que esta negativamente relacionada a la variable IsRepeatedGuest (coeficiente de -1.8457), lo cual guarda sentido, la mayoría de usuarios que han tenido cancelaciones son más probables de cancelar, en cambio, los clientes antiguos no lo son. (Ver detalle en Fig. 2)

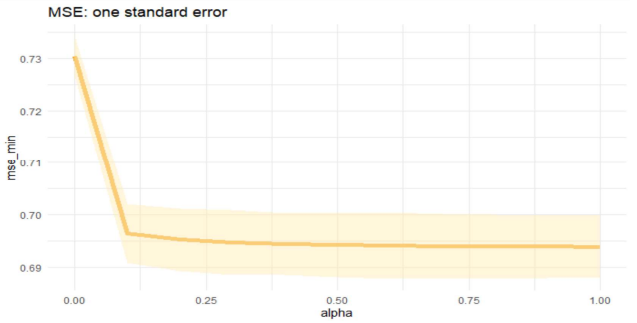
Fig: 2.

IsRepeatedGuest	-1.9424e+00	1.8949e-01	-10.2505	< 2.2e-16
PreviousCancellations	3.4662e+00	1.8117e-01	19.1320	< 2.2e-16

Como se observan algunos predictores son más relevantes, por lo que se ajustó el modelo con la selección de variables de ANOVA con la prueba del Chi-square. El modelo resultante tuvo un AIC de 25431.5 (mejor al modelo inicial).

Después, se ajustó el modelo inicial con técnicas de regularización. Para la selección del alpha se realizó un cross validation con diferentes valores de alpha para elegir el que menor error estándar tiene. (Ver Fig. 3)

Fig: 3.



Se observa que el menor error se encuentra en alpha=1, que se trata de un lasso).

Finalmente, se aplicó al modelo inicial, las técnicas de selección de variables Stepwise. Se evaluaron los tres modos "Backward", "Forward" y "Both Direction", obteniendo tres modelos con AIC 37939(Backward), 22461(Forward) y 22461(Both).

RESULTADOS Y SELECCIÓN DE MODELO

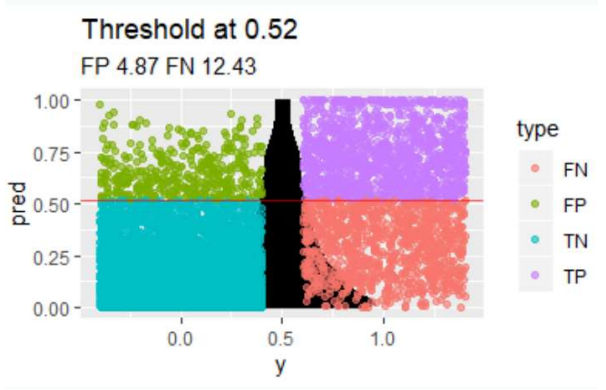
La Matriz de confusión permite comparar los valores de la muestra de pruebas (Test) y la clasificación de la predicción dependiendo del umbral.

La asignación del umbral depende de la estrategia del negocio, por ejemplo, si prefieren tener cuidado en distinguir las cancelaciones, esto puede afectar la imagen del negocio porque para reducir el error tendrían que dejar de confirmar reservas con más frecuencia a posibles clientes (muy radical), o si prefieren que sea de forma equitativa. Para la comparación de modelos se consideró la segunda opción (de forma equitativa los pesos). Con la librería "InformationValue" y la función OptimalCutoff se estableció los umbrales óptimos para cada modelo.

El modelo 2 de la reducción mediante la prueba ChiSquare obtuvo los siguientes resultados (Ver Fig. 4):

		Test	
		No	Si
Predicción	No	5116	1317
	Si	622	957

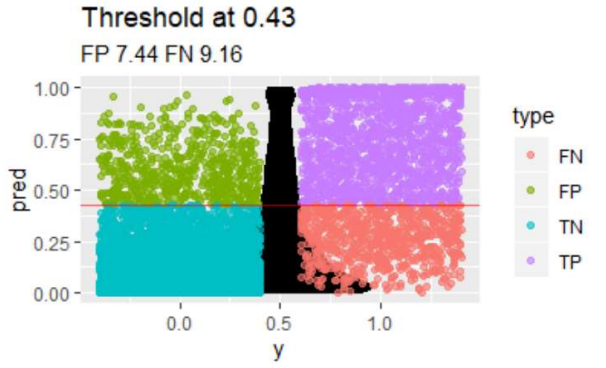
Fig: 4.



El modelo 3 de Regularización Lasso obtuvo los siguientes resultados (Ver Fig. 5):

		Test	
		No	Si
Predicción	No	5142	734
	Si	596	1540

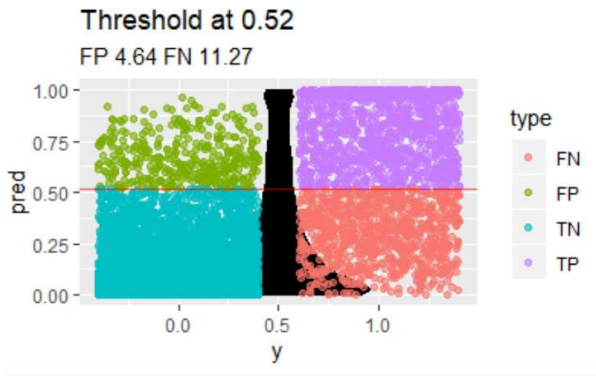
Fig: 5.



El modelo 4 de Stepwise Both y el modelo 5 de Stepwise Forward obtuvieron los mismos resultados (Ver Fig. 6):

		Test	
		No	Si
Predicción	No	5100	1258
	Si	638	1016

Fig: 6.



La tabla comparativa de los modelos muestra que el modelo con mejor Accuracy es el modelo Regularizado (Ver Fig.7)

Fig: 7.

Method	Accuracy	Precision	Recall	FScore
Full Model	0.683	0.105	0.015	0.027
Model Reducido ChiSq	0.758	0.606	0.421	0.497
Model Regularizado	0.834	0.721	0.677	0.698
Stepwise Both	0.763	0.614	0.447	0.517
Stepwise Forward	0.763	0.614	0.447	0.517

La curva ROC de los modelos nos muestra que a pesar del Regularizado tener buen accuracy, el mejor es el Stepwise Both por tener un AUC de 0.898. (Ver Fig.8)

CONCLUSIONES

- El modelo de regresión logística es eficiente para casos donde la variable es categórica y además permite tener un dataset mixto (variables numéricas y variables categóricas)

- Las técnicas de Regularizacion y Stepwise encuentran modelos óptimos dado que pueden ejecutarse mediante CrossValidation para evaluar los errores medios y obtener un mejor modelo. En este caso, se consideró a la métrica AIC.

REFERENCIAS BIBLIOGRAFICAS

- [1] Study: Cancellation rate at 40% as OTAs push free change policy, HotelManagement.net (link)
- [2] Glmnet Vignette (Regression with Regularization), (link)
- [3] Bank Marketing Prediction by Vikas Mann, (link)
- [4] Modern Regression Techniques Using R, Daniel B. Wright and Kamala London.

ANEXO

- Archivo RMarkdown en Github (MasterDS2019>Prediccion>Research>PRED-Examen20-Informe2MGN.Rmd).

Fig: 7.