LAPORAN
DATA ANALYTICS COMPETITION
FIND IT! 2020

Tim Agriminer V6.0

Disusun oleh:

Agus Hendra Nasution
Julius Christoper Siswanto
Muhammad Yusuf Sulaiman

INSTITUT PERTANIAN BOGOR

BOGOR

2020

BAB I
LATAR BELAKANG

Pada era digital jaman sekarang ini banyak kemudahan yang ditawarkan kepada konsumen, salah satunya *booking* hotel secara *online*. Dalam pemesanan kamar hotel ada sebuah kontrak (kesepakatan) antara pelanggan dengan pihak hotel. Kontrak ini memberi pelanggan hak untuk menggunakan layanan di masa depan dengan harga yang telah disepakati, biasanya kontrak ini diberikan dengan opsi untuk membatalkan kontrak sebelum layanan didapatkan. Dalam kasus industri perhotelan, opsi dalam pembatalan layanan menempatkan risiko pada hotel, karena hotel harus menjamin ketersedian kamar untuk pelanggan yang telah melakukan pemesanan, tetapi pada saat yang sama hotel harus menanggung biaya peluang untuk kamar yang kosong akibat adanya pelanggan yang membatalkan pesanan atau tidak muncul (Talluri dan Van Ryzin, 2004). Ada beberapa perbedaan antara *no-show* dan pembatalan, untuk keperluan penelitian ini keduanya akan diperlakukan sebagai pembatalan. Pembatalan terjadi ketika pelanggan mengakhiri kontrak sebelum kedatangannya dan *no-show* terjadi ketika pelanggan tidak menginformasikan hotel dan gagal *check-in*.

Pembatalan pemesanan biasanya terjadi dengan alasan yang dapat diterima, seperti: perubahan agenda pertemuan bisnis, liburan yang dijadwalkan ulang, sakit, kondisi cuaca buruk dan faktor lainnya. Tapi, seperti yang diidentifikasi oleh Chen, Schwartz, dan Vargas (2011), mereka memaparkan bahwa saat ini, sebagian besar pembatalan terjadi karena pelanggan mencari kesepakatan ideal dalam mencari penawaran terbaik. Terkadang, pelanggan terus mencari penawaran yang lebih baik dari produk / layanan yang sama dalam melakukan pemesanan. Pada beberapa kasus, pelanggan bahkan membuat banyak pemesanan untuk mendapatkan harga terbaik dan kemudian membatalkan semua pesanan kecuali satu. Reservasi dengan opsi pembatalan memberi pelanggan kebebasan untuk mendapatkan harga terbaik, selain itu manfaat dari opsi pembatalan adalah memiliki kebebasan untuk menjadwal ulang apakah rencana atau preferensi mereka berubah .

Sebagai cara untuk mengelola risiko yang terkait dengan pembatalan pemesanan, hotel menerapkan kombinasi kebijakan *overbooking* dan pembatalan

pemesanan (Mehrotra dan Ruttley, 2006; Smith et al., 2015; Talluri dan Van Ryzin, 2004). Namun, kebijakan *overbooking* dan pembatalan yang dilakukan dapat merugikan hotel. *Overbooking*, dengan tidak mengizinkan pelanggan untuk *check-in* di hotel yang sebelumnya dipesan, memaksa hotel untuk menolak penyediaan layanan kepada pelanggan, hal ini bisa menjadi pengalaman yang buruk bagi pelanggan. Pengalaman ini dapat memberikan pengaruh negatif terhadap reputasi dan pendapatan hotel (Noone dan Lee, 2011), belum lagi potensi hilangnya pendapatan di masa mendatang karena pelanggan merasa tidak puas dan tidak akan memesan lagi untuk menginap di hotel (Mehrotra dan Ruttley, 2006). Kebijakan pembatalan, terutama kebijakan yang tidak dapat dikembalikan, memiliki potensi untuk tidak hanya mengurangi jumlah pemesanan, tetapi juga untuk mengurangi pendapatan karena diskon signifikan terhadap harga (Smith et al., 2015). Namun ada pemesanan yang dibatalkan hingga mencapai 20% dari total pemesanan (Morales & Wang, 2010). Pada hotel yang berada di dekat bandara / pinggir jalan jumlah ini dapat naik hingga 60% (Liu, 2004).

Agar menghindari kerugian akibat adanya pembatalan pesanan, hotel biasanya sering menjual di atas kapasitasnya (pemesanan berlebih) (Ivanov dan Zhechev, 2012; Mehrotra dan Ruttley, 2006; Morales dan Wang, 2010). Namun, pemindahan kamar hotel juga dapat menghasilkan keuntungan bagi hotel (Hayes dan Miller, 2011; Mehrotra dan Ruttley, 2006): realokasi pelanggan ke hotel-hotel alternatif, kompensasi tunai, atau reputasi sosial dapat digunakan sebagai media untuk menjamin tamu hotel dapat menginap kembali di masa depan.

BAB II
TUJUAN DAN MANFAAT

Tujuan :

1. Mengidentifikasi fitur dari *database* PMS dengan kemampuan prediktif mengenai probabilitas pembatalan pemesanan;
2. Membangun model yang dapat memprediksi sebab-sebab pemesanan kamar dibatalkan; dan
3. Memahami jika satu model prediksi cocok untuk semua hotel atau jika model spesifik harus dibangun untuk setiap hotel.

Manfaat:

1. Mengetahui gambaran umum dari kondisi hotel di Portugal sehingga dapat dilakukan pemodelan yang bisa membantu pihak hotel untuk mengetahui tamu mana saja yang berpotensi membatalkan pemesanan hotel; dan
2. Sebagai bentuk media pembelajaran dalam melakukan pemodelan dari sebuah data.

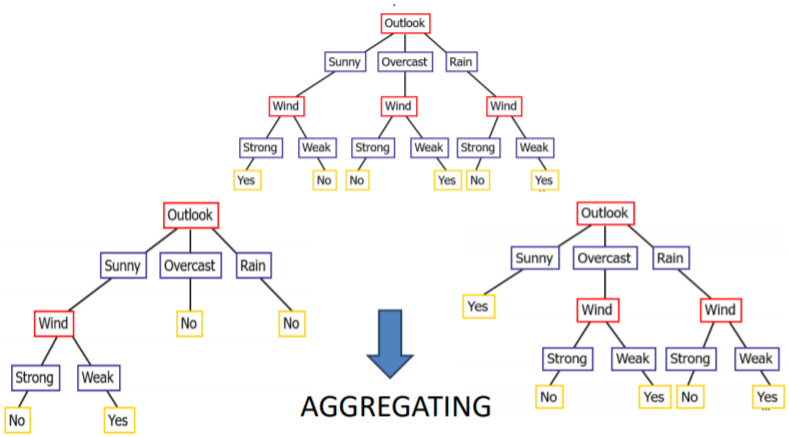
BAB III
METODE ANALISIS DATA

a. Klasifikasi

Klasifikasi pada dasarnya merupakan proses menilai objek dan pengelompokan peubah ke dalam kelas yang sudah ditentukan (Linoff dan Berry 2011). Target yang akan diprediksi dalam klasifikasi merupakan peubah yang bersifat kategorik. Metode klasifikasi dibedakan ke dalam dua kelompok, yaitu non parametrik dan parametrik. Algoritma klasifikasi non parametrik tidak membutuhkan asumsi distribusi. Ada berbagai macam teknik klasifikasi yang termasuk ke dalam kelompok non-parametrik, antara lain CART, *Neural Network*, *Support Vector Machine*, dan *Random Forest*, sedangkan yang termasuk ke dalam kelompok parametrik antara lain regresi logistik dan analisis diskriminan. Terdapat dua tahapan atau proses dalam melakukan klasifikasi, yaitu proses pelatihan dan proses pengujian. Pada proses pelatihan, model dibangun dengan menggunakan data latih yang telah diketahui labelnya, sedangkan proses pengujian dilakukan untuk mengetahui kinerja model yang telah dibangun dengan menggunakan data uji.

b. *Bagging*

Bagging merupakan merupakan salah satu metode *ensemble* dengan cara menggabungkan antara proses *bootstrapping* dan *aggregating* dalam metode pohon klasifikasi. *Bootstrapping* adalah pengambilan data contoh dari sampel yang dimiliki (*resampling*). *Bootstrap aggregating* atau *bagging* adalah suatu prosedur yang bertujuan untuk mengurangi keragaman yang dihasilkan dari metode statistika yang digunakan. Untuk kasus klasifikasi digunakan *majority vote* atau pemilihan suara terbanyak, sedangkan untuk masalah regresi digunakan rata-rata. Penggunaan *bagging* ini sangat membantu terutama mengatasi sifat ketidakstabilan pohon klasifikasi. Proses *bagging* dapat mengurangi galat baku dugaan yang dihasilkan oleh pohon tunggal (Hastie et al, 2008). *Bagging* diilustrasikan seperti pada Gambar 1.



Gambar 1. Ilustrasi Bagging

Random Forest (RF) adalah suatu algoritma yang digunakan pada klasifikasi data dalam jumlah yang besar. Klasifikasi random forest dilakukan melalui penggabungan pohon (*tree*) dengan melakukan *training* pada sampel data yang dimiliki. *Random Forest* adalah pohon dari metode *bagging* yang menggunakan pohon-pohon yang tidak berkorelasi. Sama seperti *bagging*, dibentuk banyak pohon keputusan saat proses *bootstrap* (James et al, 2003).

Oleh karena itu, metode penggunaan metode *Random Forest* ini berupaya untuk memperbaiki proses pendugaan yang dilakukan menggunakan metode *bagging* (Breiman, 2001). Penggunaan pohon (*tree*) yang semakin banyak akan mempengaruhi akurasi yang akan didapatkan menjadi lebih baik. Penentuan klasifikasi dengan *Random Forest* diambil berdasarkan hasil voting dari *tree* yang terbentuk.

c. *Extra Trees*

Klasifikasi dengan metode *Extra Trees* atau yang disebut juga sebagai *Extremely Randomized Trees* merupakan varian pengembangan dari *decision tree* acak pada berbagai sub bagian dataset dan menghitung rata-ratanya untuk meningkatkan akurasi prediksi dan pengendalian *overfitting*. Berbeda dengan *Random Forest* yang pada setiap tahapannya, sampel dan keputusan diambil secara acak dan bukan diambil dari yang terbaik, *Extra Trees* membangun *group decision tree* sesuai dengan prosedur *top-down*. Selanjutnya, lakukan pemodelan untuk masing-masing model yang dibangun dari n data latih hasil *V-Cross Validation*.

Untuk setiap model yang telah terbentuk, gunakan model tersebut untuk memprediksi data uji yang telah dibuat. Setelah itu hitung *risk* untuk setiap model. Semakin tinggi *risk* yang di dapat dari model tersebut, maka semakin kecil bobot yang diberikan untuk model tersebut. Untuk peubah respon berupa data numerik, dapat menghitung nilai MSE (*Mean Squared Error*). Jika peubah responnya berupa data kategorik, *risk* dari model dapat diketahui dengan menghitung nilai dari (1-AUC). AUC di dapat dari plot sensitivitas dan 1- spesifisitas. Setelah diketahui *risk* dari masing-masing model, dapat dipilih algoritma yang memiliki *risk* terkecil.

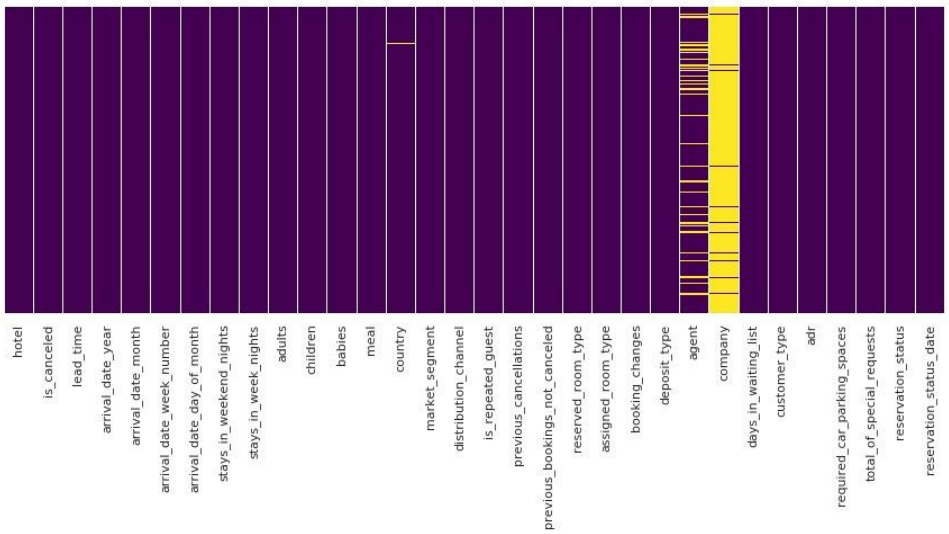
d. ROC Curve

Kurva ROC adalah tampilan grafis sensitivitas (TPR) pada sumbu y dan (1 - spesifisitas) (FPR) pada sumbu x untuk berbagai titik batas nilai tes. Area di bawah kurva (*Area Under Curve* / AUC) adalah ukuran sensitivitas yang efektif dan kombinasi dan spesifisitas untuk menilai validitas inheren atas diagnostik. Nilai maksimum AUC adalah 1, menandakan tes diagnostik sangat sempurna dalam membedakan pembatalan booking hotel dengan subjek yang tidak melakukan pembatalan.

BAB IV
ANALISIS

A. Praproses Data dan *Feature Engineering*

Pada penelitian ini, kami menggunakan dataset "Hotel Booking" dari kasus pembatalan pemesanan hotel di Portugal. Terdapat dua jenis hotel yaitu “City Hotel” yang berada di Lisbon dan “Resort Hotel” yang berada di Algarve. Dataset ini memiliki 32 fitur (atribut atau variabel) dan 119.390 observasi yang terbagi atas 40.060 observasi pada “Resort Hotel” dan 79.330 observasi pada data “City Hotel”. Data *booking* kamar hotel ini mengandung beberapa *missing value* atau data yang hilang. Data hilang tersebut terdapat pada beberapa peubah, seperti children (4 observasi) , agent (16340 observasi), country (488 observasi), dan company (112593 observasi). Ringkasan banyaknya *missing value* dari tiap peubah di dalam data booking hotel di Portugal ditampilkan pada gambar 2.



Gambar 2. Missing Value

Pengisian *missing value* dilakukan berdasarkan skala pengukuran dari peubahnya. Untuk data dengan skala pengukuran nominal, maka dilakukan imputasi dengan menggunakan modus. Untuk atribut agent, country, dan company, lebih baik dibuang dari dataset karena informasi yang hilang terlalu banyak.

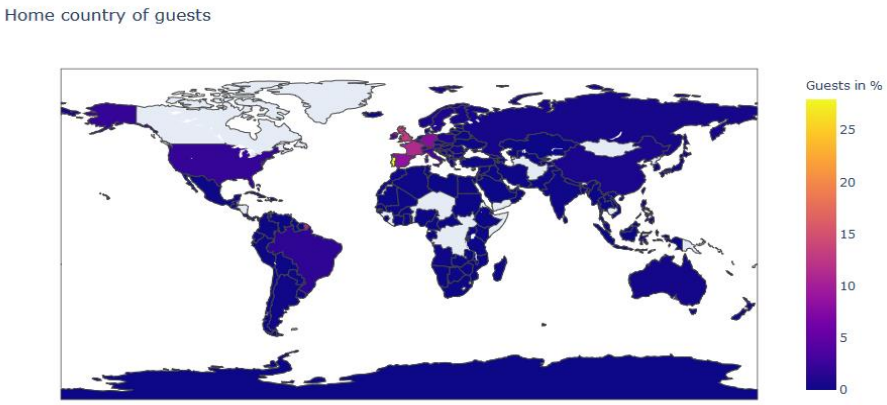
Untuk meningkatkan akurasi pada proses klasifikasi, salah satu cara yang dapat dilakukan adalah dengan menambahkan peubah baru ke dalam data berdasarkan peubah yang ada. Informasi mengenai peubah baru tersebut dapat dilihat pada tabel 2.

Tabel 1. Peubah Baru yang Ditambahkan

No	Nama Peubah	Penjelasan	Skala Pengukuran
1	total_duration	merupakan lamanya tamu menginap di hotel	Numeric
2	country_discr	berisi negara-negara pengunjung hotel yang dikelompokkan berdasarkan jumlahnya	Kategorik

B. Eksplorasi dan Analisis Data

1. Sebaran asal negara tamu hotel



Gambar 3. Peta asal negara tamu hotel

country	Number of Guests	Guests in %
PRT	20977	27.97
GBR	9668	12.89
FRA	8468	11.29
ESP	6383	8.51
DEU	6067	8.09

Gambar 4. Proporsi pengunjung dari masing - masing negara

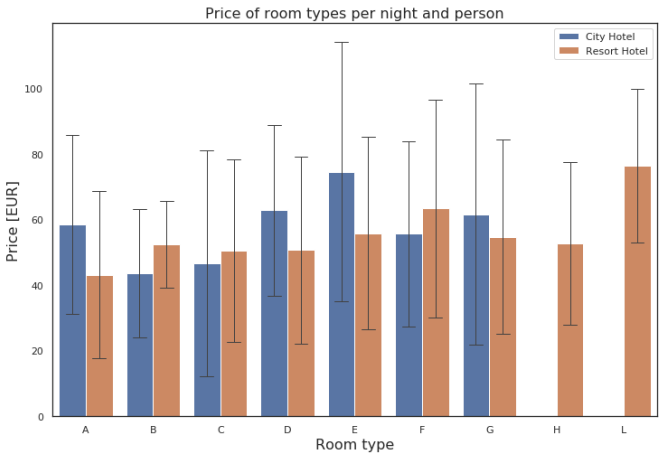
Berdasarkan hasil diatas, warga Portugal paling mendominasi jumlah tamu di hotel, baik di hotel kota, maupun hotel resor dengan menguasai 27.97% dari jumlah tamu keseluruhan. Selain Portugal, sebagian besar tamu hotel berasal dari benua Eropa, karena Portugal diklaim memiliki iklim yang cukup enak bagi warga Eropa untuk berlibur, terutama di di musim panas. Selain itu, di musim tertentu, beberapa kota di Portugal panen anggur (untuk dijadikan *wine*) (dikutip dari https://santorinidave.com/best-time-to-visit-portugal)

2. Jumlah tamu pada masing - masing hotel

Gambar 5. Perbandingan jumlah tamu pada masing - masing hotel

Secara keseluruhan, proporsi jumlah booking yang tidak dibatalkan dengan yang dibatalkan sekitar 5 : 3. Jika dilihat berdasarkan jenis hotel, hotel resor memiliki proporsi pembatalan booking yang lebih kecil (27.77%) daripada hotel kota (41.41%).

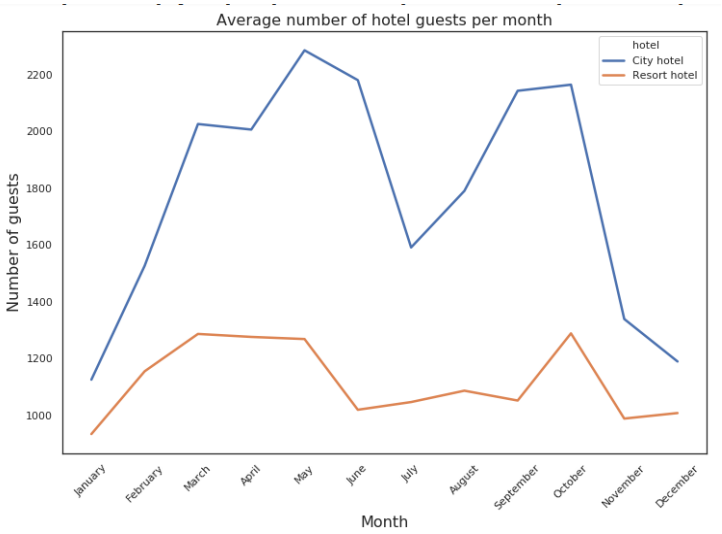
3. Harga kamar yang dibayar tamu untuk menginap per malam



Gambar 6. Harga tipe kamar per malam

Harga kamar untuk masing - masing hotel relatif ideal, untuk city hotel didapati ada 7 tipe kamar dengan harga kamar per malam termurah ada pada tipe kamar B dan harga tertinggi ada pada kamar bertipe E. Setiap tipe kamar diasumsikan memiliki pelayanan yang berbeda seperti semakin tinggi tingkat lantai yang didiami kamar maka harganya juga relatif berbeda. Ada tiga tipe kamar di hotel resor yang tidak terdapat pada hotel kota. Mungkin ketiga tipe kamar tersebut merupakan kamar khusus dengan *benefit* khusus, dan punya harga yang khusus juga (misal dari kamar tersebut, kita bisa melihat *sunrise* atau *sunset*). Disini, tipe kamar dibuat anonim dari asalnya, sehingga cukup sulit untuk melakukan eksplorasi lebih jauh, tetapi tipe kamar L paling digemari untuk resort hotel.

4. Pengunjung hotel pada setiap bulan



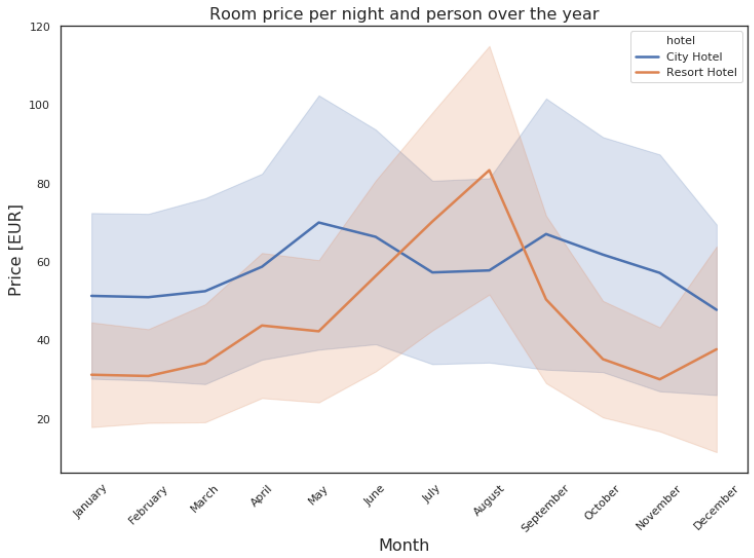
Gambar 7. Jumlah pengunjung pada setiap bulan

Berdasarkan pola di atas, ada beberapa pola yang bisa kita tangkap: (dikutip dari https://santorinidave.com/best-time-to-visit-portugal)

a. Puncak kunjungan untuk hotel kota terjadi pada sekitar bulan April-Juni dan September-Oktober. Bulan-bulan ini merupakan waktu terbaik untuk menikmati jalan-jalan di seluruh Portugal, didukung cuacanya menyenangkan dan cukup ringan (22 ^{o}C - 24 ^{o}C). Banyak festival dan acara berlangsung selama waktu ini. Beberapa tarif penerbangan dan tarif kamar ada yang lebih rendah dari biasanya. Pada bulan Juni dan September, ini adalah waktu terbaik untuk mengunjungi pantai-pantai Lisbon untuk cuaca yang baik dan hangat. Mereka yang ingin berenang atau *surfing* mungkin dapat melakukannya paling awal Mei dan paling lambat Oktober.
b. Puncak kunjungan untuk hotel resort yang berada di Algarve berada pada bulan September-Oktober dan Maret-Mei. Algarve paling baik dinikmati pada akhir April hingga awal Mei atau pertengahan September hingga akhir Oktober. Baik akhir musim semi dan awal musim gugur menawarkan suhu cukup hangat untuk bersantai di pantai, menjelajahi tebing pasir yang dramatis, dan menikmati kota-kota kecil bersejarah yang menawan. Awal musim gugur dapat menjadi tempat yang ideal untuk berenang di perairan sekitar Algarve.
c. Terjadi penurunan jumlah tamu hotel kota di sekitar bulan pertengahan Juni hingga Agustus, karena di Portugal, bulan-bulan ini termasuk *High Season*. Jalanan, pantai, dan area wisata pemandangan Portugal cukup sibuk di musim panas, ramai dengan orang Eropa yang mengambil liburan tahunan mereka. Harga akomodasi meningkat 30%. Wisatawan turun ke daerah resor Algarve dan pesisir untuk menikmati suhu puncak musim panas dan air laut yang hangat (inilah alasan terjadi kenaikan jumlah pengunjung hotel resor dari bulan Juni ke Agustus).
d. Jumlah tamu hotel cukup rendah di sekitar bulan November dan Maret dikarenakan musim hujan dimulai pada pertengahan November dan berlanjut hingga Maret, serta beberapa daerah di Portugal mulai membeku dan banyak resor ditutup saat musim dingin. Fasilitas ski tersedia antara Desember dan Maret.

e. Terjadi peningkatan jumlah tamu hotel secara signifikan antara Januari sampai Maret, karena banyak festival dan karnaval diselenggarakan. Di bulan Februari, ada balapan sepeda terkenal, *Tour of the Algarve* dengan panjang lintasan lebih dari 700 km yang terbagi menjadi beberapa tahap. Mendekati bulan Maret, perlombaan paling populer di Portugal, *Lisbon Half Marathon*, membawa lebih dari 35.000 pelari dari seluruh dunia untuk berlari melewati lintasan yang terbentang sepanjang sungai dan pantai.

5. Variasi harga kamar tiap bulan



Gambar 8. Harga kamar pada setiap bulan

Berdasarkan plot di atas, terlihat bahwa harga kamar untuk hotel kota tidak terlalu berfluktuasi, namun harga kamar untuk hotel resort cukup tinggi di sekitar bulan Juni sampai pertengahan Agustus. Hal ini berkaitan dengan kondisi *High Season* di sekitar bulan tersebut.

6. Pengaruh agent dan company terhadap perilaku tamu

Sebelumnya, saat melihat sekilas beberapa nilai dari atribut agent dan company, cukup banyak ditemukan nilai “Null”. Jika disesuaikan dengan deskripsi masing-masing atribut, maka : i) Jika atribut agent dan company bernilai “Null”, maka tamu tersebut melakukan pembayaran dan booking sendiri; ii) Jika atribut

agent bernilai “Null” sedangkan atribut company tidak “Null”, maka tamu tersebut melakukan booking dan pembayaran melalui perusahaannya; iii) Jika atribut company bernilai “Null” sedangkan atribut agent tidak “Null”, maka tamu tersebut melakukan booking melalui suatu agen, namun biayanya ditanggung si tamu.

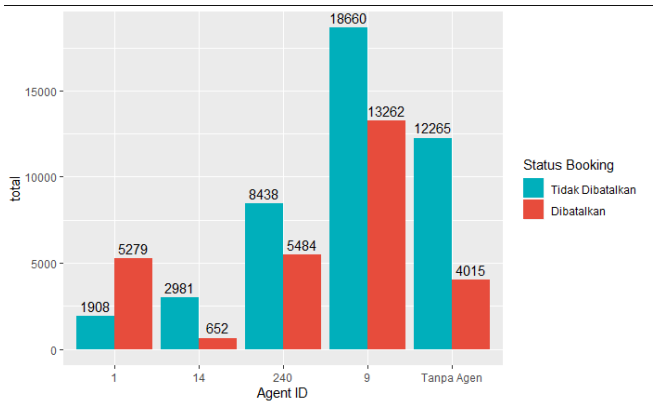
- Dari sisi agent

agent <fctr>	total <int>
9	31922
NULL	16280
240	13922
1	7187
14	3633
7	3532
6	3290
250	2870
241	1721
28	1657

1-10 of 334 rows

Gambar 9. Agen dengan penjualan paling banyak

Bisa kita lihat bahwa hampir 26.78% tamu melakukan booking kamar hotel melalui agent dengan ID 9. Cukup banyak juga tamu yang melakukan booking kamar tanpa agent.



Gambar 10. Perbandingan status booking pada masing - masing agen

Jika dilihat jumlah pembatalan booking untuk 5 hasil teratas pada atribut agent, terlihat bahwa agent ID 1 memiliki proporsi jumlah pembatalan booking terbesar diantara yang lain.

- Dari sisi company

company <fctr>	total <int>
NULL	112441
40	924
223	784
67	267
45	249
153	213
174	147
219	141
281	138
154	133

1-10 of 349 rows

company <fctr>	is_canceled <int>	total <int>
223	0	665
223	1	119
40	0	847
40	1	77
NULL	0	69423
NULL	1	43018

Previ 6 rows

Gambar 11. Company dengan pemesanan paling banyak

Bisa kita lihat bahwa lebih dari 94% orang melakukan pembayaran sendiri. Dua perusahaan yang cukup sering melakukan pembayaran booking kamar hotel adalah perusahaan dengan ID 40 dan 223. Berdasarkan hasil kedua, bisa kita lihat bahwa jika kamar tamu tersebut dibayar oleh perusahaan, maka tamu tersebut cenderung tidak membatalkan booking.

C. Anomali Data

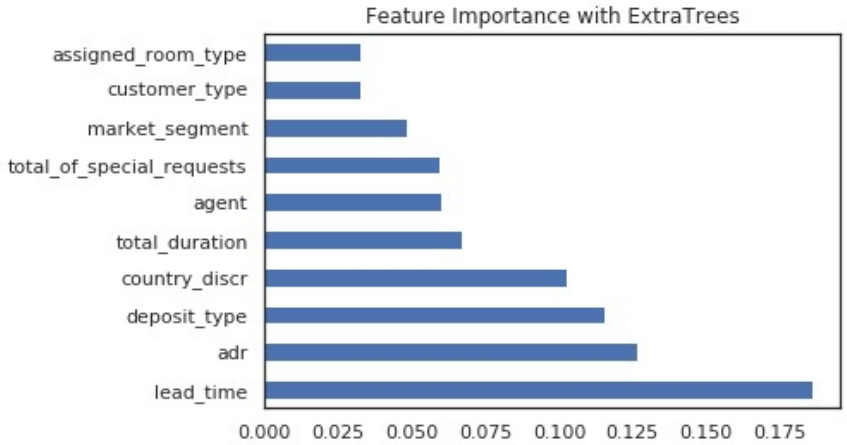
Saat melakukan eksplorasi, kami melihat adanya beberapa *outlier* ataupun *noise* pada beberapa atribut data (yang selanjutnya akan dibuang demi kebaikan model), seperti :

- Banyaknya data yang bernilai nol, bahkan ada yang negatif, pada atribut adr. Ada juga beberapa data dimana nilai atribut adults, children, babies, dan adr bernilai nol.
- Ada sekitar dua data outlier di mana nilai dari atribut lead_time lebih dari 650. Ada juga beberapa data outlier dimana nilai atribut total_duration lebih dari 40.
- Adanya data *noise* dimana nilai dari atribut adults, children, dan babies bernilai nol, tetapi reservation_status bernilai “Check-Out”. Untuk atribut adults, ada suatu noise yang jika nilainya lebih dari 4, maka dipastikan pesanan kamar dibatalkan. Untuk atribut children dan babies, ada suatu outlier dimana jika nilainya lebih dari 3, maka sudah pasti membatalkan pesanan.

- Adanya data yang bocor, yaitu atribut reservation_status. Atribut ini sudah pasti bisa memprediksi apakah tamu tersebut membatalkan pesanan atau tidak.

D. *Feature Selection*

Feature Selection digunakan untuk memilih variabel *predictor* yang dianggap memiliki tingkat pengaruh yang lebih tinggi dibanding variabel lainnya. Pada tahapan ini, metode seleksi yang digunakan adalah *feature importance* berdasarkan *Extra Trees*. Dalam dataset yang telah melalui praproses, terdapat 25 variabel *predictor*. Berikut adalah hasil dari *feature selection*:



Gambar 12. Output *Feature Importance*

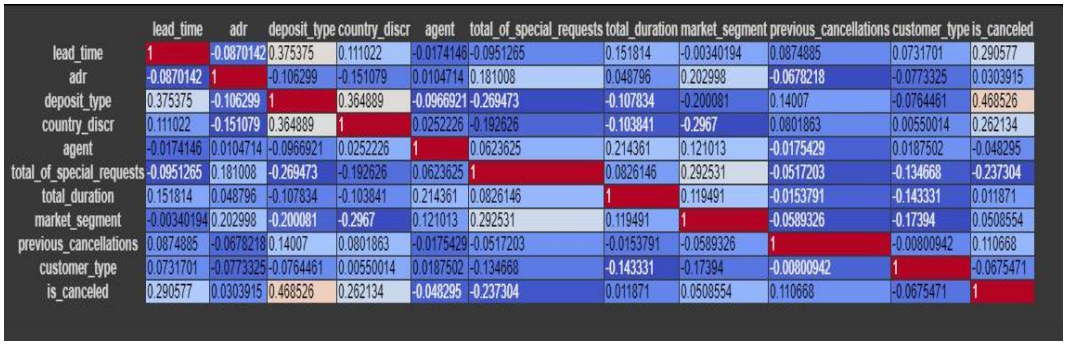
Hasil di atas menunjukkan bahwa terdapat 10 variabel yang terpilih diantaranya feature lead_time, adr deposit_type, country_discr, agent, total_of_special_requests, customer_type, total_duration, market_segment, dan previous_cancellations. Variabel yang memiliki tingkat kepentingan paling tinggi terhadap data ini adalah lead_time atau jumlah waktu tunggu mulai dari proses booking sampai check-in. Pada pengolahan selanjutnya akan menggunakan kesepuluh variabel *predictor* ini.

Tabel 2. Peubah penjelas yang terpilih

Variabel	Feature Description	Tipe Data
X1	lead_time	Numeric
X2	adr	Numeric
X3	deposite_type	Kategorik
X4	country_discr	Kategorik
X5	agent	Kategorik
X6	total_of_special_requests	Numeric
X7	total_duration	Numeric
X8	market_segment	Kategorik
X9	previous_cancellations	Kategorik
X10	customer_type	Kategorik

Untuk pemodelan berikutnya akan menggunakan ke 10 feature tersebut, hubungan

E. Matriks Korelasi



	lead_time	adr	deposit_type	country_discr	agent	total_of_special_requests	total_duration	market_segment	previous_cancellations	customer_type	is_canceled
lead_time	1	-0.0870142	0.375375	0.111022	-0.0174146	-0.0951265	0.151814	-0.00340194	0.0874885	0.0731701	0.290577
adr	-0.0870142	1	-0.106299	-0.151079	0.0104714	0.181008	0.048796	0.202998	-0.0678218	-0.0773325	0.0303915
deposit_type	0.375375	-0.106299	1	0.364889	-0.0966921	-0.269473	-0.107834	-0.200081	0.14007	-0.0764461	0.468526
country_discr	0.111022	-0.151079	0.364889	1	0.0252226	-0.192626	-0.103841	-0.2967	0.0801863	0.00550014	0.262134
agent	-0.0174146	0.0104714	-0.0966921	0.0252226	1	0.0623625	0.214361	0.121013	-0.0175429	0.0187502	-0.048295
total_of_special_requests	-0.0951265	0.181008	-0.269473	-0.192626	0.0623625	1	0.0826146	0.292531	-0.0517203	-0.134668	-0.237304
total_duration	0.151814	0.048796	-0.107834	-0.103841	0.214361	0.0826146	1	0.119491	-0.0153791	-0.143331	0.011871
market_segment	-0.00340194	0.202998	-0.200081	-0.2967	0.121013	0.292531	0.119491	1	-0.0589326	-0.17394	0.0508554
previous_cancellations	0.0874885	-0.0678218	0.14007	0.0801863	-0.0175429	-0.0517203	-0.0153791	-0.0589326	1	-0.00800942	0.110668
customer_type	0.0731701	-0.0773325	-0.0764461	0.00550014	0.0187502	-0.134668	-0.143331	-0.17394	-0.00800942	1	-0.0675471
is_canceled	0.290577	0.0303915	0.468526	0.262134	-0.048295	-0.237304	0.011871	0.0508554	0.110668	-0.0675471	1

Gambar 13. Matriks Korelasi

Setelah melewati tahap praproses, kita akan melihat matriks korelasi untuk melihat hubungan dari masing - masing feature. Berdasarkan matriks korelasi di atas, setidaknya ada 7 atribut yang memiliki hubungan positif dengan respon atribut target (is_canceled), yaitu lead_time, adr, deposit_time, country_discr, total_duration, market_segment, dan previous_cancellations. Untuk atribut yang

memiliki hubungan negatif ada 3 atribut yaitu agent, total_of_special_requests dan customer_type. Kita juga bisa melihat adanya korelasi antar atribut penjelas, yaitu atribut lead_time dengan deposit_type (berkorelasi positif), total_of_special_requests dengan market_segment (berkorelasi positif), dan country_discr dengan deposit_type (berkorelasi positif).

F. Pencocokan Model

Proses klasifikasi "Hotel Booking" dicoba dengan menggunakan beberapa metode, seperti *Extra Trees*, *Random Forest*, *AdaBoost*, KNN, dan *LightGBM* untuk semua data. Tentu saja data yang akan diklasifikasikan harus sudah dirapikan terlebih dahulu seperti, mengubah peubah kategorik menjadi peubah boneka serta memilih peubah-peubah yang akan dimasukkan ke dalam model sesuai dengan hasil dari feature selection. Selanjutnya, dilakukan pemisahan data latih dan data uji sebesar 70% untuk data latih dan 30% untuk data uji, berikut hasil prediksi dari masing - masing model :

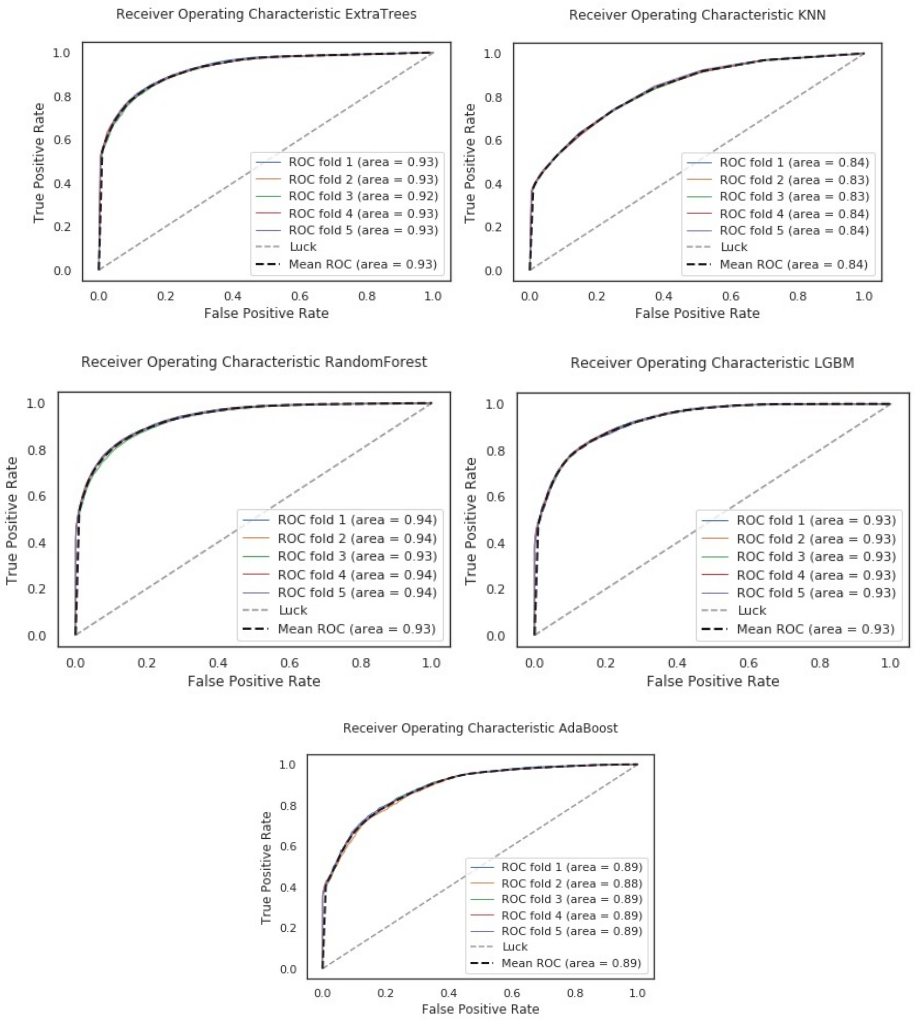
Tabel 3. *Evaluation Metrics* untuk Tiap Metode

Model	Accuracy	Precision	Recall	F1 Score	AUC
KNN	0.773	0.726	0.635	0.751	0.84
AdaBoost	0.813	0.805	0.661	0.792	0.89
Random Forest	0.863	0.833	0.793	0.852	0.93
LightGBM	0.854	0.844	0.749	0.840	0.93
Extra Trees	0.858	0.826	0.787	0.847	0.93

Tabel di atas menunjukkan hasil akurasi, presisi, *recalls*, dan AUC dengan *hold-out method*. Terdapat tiga metode yang memiliki tingkat akurasi dan AUC tertinggi yakni *LightGBM*, *Random Forest*, dan *Extra Trees*.

G. Evaluasi

Untuk mengevaluasi model penulis membandingkan hasil nilai akurasi dan AUC yang menggunakan *hold-out method* dengan *cross validation*. Berikut adalah hasil akurasi dan AUC antara hold out dan cross validation :



Gambar 14. Kurva ROC untuk Tiap Metode

Tabel 4. Perbandingan nilai *Accuracy* berdasarkan *Hold-Out Method* dan *Cross-Validation*

Model	Accuracy	Cross Validation
KNN	0.773	0.77
AdaBoost	0.813	0.84

Random Forest	0.863	0.85
LightGB	0.854	0.85
Extra Trees	0.858	0.86

Hasil akurasi pada setiap metode dengan *cross validation* lebih tinggi dibandingkan dengan *hold-out method*. Kecuali untuk model KNN. Sedangkan untuk ukuran nilai kebaikan model bisa dilihat pada tabel dibawah :

Tabel 5. Nilai Galat dari Tiap Model Sebelum Tuning

Model	MAE	MSE	RMSE
KNN	0.226	0.226	0.475
AdaBoost	0.186	0.186	0.432
Random Forest	0.136	0.136	0.369
LightGBM	0.145	0.145	0.381
Extra Trees	0.141	0.141	0.376

H. Hyperparameter Tuning

Hyperparameter Tuning hanya penulis gunakan untuk metode dengan nilai akurasi yang terbaik, yaitu metode *LightGBM*, *Random Forest*, AdaBoost dan *Extra Trees*. Berikut adalah *hyperparameter tuning* pada masing-masing metode.

1. *Extra Trees*

   Parameter terbaik untuk metode *Extra Trees* yang digunakan dalam *Hyperparameter Tuning* adalah class_weight yang *balanced*, dengan *criterion* adalah *gini*, max_depth sebesar None, max_feature adalah *3*, min_samples_split adalah 10, min_samples_leaf adalah 3, bootstrap adalah True, dan n_estimators sebesar 300. *Extra Trees* yang parameternya sudah

ditentukan melalui *Hyperparameter Tuning* memiliki *Best Score* sebesar 0,857.

2. *Random Forest*

   Parameter terbaik untuk metode *Random Forest* adalah *criterion* dengan *entropy*, max_depth sebesar None, max_feature adalah *1*, min_samples_split adalah 10, min_samples_leaf adalah 10, bootstrap adalah False, dan n_estimators sebesar 300. *Best Score* yang diperoleh sebesar 0.856.

3. *LightGBM*

   Parameter terbaik untuk metode *LightGBM* adalah loss dengan deviance, n_estimators sebesar 300, learning_rate adalah 0,1, max_depth adalah 8, min_samples_leaf adalah 100, dan max_features adalah 0,1.Hasil best score untuk LightGBM sebesar 0.853.

H. Model Terbaik

Nilai akurasi, presisi, *recalls*, dan AUC pada model setelah dilakukan *hyperparameter tuning* adalah sebagai berikut :

Tabel 6. Perbandingan *Evaluation Metrics* untuk Tiap Metode

Algoritma	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.856	0.873	0.723	0.841	0.84
LightGBM	0.853	0.841	0.749	0.839	0.89
Extra Trees	0.857	0.811	0.806	0.847	0.93

Tabel 7. Nilai Galat untuk Tiap Model

Model	MAE	MSE	RMSE
Random Forest	0.143	0.143	0.378
LightGBM	0.146	0.146	0.383
Extra Trees	0.142	0.142	0.377

Nilai akurasi, presisi, *recall, F1-score*, serta AUC pada ketiga metode ini sangat besar dan hasil dari *accuracy* dan *F1-score* diatas 82% untuk *Random Forest* dan *Extra Trees* yang artinya nilai data yang diujikan cukup baik untuk memprediksi data aktual yang *imbalance*, serta tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi juga cukup baik, dapat dilihat dari nilai *recalls* Extra Trees diatas 80%. Baik nilai akurasi, *recall*, presisi, *F1-score* maupun AUC yang tinggi dapat menjadi hal baik, namun patut dikoreksi dan ditinjau kembali pada bagian pengolahan data karena nilai *recall* dari masing-masing model mayoritas di bawah 80%. Namun, penulis telah mencoba metode-metode klasifikasi tersebut dengan data yang sudah diseimbangkan atau data yang *balance*. Nilai akurasi yang dihasilkan meningkat signifikan untuk model *Extra Trees*. Sehingga, dapat diselidiki lebih lanjut mengenai apa yang menyebabkan besarnya tingkat *recall* pada penelitian ini kecil. Setelah dilakuakan hyperparameter tunning didapatkan nilai galat pada Extra Trees adalah yang terkecil dengan nilai MAE, MSE, dan RMSE sebesar 1.42, 1.42, dan 3.77. Hal ini menjadikan Extra Trees sebagai model yang tepat untuk penelitian ini.

BAB V
KESIMPULAN

Dari hasil dan pembahasan yang telah dilakukan dalam prediksi faktor-faktor penyebab tamu melakukan pembatalan pemesanan didapat bahwa sebanyak 10 fitur memiliki kemampuan prediktif mengenai probabilitas pembatalan pemesanan dengan keakuratan yang sangat baik dan tidak terjadi *overfitting* pada model Extra Trees. Untuk rata-rata ukuran akurasi, presisi, recall, dan f1-score dari keseluruhan gugus data mengalami kenaikan walaupun tidak terlalu besar setelah dilakukan hyperparameter tunning. Faktor yang diyakini berpengaruh untuk kedua hotel yaitu lead_time, adr deposit_type, country_discr, agent, total_of_special_requests, customer_type, total_duration, market_segment, dan previous_cancellations. Dari faktor-faktor tersebut, faktor lead_time merupakan faktor yang paling penting, diikuti dengan faktor adr, dan deposite_tipe dalam mempengaruhi kebaikan model *Extra Trees*. Model *Extra Trees* diketahui memiliki nilai *accuracy*, *recall*, *presisi* dan *F1-score* paling tinggi dibandingkan dengan model lain, hal ini bisa dianggap bahwa model *Extra Trees* sanggup memprediksi untuk kedua hotel tersebut.

BAB VI
DAFTAR PUSTAKA

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), pp. 5-32.

Chen, C. C., Schwartz, Z. and Vargas, P. (2011). The Search for The Best Deal: How Hotel Cancellation Policies Affect the Search and Booking Decisions of Deal-Seeking Customers. *International Journal of Hospitality Management*, 30(1), pp. 129–135.

Hastie, T. J., Tibshirani, R. J. and Friedman JH. 2008. *The Elements of Statistical Learning: Data-Mining, Inference and Prediction.* 2nd ed. New York: Springer-Verlag

Hayes, D. K. and Miller, A. A. (2011). *Revenue Management for The Hospitality Industry*. Hoboken, NJ: John Wiley & Sons, Inc.

Ivanov, S. (2014). *Hotel Revenue Management: From Theory to Practice*. Varna: Zangador.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* . New York: Springer.

Linoff, G. S. and Berry, M. J. 2011. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. New York: John Wiley & Sons.

Liu, P. H. (2004). Hotel Demand/Cancellation Analysis and Estimation of Unconstrained Demand using Statistical Methods. In: I. Yeoman and U. McMahon-Beattie, ed., *Revenue Management and Pricing: Case Studies and Applications* (pp. 91–101). Cengage Learning EMEA.

Mehrotra, R. and Ruttley, J. (2006). *Revenue Management.* 2nd ed. Washington, DC: American Hotel & Lodging Association (AHLA).

Morales, D. R. and Wang, J. (2010). Forecasting Cancellation Rates for Services Booking Revenue Management using Data Mining. *European Journal of Operational Research*, 202(2), pp. 554–562.

Noone, B. M. and Lee, C. H. (2010). Hotel Overbooking: The Effect of Overcompensation on Customers' Reactions to Denied Service. *Journal of Hospitality & Tourism Research*, 35(3), pp. 334–357

Santorini Dave. (2020). Best Time to Visit Portugal. [online] Available at: https://santorinidave.com/best-time-to-visit-portugal [Accessed 5 Mar. 2020]

Smith, S. J., Parsa, H. G., Bujisic, M. and van der Rest, J. P. (2015). Hotel Cancelation Policies, Distributive and Procedural Fairness, and Consumer Patronage: A Study of The Lodging Industry. *Journal of Travel & Tourism Marketing*, 32(7), pp. 886–906

StayNTouch. (2019). *Overbooking in Your Hotel*. [online] Available at: https://www.stayntouch.com/blog/overbooking-your-hotel/ [Accessed 5 Mar. 2020]

Talluri, K. T. and Van Ryzin, G. (2004). *The Theory and Practice of Revenue Management*. Boston: Kluwer Academic Publishers.