به نام خداوند بخشنده و مهربان

درس داده کاوی

پروژه پایانی درس

استاد درس: دکتر ناظرفرد

نام دانشجو:

روزبه قاسمی ۹۵۳۱۴۲۴

تیر ۱۳۹۹

فهرست مطالب

مقدمه

هدف از انجام این پروژه آشنایی با مفهوم تجزیه و تحلیل داده‌های اکتشافی[1] آشنا شویم. تجزیه و تحلیل داده‌های اکتشافی به مجموعه‌ای از تکنیک‌ها اشاره دارد که در اصل توسط جان توکی[2] استفاده شده‌است تا داده‌ها را به گونه‌ای نمایش دهد که ویژگی‌های جالب آشکار شوند. بر خلاف روش‌های کلاسیک که معمولا با یک مدل فرض شده برای داده‌ها شروع می‌شوند، از تکنیک‌های EDA برای تشویق داده‌ها برای پیشنهاد مدل‌هایی استفاده می‌شود که ممکن است مناسب باشند. این مفهوم به ما کمک می‌کنند تا بتوانیم زمانی که با یک مجموعه‌داده مواجه می‌شویم، چگونه آنرا بررسی و تحلیل کنیم. تحلیل داده‌ها به ما کمک می‌کند تا بتوانیم از هزاران داده و ویژگی موجود در مجموعه‌داده اطلاعات مفیدی را بدست آوریم. در مرحله بعدی ما باید بتوانیم ویژگی‌هایی را از میان ویژگی‌های موجود انتخاب کرده و آن‌ها را جدا کنیم و در نهایت برای پیش‌بینی این که راهی برای پیش‌بینی احتمال لغو رزرو هتل به طور اختصاصی پیدا کنیم.

مرحله اول؛ ارزیابی داده‌ها

در ابتدا پس از آدرس‌دهی کردن محل ذخیره مجموعه دادن و خواندن آن به کمک کتابخانه Pandas، با دستور Head، ۵ رکورد اول آنرا نشان می‌دهیم.

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	meal	country
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	0.0	0	BB	PRT
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	0.0	0	BB	PRT
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	0.0	0	BB	GBR
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	0.0	0	BB	GBR
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	0.0	0	BB	GBR

[1] Exploratory Data Analysis
[2] John Tukey

سپس اطلاعات دیتاست را بدست می‌آوریم:

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

از تابع اطلاعات برای چاپ یک خلاصه مختصر از یک DataFrame استفاده می‌شود. این روش اطلاعات را در مورد یک DataFrame از جمله the شاخص و dtypes ستون، مقادیر غیر صفر و کاربرد حافظه را چاپ می‌کند.

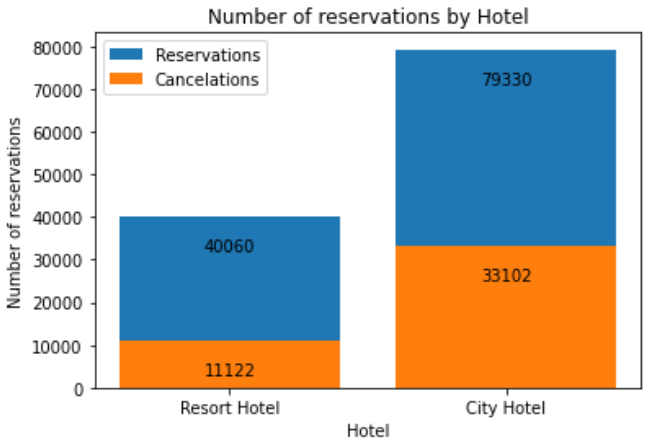
سپس در ادامه، آن ویژگی‌هایی از مجموعه‌داده که نیاز است اصلاح شوند، را اصلاح می‌کنیم.

```
dataset.isnull().sum()
```

```
hotel                               0
is_canceled                         0
lead_time                           0
arrival_date_year                   0
arrival_date_month                  0
arrival_date_week_number            0
arrival_date_day_of_month           0
stays_in_weekend_nights             0
stays_in_week_nights                0
adults                              0
children                            4
babies                              0
meal                                0
country                           488
market_segment                      0
distribution_channel                0
is_repeated_guest                   0
previous_cancellations              0
previous_bookings_not_canceled      0
reserved_room_type                  0
assigned_room_type                  0
booking_changes                     0
deposit_type                        0
agent                           16340
company                        112593
days_in_waiting_list                0
customer_type                       0
adr                                 0
required_car_parking_spaces         0
total_of_special_requests           0
reservation_status                  0
reservation_status_date             0
dtype: int64
```

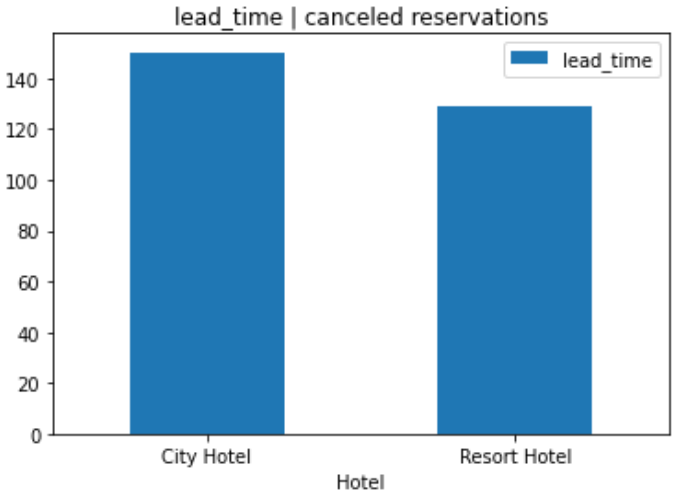
```
#children: NA -> 0
#country: NA -> 'N.A.'
#agent: NA -> 0
#company: NA -> 0

dataset.fillna({'children' : 0, 'country' : 'N. A.', 'agent' : 0, 'company' : 0}, inplace = True)
```

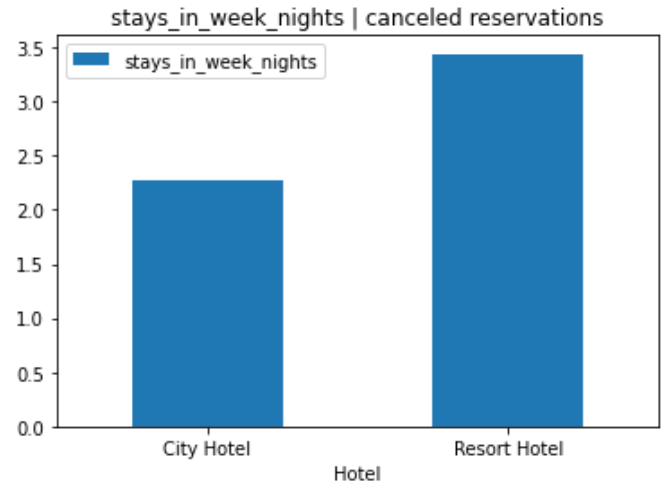
سپس در ادامه، به ترتیب اطلاعاتی را بدست می‌آوریم که برای جلوگیری از طولانی شدن گزارش صرفاً نتایج بدست‌آمده را به صورت خلاصه تحلیل می‌کنیم و کد را دیگر در گزارش نمی‌آوریم.



در شکل بالا، رزروهای صورت گرفته بر اساس نوع هتل‌ها مشخص شده است. در این شکل مشخص می‌کند که چه تعداد از رزرواسیون‌ها کنسل شده است و در کل برای هر نوع هتل، چند رزرواسیون صورت گرفته است.

در ادامه، برای پیدا کردن ویژگی‌هایی که بیشترین تاثییر در کنسل کردن یک رزرو هتل دارند را بررسی می‌کنیم.

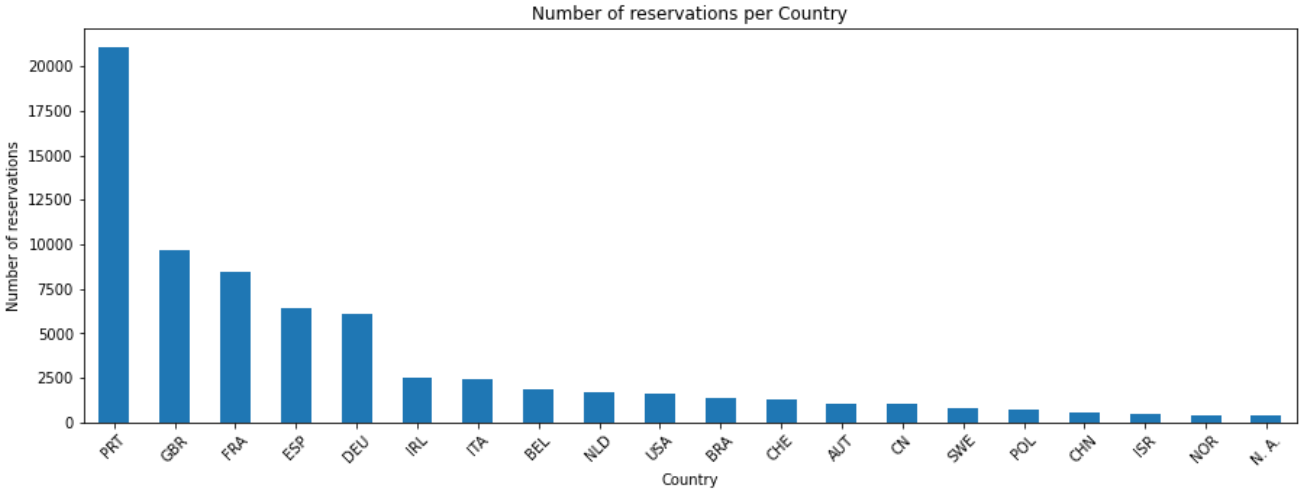




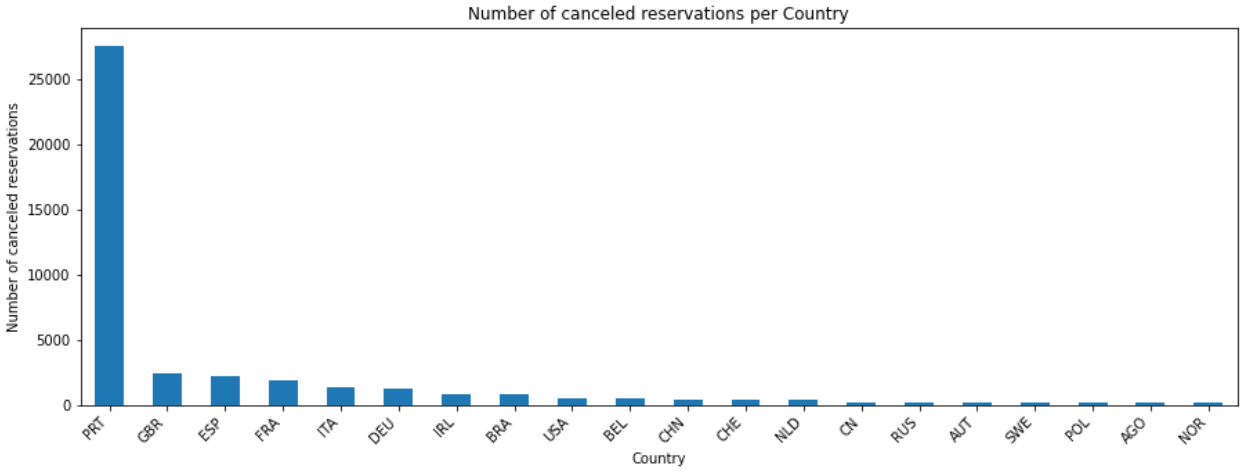
در ادامه، بر اساس ارزش هر کدام از ویژگی‌ها و بر اساس نوع هتل آن‌ها را لیست می‌کنیم.

	hotel	attribute	type	value
0	City Hotel	lead_time	mean	150.281222
1	Resort Hotel	lead_time	mean	128.680543
2	City Hotel	stays_in_weekend_nights	mean	0.787505
3	Resort Hotel	stays_in_weekend_nights	mean	1.335281
4	City Hotel	stays_in_week_nights	mean	2.266781
5	Resort Hotel	stays_in_week_nights	mean	3.440299
6	City Hotel	adults	mean	1.882907
7	Resort Hotel	adults	mean	1.957741
8	City Hotel	children	mean	0.079451
9	Resort Hotel	children	mean	0.187017
10	City Hotel	babies	mean	0.001933
11	Resort Hotel	babies	mean	0.009441
12	City Hotel	is_repeated_guest	mean	0.013322
13	Resort Hotel	is_repeated_guest	mean	0.009980
14	City Hotel	previous_cancellations	mean	0.161561
15	Resort Hotel	previous_cancellations	mean	0.347599
16	City Hotel	previous_bookings_not_canceled	mean	0.026041
17	Resort Hotel	previous_bookings_not_canceled	mean	0.022388
18	City Hotel	booking_changes	mean	0.079844
19	Resort Hotel	booking_changes	mean	0.153390
20	City Hotel	days_in_waiting_list	mean	4.730409
21	Resort Hotel	days_in_waiting_list	mean	0.092789
22	City Hotel	adr	mean	104.687920
23	Resort Hotel	adr	mean	105.787010
24	City Hotel	required_car_parking_spaces	mean	0.000000
25	Resort Hotel	required_car_parking_spaces	mean	0.000000
26	City Hotel	total_of_special_requests	mean	0.275754
27	Resort Hotel	total_of_special_requests	mean	0.486783
28	City Hotel	meal	count	33102.000000
29	Resort Hotel	meal	count	11122.000000

در ادامه، رزروهای صورت گرفته بر اساس کشورها در شکل زیر نمایش داده می‌شود:

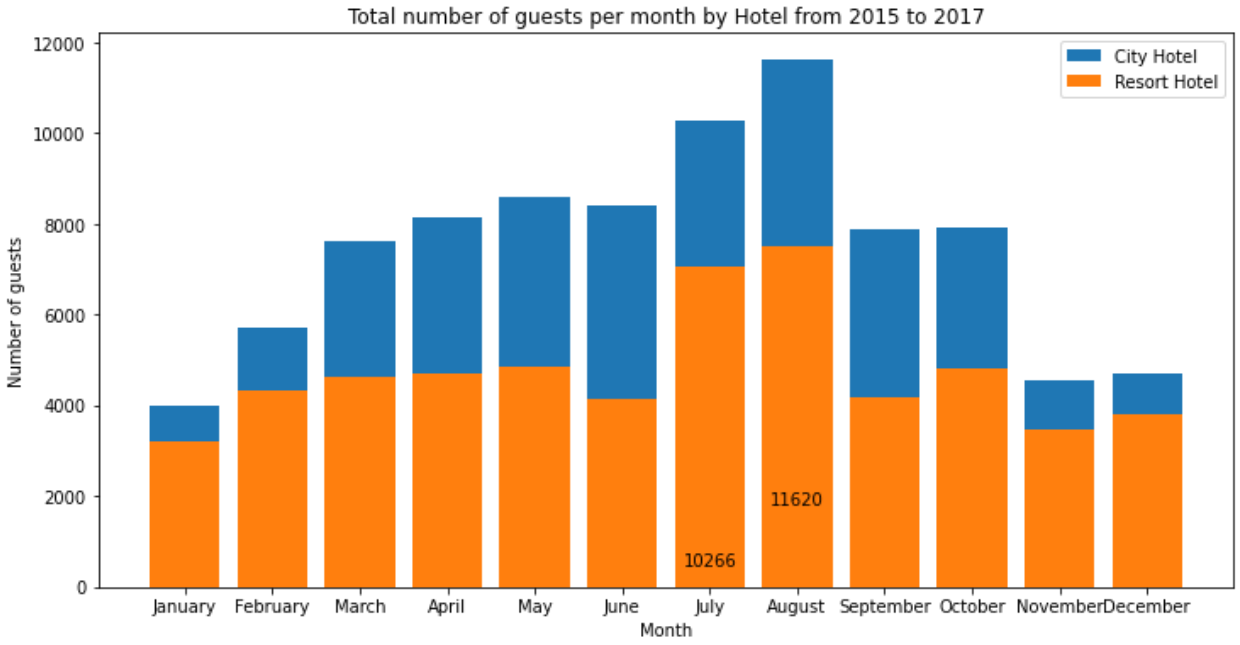


سپس، رزروهای کنسل شده بر اساس کشورها را در شکل زیر مشاهده می‌کنیم:



همانطور که در دو شکل بالا مشخص است، کشور پرتغال بیشترین نرخ کنسل کردن رزرواسیون‌ها را دارد. همچنین پرتغال دارای بالاترین نرخ رزرواسیون نیز می‌باشد!

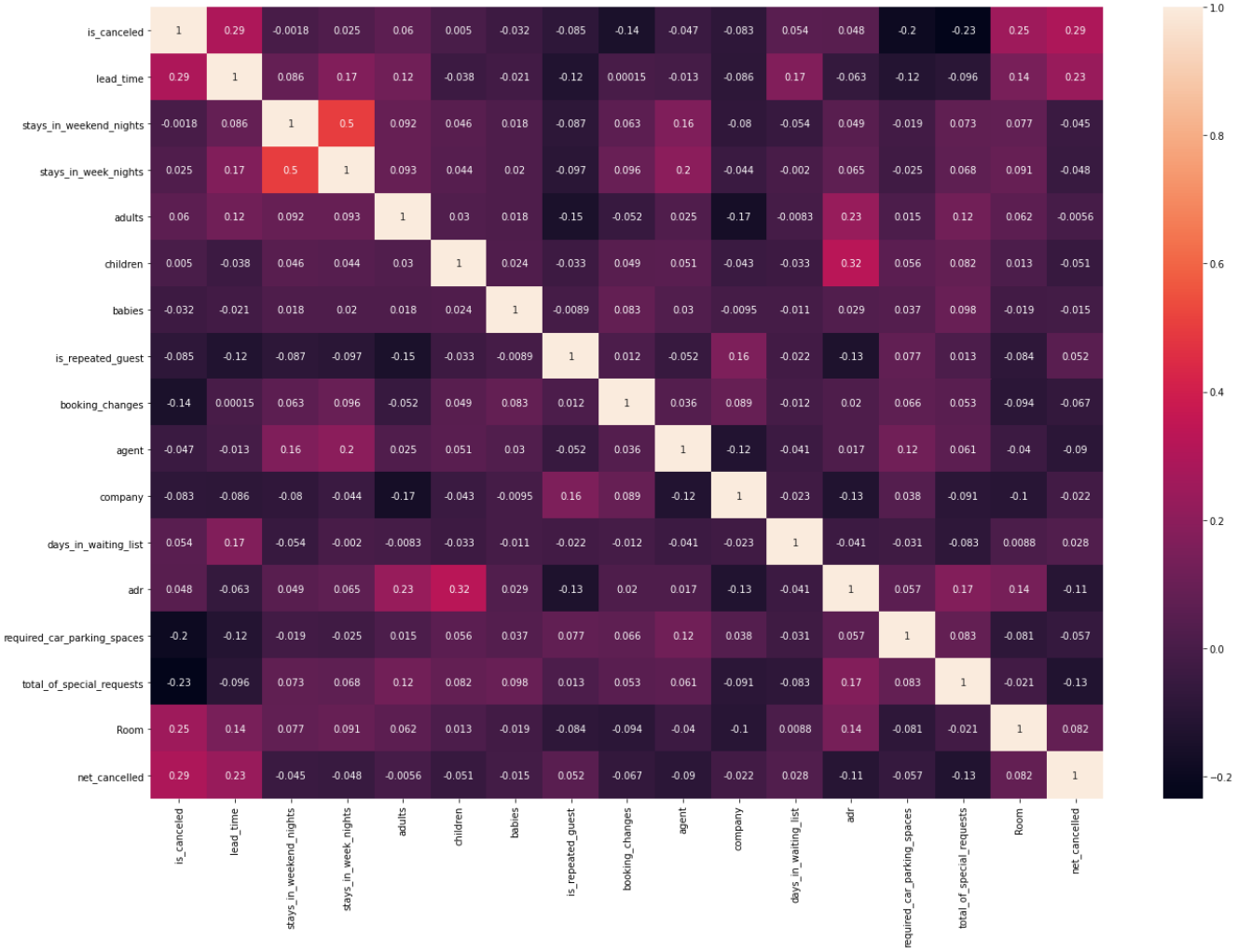
در شکل زیر، تعداد مهمان‌های هر هتل از ابتدای سال ۲۰۱۵ تا ابتدای سال ۲۰۱۷ یعنی تا دسامبر ۲۰۱۶ آورده شده است.



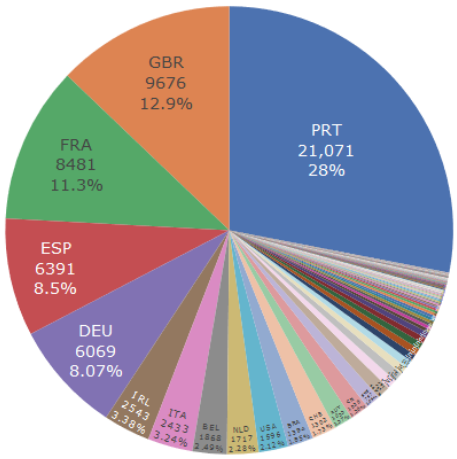
سپس در ادامه نشان دادم که به طور متوسط هر مهمان در هر نوع از هتل‌ها چند شب مهمان بوده است:

	nights
hotel	
City Hotel	2.923618
Resort Hotel	4.142892

سپس در شکل زیر یک heatmap از correlation میان ویژگی‌های این مجموعه‌داده آورده شده است:

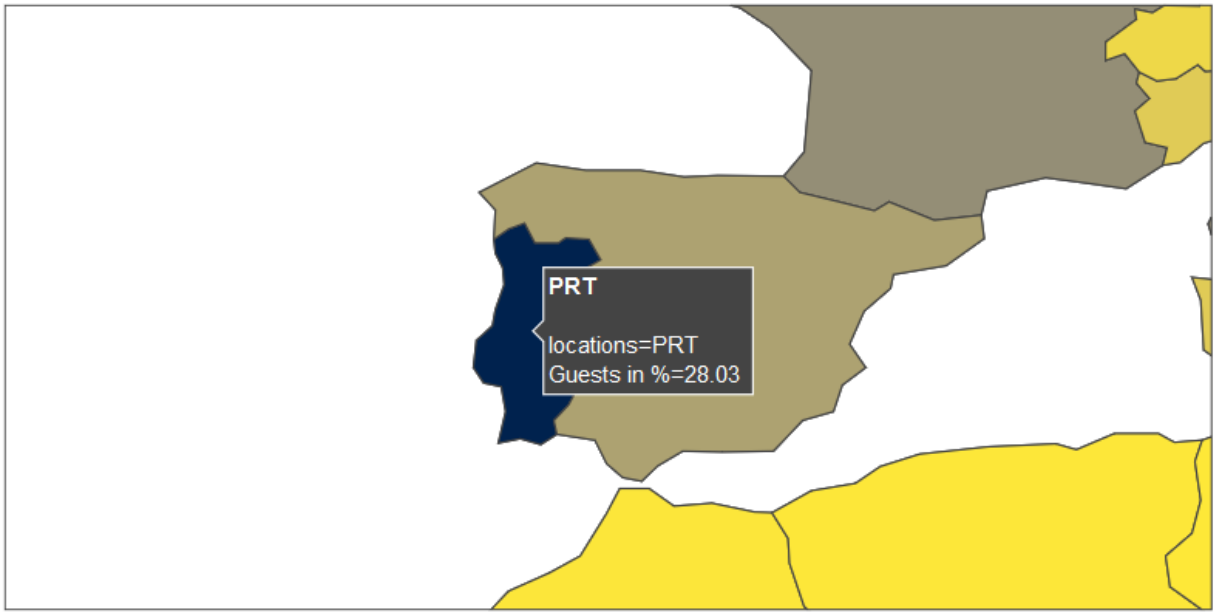


سپس در شکل زیر، توزیع مهمان‌های هتل‌ها بر اساس کشوری که از آنجا می‌آید آورده شده است:

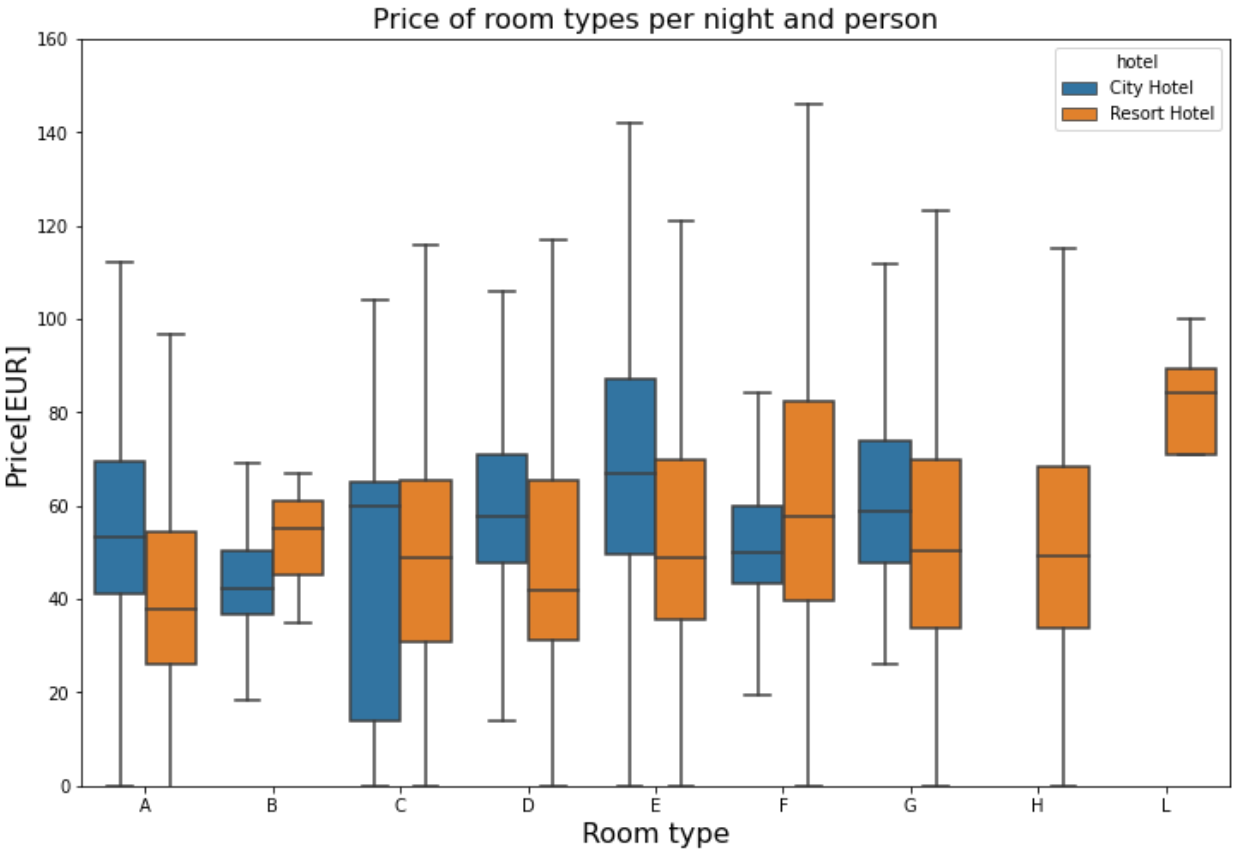


سپس مشابه شکل بالا، بر روی نقشه جهان نشان می‌دهیم:

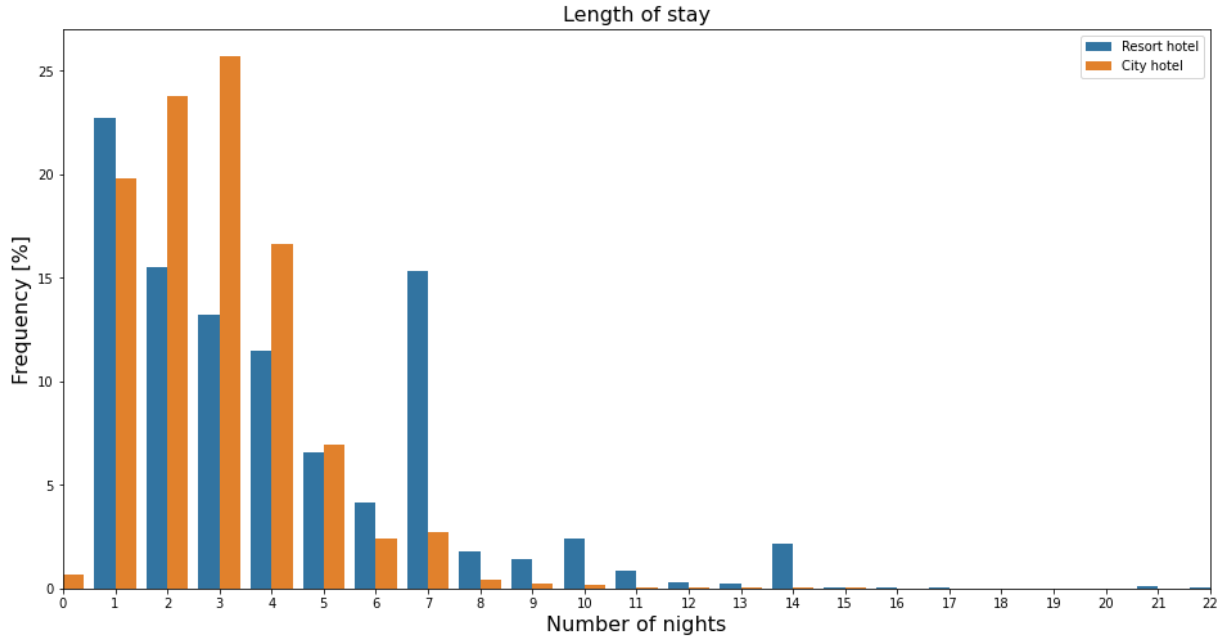
به این صورت که هر چه تعداد مهمان‌ها از آن کشورها بیشتر باشد، رنگ تیره‌تری دارد. همانطور که در قبل اشاره شد، کشور پرتغال دارای بیشترین متقاضی رزرواسیون هتل هاست که در شکل زیر نشان داده شده است:



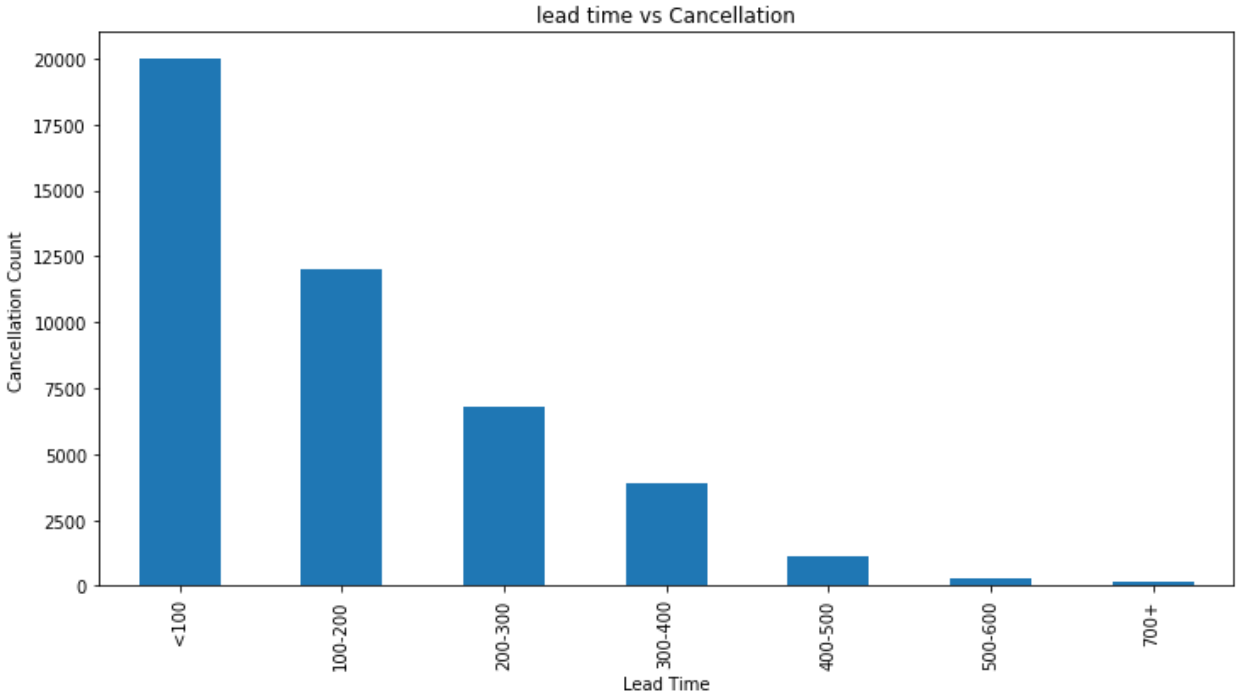
سپس در قسمت زیر، قیمت نوع اتاق‌ها را بر اساس نوع هتل‌ها به یورو نشان می‌دهم:



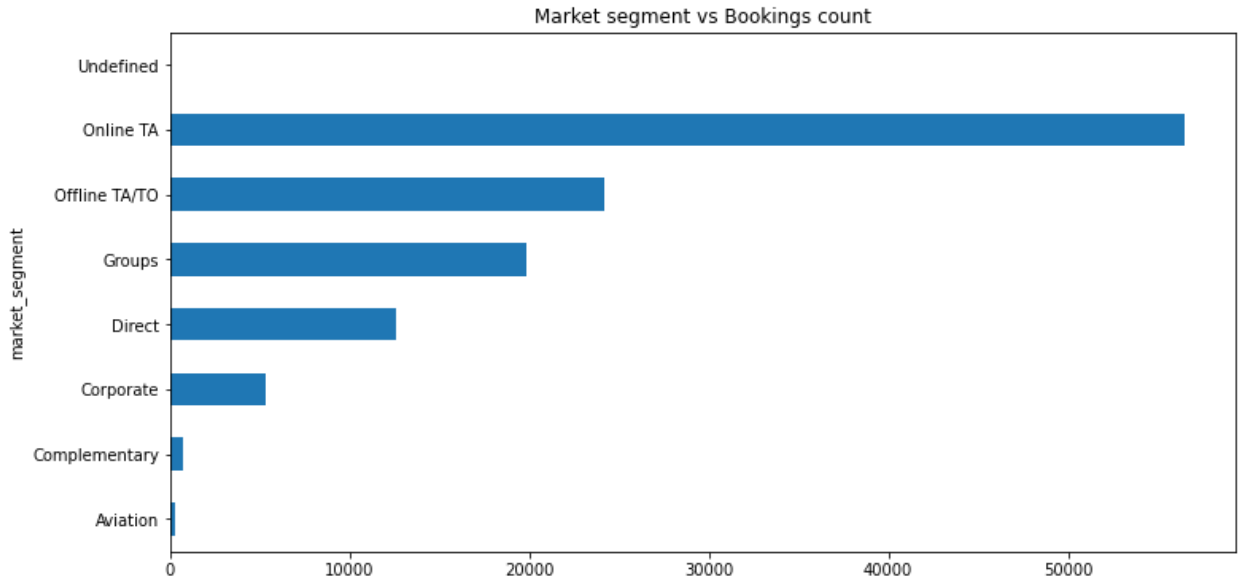
سپس در شکل زیر، فرکانس طول اقامت مهمان‌ها نشان داده شده است:



سپس بررسی می‌کنیم که آیا لغو رزرو تحت تأثیر عوامل دیگری است؟

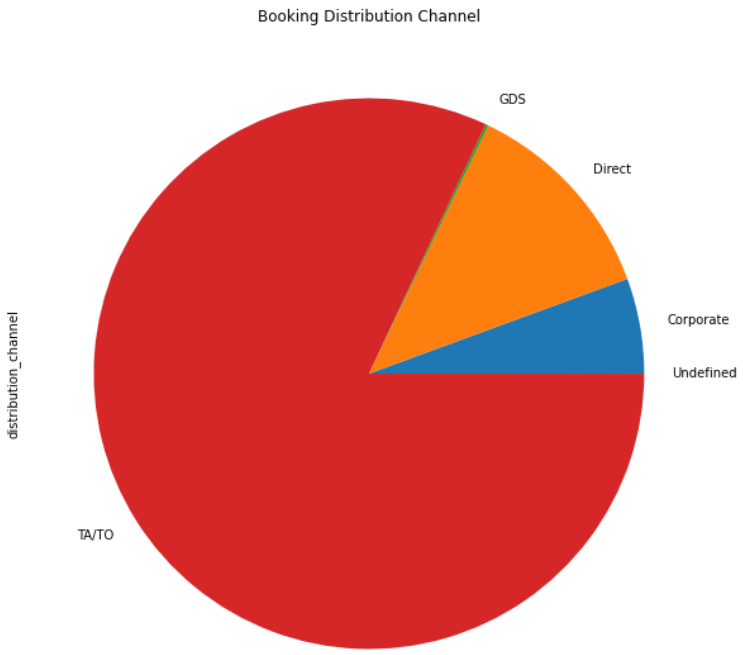


در ادامه، بررسی می‌کنیم که حداکثر تعداد رزروهای انجام شده در کدام بخش بازار است؟



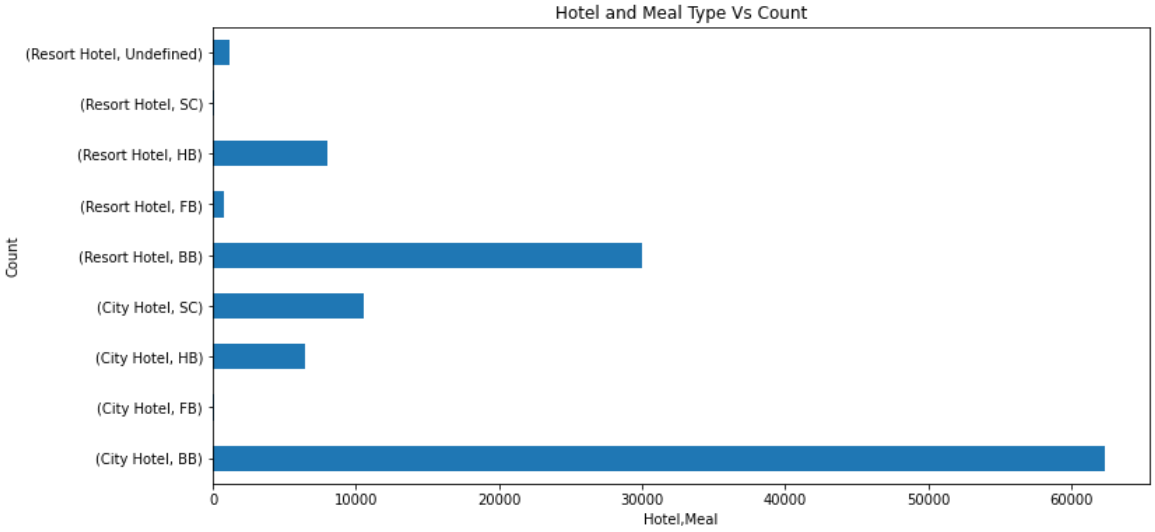
همانطور که مشخص است، بیشترین رزرو از طریق آژانس مسافرتی آنلاین انجام می شود.

در ادامه بررسی می‌کنم که بیشتر از کدام روش برای رزرو هتل استفاده می‌شود؟



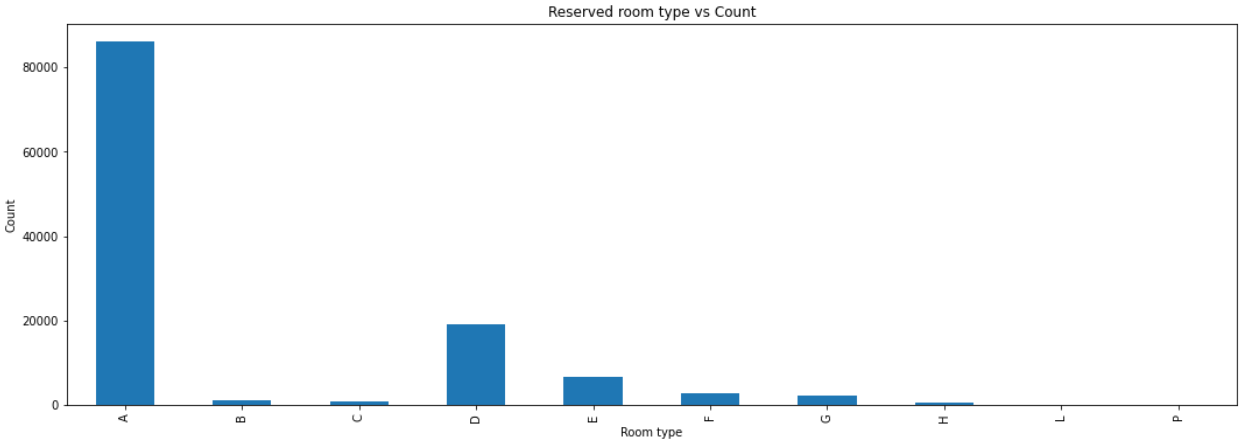
همانطور که مشخص است، بیشتر رزروها از طریق آژانس مسافرتی / اپراتور تور انجام می‌شود.

در ادامه بررسی می‌کنیم که چه نوع غذای combo توسط مردم به شدت ترجیح داده شد؟



اکثر مردم "BB" را در هر دو نوع هتل ترجیح می‌دهند. همچنین افراد کمی ترجیح می‌دهند "بسته غذا" را در هتل Resort ترجیح دهند. بنابراین اغلب آن‌ها غذا را در هتل Resort ترجیح می‌دهند.

همچنین بررسی می‌کنم که در زمان رزرو کدام نوع اتاق را افراد ترجیح می دهند؟



همانطور که مشخص است نوع اتاق "A" بیشترین متقاضی را دارد.

در آخر بررسی می‌کنم که آیا در زمان ثبت نام در اتاق، به همه اتاقهای مشابه اتاق انتخاب شده توسط آنها اختصاص داده شده است؟

reserved_assigned_status	0	1	All
hotel			
City Hotel	7192	72138	79330
Resort Hotel	7725	32335	40060
All	14917	104473	119390

همانطور که از نتایج مشخص است، در City Hotel، ۹ ٪ از bookings در زمان چک کردن به همان نوع اتاق نمی‌رسیدند. در Resort Hotel ، ۱۹ درصد از bookings در زمان چک کردن به همان نوع اتاق نمی‌رسیدند. این هتل‌ها نیاز به برداشتن گام‌های لازم برای تخصیص نوع اتاق ارجح به مشتریان دارند چون از پریشانی غیر ضروری در بین مشتریان اجتناب می‌کنند و همچنین به آن‌ها کمک می‌کند تا اعتماد مشتری را به دست آورند.

مرحله دوم؛ استفاده از روش‌های هوش مصنوعی برای پیش‌بینی

همانطور که در مرحله اول بدست آوردیم، نرخ لغو برای رزرو هتل در صنعت رزرواسیون آنلاین هتل‌ها بسیار بالا است. هنگامی که رزرو لغو شد، تقریبا هیچ کاری نمی‌شود کرد. این باعث ناراحتی بسیاری از هتل‌ها می‌شود و اشتیاق به انجام اقدامات احتیاطی را ایجاد می‌کند. بنابراین، پیش‌بینی reservations که می‌تواند لغو شود و جلوگیری از این حذف لغو شود یک مقدار اضافی برای هتل‌ها ایجاد خواهد کرد. لغو رزرو کردن تاثیر قابل‌توجهی بر تصمیمات مدیریت تقاضا در صنعت مهمان‌نوازی دارد. برای کاهش تاثیر حذف، هتل‌ها سیاست‌های لغو سخت و تاکتیک‌های overbooking را اجرا می‌کنند که به نوبه خود می‌تواند تاثیر منفی بر درآمد و شهرت هتل داشته باشد. برای کاهش این اثر، یک نمونه اولیه سیستم مبتنی بر یادگیری ماشین توسعه یافت. این سیستم از داده‌های سیستم‌های مدیریت املاک هتل استفاده می‌کند و هر روز یک مدل طبقه‌بندی را آموزش می‌دهد تا پیش‌بینی کند که رزرو شده "به احتمال زیاد" و با محاسبه تقاضای خالص است. این نمونه اولیه که در محیط تولید در دو هتل مستقر شده‌است، با اجرای آزمایش A/B، اندازه‌گیری تاثیر اقدامات انجام‌شده بر روی bookings را به احتمال زیاد لغو می‌کند. نتایج نشان می‌دهند که عملکرد نمونه اولیه خوب است و نشانه‌های مهمی برای پیشرفت تحقیقات ارایه می‌دهد در حالی که این است که bookings با هتل‌ها تماس گرفته تا با هواپیما تماس گرفته نشده تا با آن‌ها تماس گرفته نشده باشد.

حال برای انجام این مرحله، ابتدا باید از جدول Heatmap داده شده، ویژگی‌هایی که تاثیر بیشتری داشت را جدا می‌کنیم. (۵ ویژگی برتر)

```
is_canceled                     1.000000
lead_time                       0.293123
reserved_assigned_status        0.247770
total_of_special_requests       0.234658
required_car_parking_spaces     0.195498
booking_changes                 0.144381
previous_cancellations          0.110133
is_repeated_guest               0.084793
company                         0.082995
adults                          0.060017
previous_bookings_not_canceled  0.057358
days_in_waiting_list            0.054186
adr                             0.047557
agent                           0.046529
babies                          0.032491
stays_in_week_nights            0.024765
adr_pp                          0.017808
arrival_date_year               0.016660
arrival_date_week_number        0.008148
arrival_date_day_of_month       0.006130
children                        0.005036
stays_in_weekend_nights         0.001791
Name: is_canceled, dtype: float64
```

سپس برای این پروژه من از دو روش استفاده کردم:

دسته‌بندی بیز ساده

در حوزه یادگیری ماشین تکنیک و روش دسته بند بیز ساده، با بکارگیری قضیه بیز و فرض استقلال بین متغیرها، به عنوان عضوی از خانواده دسته‌بندهای برمبنای احتمال (Probabilistic Classifiers) قرار می‌گیرد. در سال‌های ۱۹۶۰ تحقیق و بررسی‌های زیادی پیرامون بیز ساده بخصوص در زمینه بازیابی متن صورت گرفت و حتی امروز هم به عنوان ابزاری برای دسته‌بندی متن (Text Categorization) برای حل مسائلی مانند تشخیص هرزنامه‌ها (Spam Mails) به کار می‌رود. معمولا این کار به کمک برآورد تابع احتمال و از طریق فراوانی یا فراوانی نسبی کلمات در اسناد متنی صورت می‌گیرد. به این ترتیب به کمک حداکثرسازی تابع درست‌نمایی(Likelihood maximization) برآورد پارامترهای مدل میسر می‌شود. در حوزه آمار و دانش رایانه، مدل بیز ساده با نام‌های دیگری نظیر بیز ساده (Simple Bayes) و بیز مستقل نیز شناخته می‌شود که در بسیاری از حوزه‌های دیگر نیز کاربرد دارد. به منظور آشنایی خوانندگان با این زمینه کاربردی بخصوص برای افرادی که به تازگی خواستار ورود به گروه متخصصان علم داده هستند، متن حاضر را تهیه کرده‌ایم.

برای تقسیم‌بندی داده‌های train و تست از k-fold cross validation استفاده کردم و $k = 10$ را انتخاب کردم به این صورت که هر بار ۹ بخش را به عنوان داده train و ۱ بخش را به عنوان داده test در نظر می‌گیریم. سپس error میانگین را به دست آورده و accuracy را بدست می‌آوریم.

دقت بدست‌آمده برابر است با:

	precision	recall	f1-score	support
0	0.90	0.18	0.30	75166
1	0.41	0.97	0.58	44224
accuracy			0.47	119390
macro avg	0.66	0.57	0.44	119390
weighted avg	0.72	0.47	0.40	119390

دقت در این مرحله بدست آمده برابر با ۴۷ درصد است.

حال در همین روش، همانطور که در توضیح پروژه خواسته شده بود، از الگورتیم PCA برای پیدا کردن ویژگی‌هایی که بیشترین تاثیر دارد را به صورت یک الگوریتم feature extraction based انتخاب شده است و یک سری ویژگی جدید در واقع تولید می‌شود.

نتایج الگوریتم بیزساده بر اساس استفاده از PCA به شکل زیر است:

	precision	recall	f1-score	support
0	0.68	0.85	0.76	75166
1	0.57	0.33	0.42	44224
accuracy			0.66	119390
macro avg	0.63	0.59	0.59	119390
weighted avg	0.64	0.66	0.63	119390

همانطور که مشخص است، دقت بهبود پیدا کرد! این نشان از عملکرد قوی الگوریتم‌های کاهش ابعاد مثل PCA هست.

دسته‌بندی جنگل تصادفی

الگوریتم جنگل تصادفی(Random Forest) یک الگوریتم یادگیری ماشین با قابلیت استفاده آسان است که اغلب اوقات نتایج بسیار خوبی را حتی بدون تنظیم فراپارامترهای آن، فراهم می‌کند. این الگوریتم به دلیل سادگی و قابلیت استفاده، هم برای دسته‌بندی (Classification) و هم رگرسیون، یکی از پر کاربردترین الگوریتم‌های یادگیری ماشین محسوب می‌شود. در این مطلب، چگونگی عملکرد جنگل تصادفی و دیگر مباحث مهم پیرامون آن مورد بررسی قرار خواهند گرفت. برای درک چگونگی عملکرد جنگل تصادفی، ابتدا باید الگوریتم درخت تصمیم (Decision Tree)که بلوک سازنده جنگل تصادفی است را آموخت. انسان‌ها همه روزه از درخت تصمیم برای تصمیم‌گیری‌ها و انتخاب‌های خود استفاده می‌کنند، حتی اگر ندانند آنچه که از آن بهره می‌برند نوعی الگوریتم یادگیری ماشین است.

	precision	recall	f1-score	support
0	0.79	0.87	0.83	75166
1	0.74	0.60	0.66	44224
accuracy			0.77	119390
macro avg	0.76	0.74	0.74	119390
weighted avg	0.77	0.77	0.77	119390

همانطور که در شکل بالا مشخص است، دقت بیشتری نسبت به بیز ساده بدست آورد که بر اساس قواعد یادگیری ماشین این نتیجه قابل انتظار بود.

پایان