UNIVERSIDAD DE LA FRONTERA

Junio 29, 2022

HOTEL BOOKING

Giselle Acuña
Bastian Barrientos
Alex Muñoz

Problema

- El equipo de T4 fue contratado para realizar el análisis de un set de datos de reservas hoteleras, del cual se espera obtener la mayor información sobre el comportamiento de los huéspedes que reservan y cómo optimizar la ganancia de los hoteles.

Estructura

Reserva de Hotel

2 DataSets con la misma estructura

-> Resort Hotel (H1) - 40,060 Observaciones
-> City Hotel (H2) - 79,330 Observaciones

119,390 observaciones
36 variables

Cada observacion representa una reserva
de hotel realizada entre
01 de Julio 2015 ~ 31 de Agosto 2017

- ¿Cómo fueron adquiridos los datos?:
  Extracción de las bases de datos SQL del sistema de gestión de propiedades (PMS) de los hoteles.

- Ubicación de la fuente de datos:
  Ambos hoteles están ubicados en Portugal:
  H1 en la región turística de Algarve y
  H2 en la ciudad de Lisboa.

- Formato de los datos:
  Mixto (raw and preprocessed).

Relevancia y Motivacion

- Predicción de la cancelación de una reserva.
- Segmentación de clientes.
- Satisfacción de clientes.
- Estacionalidad.

Objetivos

Predicción de la cancelación de una reserva.

- Analisis exploratorio
- Modelos de clasificacion supervisada

Arbol de decision
Naïve Bayes
Support Vector Machines

Preguntas y problemas

Preguntas

- ¿De qué manera se podría predecir la cancelación de una reserva por parte del cliente a partir de los datos disponibles? ¿Qué beneficios aportará a los hoteles esta predicción?

- ¿Que tipo de cliente (adultos, adultos con niños, adultos con bebes, etc) produce una mayor tarifa diaria para los hoteles (campo adr)? ¿Qué beneficios aportaría a los hoteles esta información?

- ¿Qué tipo de cliente tiene un tiempo de llegada mayor luego de realizar la reserva a un hotel? ¿Sería de utilidad dicha información para los hoteles?

- ¿A qué tipo de cliente deberían dirigirse los hoteles para mejorar sus ingresos?

Datos de interés

Campos de interés:

- adr: Tarifa media diaria (calculada dividiendo la suma de todas las transacciones de alojamiento por el número total de noches de estancia).
- adults: Número de adultos en la reserva.
- children: Número de niños en la reserva.
- babies: Número de bebés en la reserva.
- is_canceled: Valor que indica si la reserva fue cancelada (1) o no (0).
- lead_time: Número de días transcurridos entre la fecha de entrada de la reserva en el PMS y la fecha de llegada.

Preprocesamiento

Atributos eliminados:

- Company
- Agent

Atributos 'arreglados':

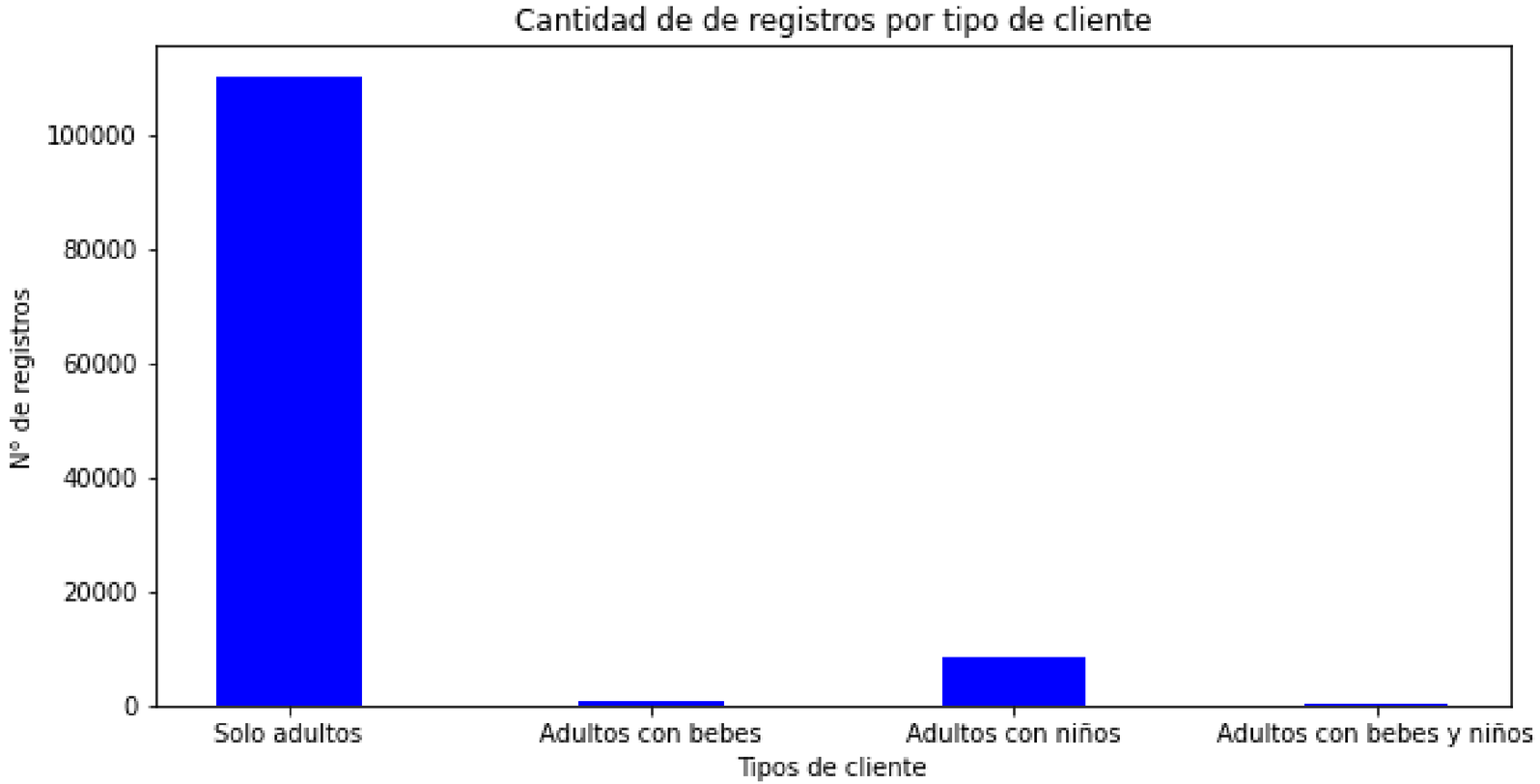
- Country
- Children

Atributos creados

- Solo adultos
- Adultos con niños
- Adultos con bebés
- Adultos con bebés y niños.

	adults	children	babies	adult_only	adultWithBabies	adultWithChildrens	AdultWithSons
0	2	0.0	0	1	0	0	0
1	2	0.0	0	1	0	0	0
2	1	0.0	0	1	0	0	0
3	1	0.0	0	1	0	0	0
4	2	0.0	0	1	0	0	0

Frecuencia tipo de



Estadísticas tipo de cliente

Promedio de tarifa diaria solo adultos: 97.37077704483092
Promedio de tarifa diaria adultos con bebes: 112.22223719676549
Promedio de tarifa diaria adultos con niños: 158.20546405228757
Promedio de tarifa diaria adultos con bebes y niños: 152.09565714285714

Promedio de dias de llegada solo adultos: 105.3226026277054
Promedio de dias de llegada adultos con bebes: 79.75471698113208
Promedio de dias de llegada adultos con niños: 89.73761140819964
Promedio de dias de llegada adultos con bebes y niños: 68.61714285714285

Propuesta experimental

¿De qué manera se podría predecir la cancelación de una reserva por parte del cliente a partir de los datos disponibles? ¿Qué beneficios aportará a los hoteles esta predicción?

- Arbol de decision
- Naïve Bayes
- SVM
- KNN (descartado)

Propuesta experimental

```
Cantidad de instancias por clase en el conjunto de entrenamiento:
Clase 1: 22112
Clase 0: 37583
```

Balance de clases:

- Conjunto de entrenamiento original.
- Conjunto de entrenamiento Sobre-muestreado
- Conjunto de entrenamiento Sub-muestreado

Resultado preliminar

Datos originales
Puntaje de métricas Árbol de decisión con datos originales:
Accuracy: 0.7932992712957534
Precision: 0.7181960258986381
Recall: 0.7273878437047757
F1: 0.722762711483587

Datos Sobre muestreados
Puntaje de métricas Árbol de decisión con datos sobre muestreados:
Accuracy: 0.7907027389228578
Precision: 0.7144003566651805
Recall: 0.7246743849493488
F1: 0.7195006959723408

Datos sub muestreados
Puntaje de métricas Árbol de decisión con datos sub muestreados:
Accuracy: 0.7686070860206048
Precision: 0.6600208244957773
Recall: 0.7740141099855282
F1: 0.7124867306371376

Resultado preliminar

Datos originales
Puntaje de métricas modelo Naïve Bayes con datos originales:
Accuracy: 0.4786330513443337
Precision: 0.4125807640815693
Recall: 0.9616497829232996
F1: 0.577425968418623

Datos Sobre muestreados
Puntaje de métricas modelo Naïve Bayes con datos sobre muestreados:
Accuracy: 0.4762375408325655 3
Precision: 0.411613625830372 3
Recall: 0.9639562228654125
F1: 0.5768918480025982

Datos sub muestreados
Puntaje de métricas modelo Naïve Bayes con datos sub muestreados:
Accuracy: 0.475081665131083
Precision: 0.4111926358156643
Recall: 0.965629522431259
F1: 0.5767771039587244

Resultado preliminar

Puntaje de métricas modelo Support Vector Machines con datos originales:
Accuracy: 0.4103023703827791
Precision: 0.384803576457336
Recall: 0.9887391461649783
F1: 0.553998580985201 7

Puntaje de métricas modelo Support Vector Machines con datos sobre muestreados:
Accuracy: 0.4344249937180668 7
Precision: 0.394256253857604 5
Recall: 0.9821816208393632
F1: 0.562657063654498 9

Datos sub muestreados
Puntaje de métricas modelo Support Vector Machines con datos sub muestreados:
Accuracy: 0.6365022196163833
Precision: 0.9578713968957872
Recall: 0.01953690303907380 5
F1: 0.038292780215396886

Conclusiones [Analisis]

- **Un modelo de clasificación supervisada puede predecir la cancelación de una reserva.**

- **A partir del analisis del comportamiento de cada cliente, los hoteles deberian enfocarse en atraer adultos acompañados de niños, bebes o ambos. (En general otorgan mayor ganancia y un tiempo menor de llegada).**

Conclusiones [Equipo]

- **Las preguntas o problemáticas que se plantearon en general eran muy simples, debido a que 3 de 4 preguntas se pudieron responder en el EDA.**

- **Se pudieron haber evaluado más modelos de clasificación sino fuera por una mala gestión del tiempo.**

Trabajos futuros

- **Una mejora en los modelos de clasificación se puede realizar a través de seleccionar los mejores hiperparámetros para cada modelo, de esta manera se puede mejorar el puntaje en las métricas.**

- **Se pueden eliminar los valores outliers o ruido encontrados en el diagrama de cajas y bigotes, lo cual queda para un trabajo futuro con tal de ver si influye en los resultados de las métricas.**