

دیگر اتاق‌ها خالی نمی‌مانند

مدل سازی توزیع احتمالاتی و توصیف عدم قطعیت داده‌های هتل‌داری

با محوریت بررسی مسئله کلاسیک Overbooking

دانشکدگان فارابی دانشگاه تهران

پروژه درس تئوری احتمال و کاربردهای آن

استاد درس: خانم مهندس یزدی

• محمدرضا صالح زاده

• سید علیرضا میرتونی

• سید محمد صالح قافله‌باشی

• سپهر معماریان

چکیده

در این گزارش، به بررسی پایگاه داده‌ای از یک هتل پرداخته و سعی میکنیم با استفاده از اصول تئوری احتمالات و آمار، عدم قطعیت موجود در این داده‌ها را توصیف کنیم.

برای این امر، به بررسی مسئله کلاسیک بیش‌فروشی یا Overbooking می‌پردازیم و سعی میکنیم این مسئله را برای هتل مورد نظر با اتکا به داده‌های ارائه شده حل کنیم. در حین ارائه مدلی برای بیش‌فروشی در این هتل، توزیع‌های احتمالاتی و شرطی متغیرهای مختلف، به ویژه نرخ کنسلی و غیبت و پارامترهای آن‌ها را بررسی خواهیم کرد.

۱	چکیده
۳	Overbooking بیش فروشی
۳	Overbooking یا بیش فروشی چیست؟
۴	استفاده از توزیع های احتمالاتی برای مدل سازی بیش فروشی
۴	Binomial Distribution استفاده از توزیع دوجمله ای
۷	محاسبه عددی (Empirical) توزیع تعداد کنسلی های روزانه
۱۰	جمع بندی استفاده از توزیع های احتمالاتی
۱۲	استفاده از مدل های یادگیری ماشین برای پیش بینی بیش فروشی
۱۲	استفاده از رگرسیون خطی برای پیش بینی میزان بیش فروشی مناسب
۱۵	دیگر توضیحات مورد انتظار گزارش
۱۵	Data preparation آماده سازی داده ها
۱۶	Conditional Distributions توزیع های شرطی
۱۷	بررسی بصری همبستگی بین تقاضا و امتیازدهی
۱۸	منابع

Overbooking | بیش‌فروشی

بیش‌فروشی یا Overbooking چیست؟

در اواسط قرن نوزدهم، شرکت‌های ایرلاین‌های ایالات متحده آمریکا، تخمین زدند که حدود نیمی از بلیت‌ها سرانجام کنسل شده یا دارندگان آن‌ها برای پرواز حضور پیدا نمی‌کنند و خالی ماندن تعداد قابل توجهی از صندلی‌ها در هر پرواز به مسئله مهمی تبدیل شد. این شرکت‌ها به پیشنهاد تیم‌های Operation Research اقدام به فروش کنترل شده و بیش از ظرفیت بلیت نمودند، تا کنسلی‌ها و غیبت‌ها از ظرفیت مازاد کاسته و دیگر صندلی‌ها خالی نماند.

مسئله اینجاست که چقدر باید بلیت بیشتری فروخت تا ضرر کنسلی و صندلی‌های خالی، کمینه شده و از طرف دیگر احتمال بیشتر بودن تعداد مسافران حاضر شده، از ظرفیت هواپیما، کمینه شود. در هتل‌داری نیز چنین است. در بررسی داده‌های هتل، متوجه شدیم ۲۹٫۸۵ درصد^۱ از اتاق‌های رزرو شده، خالی مانده‌اند. یعنی مسافران برای اقامت حضور نیافته یا قبل از موعد تحویل اتاق، رزرو را لغو کرده‌اند.

Row Labels	Count of booking_id
Cancelled	24.83%
Checked Out	70.15%
No Show	5.02%
Grand Total	100.00%

^۱ محاسبه این مقدار با استفاده از pivot tables در Excell بدست آمده است. جدول فوق یکی از همین pivot table ها است. این جدول در فایل اکسل پیوست در شیت probability موجود است.

لذا این هتل میتواند بیشتر از ظرفیت خود اتاق اجاره دهد تا پس از کنسلی‌ها و عدم حضور برخی از مسافران، اتاق‌های کمتری خالی بمانند. اکنون به دنبال رسیدن به میزانی از پذیرش بیش از ظرفیت هستیم تا ریسک حضور دو متقاضی برای یک اتاق، کمینه و حساب شده باشد.^۲

استفاده از توزیع‌های احتمالاتی برای مدل‌سازی بیش‌فروشی

به طور کلی مدل‌های پویا و ایستا برای حل مسئله بیش‌فروشی پیشنهاد داده شده‌اند، که در این گزارش به بررسی چند مورد از مدل‌های احتمالاتی و ایستا می‌پردازیم. در مدل‌های ایستا، از دینامیک رفتار مشتریان و تاثیر هر کنسلی بر احتمال کنسل شدن بلیت‌های دیگر چشم‌پوشی می‌شود.

استفاده از توزیع دوجمله‌ای | Binomial Distribution

ساده‌ترین مدل ایستا برای تحلیل کنسلی‌ها، بر پایه مدل دوجمله‌ای استوار است که در آن، موارد «عدم مراجعه» (No-shows) نیز در دسته کنسلی‌ها لحاظ می‌شوند.^۳ یعنی هر رزرو دو رو دارد، دارنده آن یا حاضر می‌شود یا رزرو را لغو می‌کند. این مدل بر مفروضات زیر استوار است:

- استقلال تصمیم‌گیری: مشتریان به صورت مستقل از یکدیگر اقدام به لغو رزرو می‌کنند.
- احتمال یکسان: هر مشتری احتمال یکسانی برای لغو رزرو خود دارد.

^۲ می‌دانیم مسئله بیش‌فروشی زمانی مطرح می‌شود که تقاضا از ظرفیت بیشتر باشد. یعنی تقاضایی برای ظرفیت خالی مانده از کنسلی‌ها و غایبان وجود داشته باشد. معیار تقاضا برای هتل این گزارش ($\text{Demand} = \text{bookings}/\text{capacity}$) هرگز از ظرفیت عدول نکرده است؛ اما به دلیل کاربردی بودن مسئله بیش‌فروشی، از این مورد چشم‌پوشی شده و مدل بیش‌فروشی برای زمانی که تقاضا برای این هتل بیشتر از ظرفیت اتاق‌های آن شود توسعه داده شده است.

^۳ به این معنا که عدم مراجعه، صرفاً نوعی کنسلی تلقی می‌شود که در روز ارائه خدمات رخ داده است.

فرض می‌کنیم:

تعداد مجاز پذیرش‌های بیش از ظرفیت در یک روز: k

تعداد کنسلی‌ها در یک روز: C

خطا و ریسک مورد نظر: α

ظرفیت یا همان تعداد تمام اتاق‌ها: n

احتمال کنسل شدن یک رزرو: p

ما به دنبال بیشینه مقدار k هتسیم به طوری که:

$$k^* = \max\{k \in \mathbb{Z} : P(C < k) \leq \alpha\}, \quad C \sim \text{Binomial}(n, p)$$

یعنی احتمال فراتر رفتن بیش‌فروشی‌ها از کنسلی‌ها، کمتر از خطای مورد انتظار شود. خطای مورد انتظار بسته به ریسک‌پذیری و ملاحظات مدیریت هتل می‌تواند کم یا زیاد شود. برای محاسبه از تابع توزیع تجمعی تعداد کنسلی‌ها در یک روز استفاده می‌کنیم:

$$k^* = \max \left\{ k : \sum_{i=1}^k \binom{n}{i} p^i (1-p)^{n-i} \leq \alpha \right\}$$

$$k^* = F_{\text{Binomial}(n,p)}^{-1}(\alpha)$$

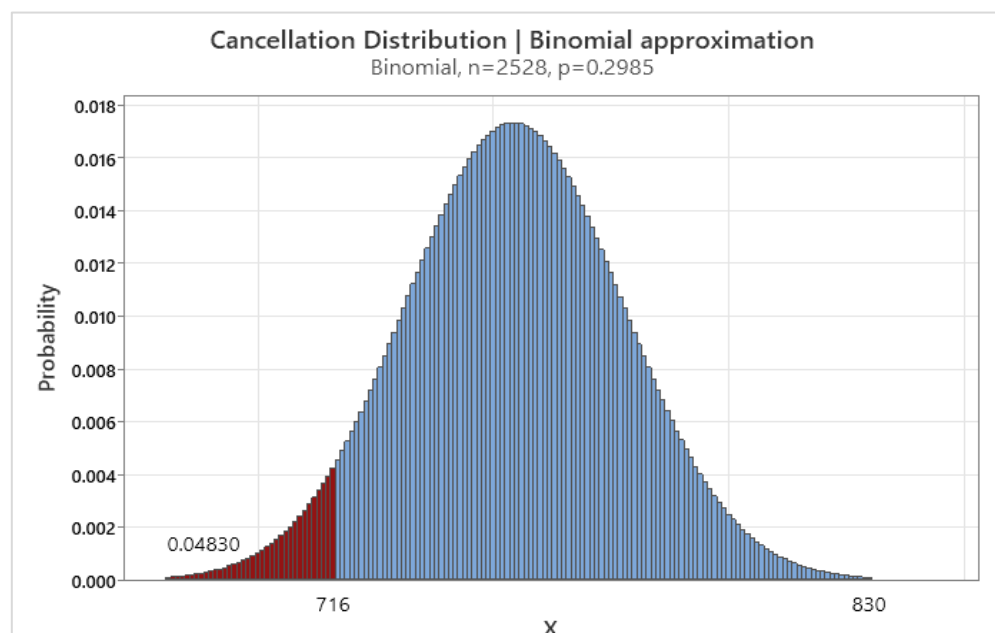
میزان بهینه‌ی بیش‌فروشی، بزرگ‌ترین عدد صحیح k تعریف می‌شود؛ به طوری که احتمال رخ دادن کمتر از k کنسلی، از سطح ریسک‌پذیری مورد انتظار α فراتر نرود.

برای تخمین کل میزان بیش‌فروشی مجاز هتل، کافی است قرار داده شود $n=2528$ ، یعنی تمام ظرفیت هتل در شعب مختلف. همچنین احتمال کنسلی نیز ۰٫۲۹۸۵ می‌باشد.

میتوان با محاسبه کامپیوتری^۴ توزیع تجمعی، برای ریسک دلخواه، میزان بیش‌فروشی مناسب را پیدا کرد. به عنوان مثال، مدل دوجمله‌ای با ریسک ۵ درصد، ادعا کرده است که میتوان ۷۱۶ اتاق را برای روز هایی که تمام ظرفیت رزرو شده است، اضافه بر ظرفیت پذیرش کرد.

K	$P(C < k)$
690	0.0025
700	0.009
716	0.05
720	0.07

جدول ۱: بیش‌فروشی بهینه اگر توزیع میزان کنسلی روزانه بر پایه توزیع دوجمله‌ای باشد



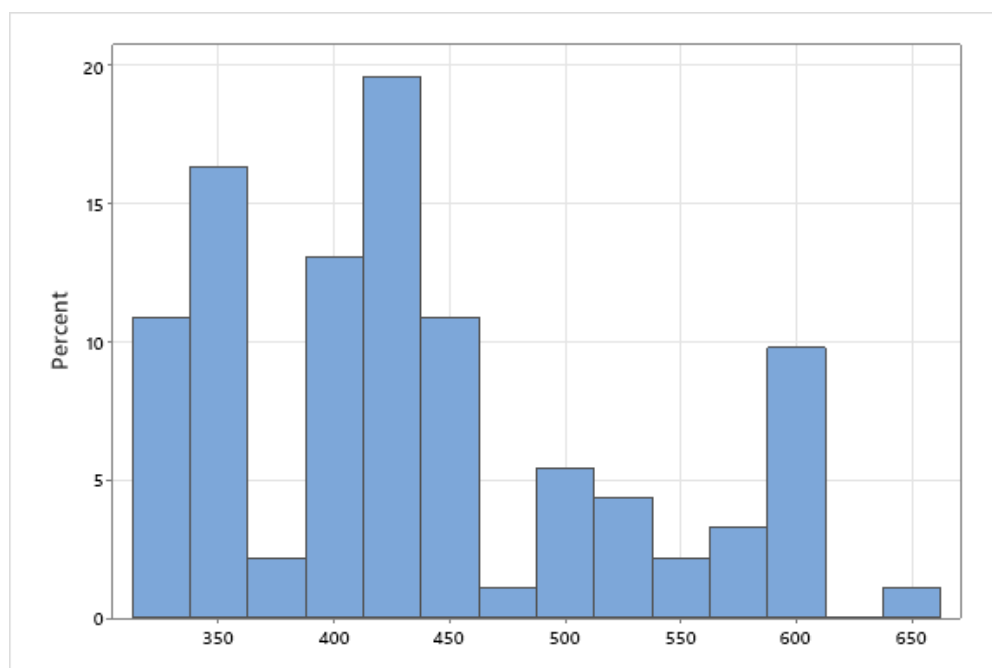
تصویر ۱: اگر توزیع میزان کنسلی روزانه بر پایه توزیع دوجمله‌ای باشد، هیستوگرام آن به شکل بالا خواهد شد. قسمت هاشور خورده، صدک پنجم و ابتدایی توزیع را نشان می‌دهد که معدل ۷۱۶ اتاق می‌باشد.

^۴ برای محاسبه سریع مقادیر جدول می‌توان از ماشین حساب آنلاین [stat trek](http://stat.trek) استفاده نمود.
^۵ برای رسم نمودار های آماری، از نرم افزار Minitab استفاده شده است. اصل نمودارها در فایل پیوست موجوداند.

محاسبه عددی (Empirical) توزیع تعداد کنسلی‌های روزانه

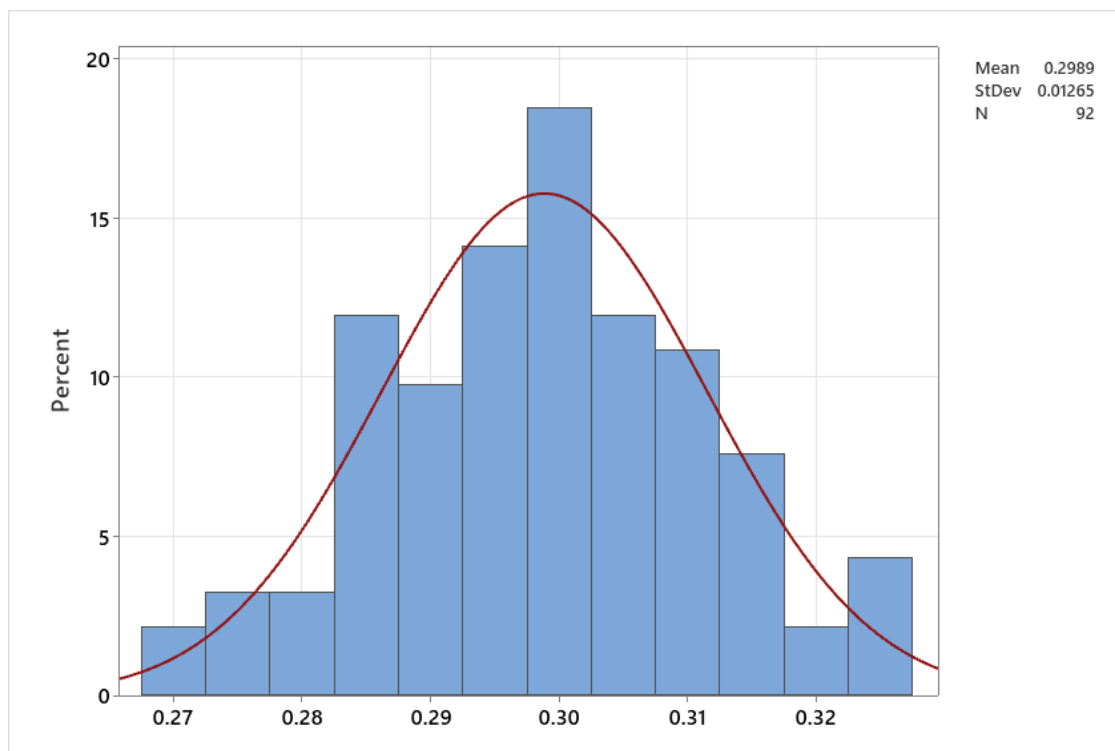
حال سوال اینجاست که آیا داده‌های موجود از هتل، دوجمله‌ای بودن توزیع کنسلی‌ها را تایید می‌کنند؟

شکل زیر، توزیع تعداد کنسلی‌ها در روز را نشان می‌دهد.



تصویر ۲: هیستوگرام تجربی تعداد کنسلی‌های روزانه

مشخص است که تعداد کنسلی‌ها از توزیع خاصی پیروی نمی‌کند. میانگین آن ۴۳۶ بوده و واریانس آن ۷۳۹۰ می‌باشد که بشدت بالاست. این نتیجه در ظاهر، پیش‌بینی میزان بیش‌فروشی بهینه را سخت‌تر می‌کند. اما در بررسی داده‌های هتل، متوجه شدیم که نسبت کنسلی‌ها به کل رزروها در هر روز، گویا تفاوت چندانی نمی‌کنند، لذا به بررسی توزیع این نسبت در روزهای مختلف پرداختیم.



تصویر ۳: توزیع نسبت روزانه‌ی کنسلی‌ها به کل رزروها

همانطور که پیداست، این نسبت توزیعی نسبتاً نرمال دارد^۶ و انحراف معیار و واریانس آن بشدت پایین است به طوری که توزیع به خوبی حول میانگین متمرکز است.

پس شاید بهتر باشد بجای تعداد، بر نرخ - نسبت کنسلی‌های روزانه تمرکز کنیم. اکنون به دنبال بیش‌فروشی حداکثر نسبتی از اتاق‌هایمان هستیم که مطمئن باشیم نسبت کنسلی در آن روز کمتر از بیش‌فروشی انجام شده نباشد. داریم:

$$R_{rate} = \frac{C_{cancellations}}{N_{bookings}}$$

$$r^* = \max\{r \in [0,1]: P(R < r) \leq \alpha\}, \quad F_R(r) = P(R \leq r)$$

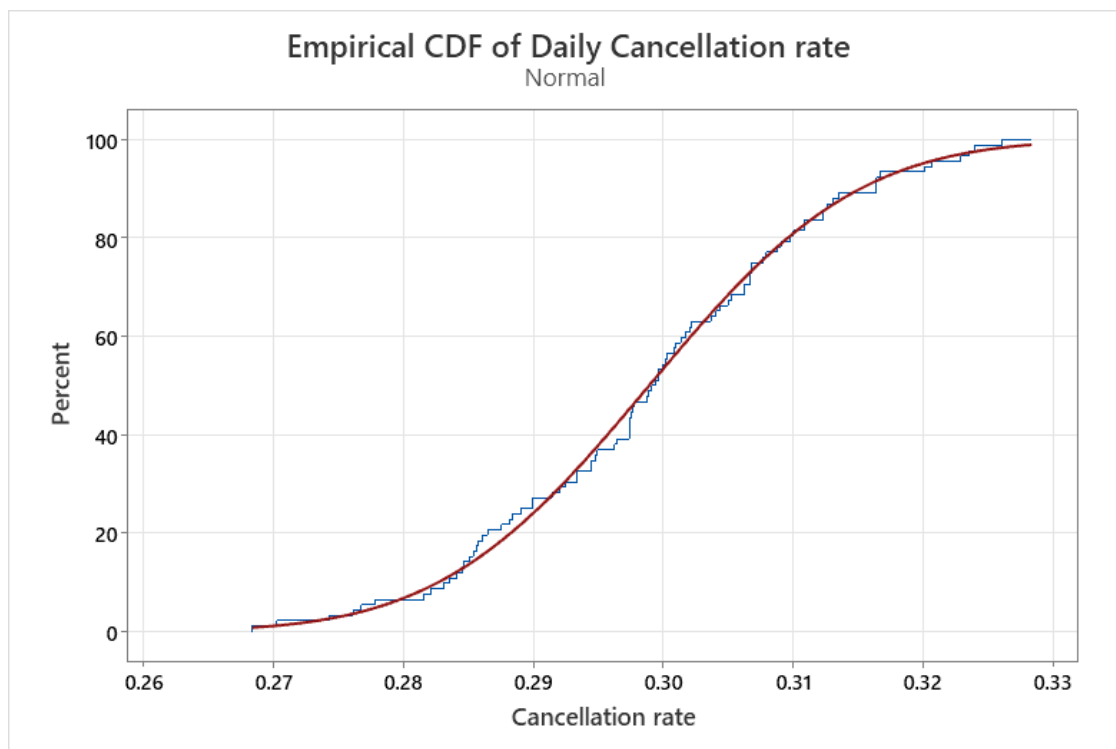
برای نرخ و تعداد^۷ بیش‌فروشی بهینه داریم:

$$r^* = \max\{r \in [0,1]: F_R(r) \leq \alpha\}, \quad k^* = N r^*$$

^۶ در آزمایش نرمال بودن، مقدار p-value حدود ۰/۸۹۵ به دست آمد که فرض نرمال نبودن را تقریباً رد می‌کند.

^۷ واضح است که تعداد بیش‌فروشی بهینه از حاصل ضرب تعداد رزروها در نرخ بهینه بدست می‌آید.

کافی است تابع توزیع تجمعی نرخ کنسلی را محاسبه کرده و با ریسک دلخواه آلفا، نرخ بهینه را بدست آوریم.



r	$r * 2528 = k^*$	$F_R(r) = \alpha$
0.273	690	2%
0.275	695	3%
0.278	703	5%

تصویر ۴ و جدول ۲: بدست آوردن تعداد بهینه پیش‌فروشی

طبق جدول ۲، که از نمودار تصویر ۴ محاسبه شده است، این هتل برای روزی که تمام ظرفیت اش رزرو شود، می‌تواند با ریسک ۵ درصد، ۷۰۳ اتاق بیش از ظرفیت یعنی ۳۲۳۱ اتاق را برای رزرو ارائه کند.

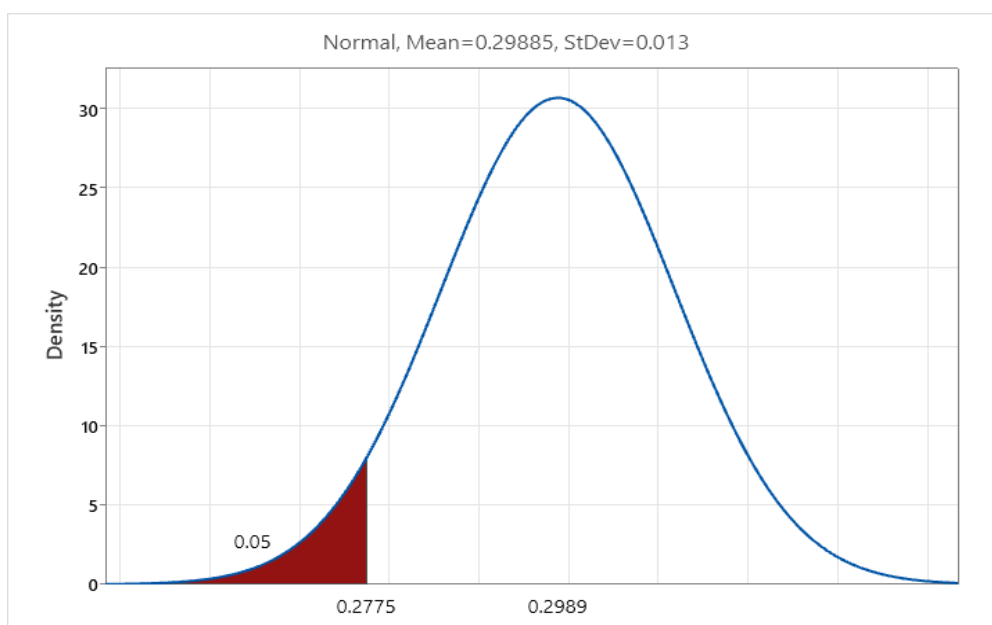
جمع‌بندی استفاده از توزیع های احتمالاتی

۱. توزیع کنسلی‌های روزانه:

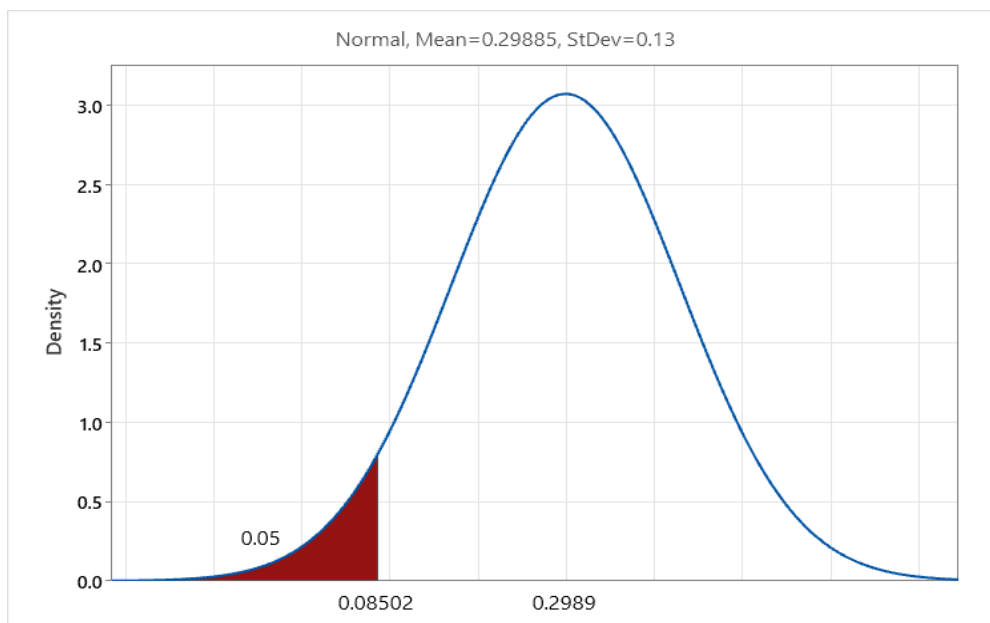
برای تعداد بیش‌فروشی، از طریق هر دو مدل، هم binomial و هم Empirical به حدود ۷۰۰ اتاق رسیدیم که نشان می‌دهد در داده‌های این هتل، واقعه لغو رزرو یا همان کنسلی، تا حد خوبی مانند پرتاب سکه، از توزیع دو جمله‌ای پیروی می‌کند. نیاز به تکرار نیست که ما عدم حضور و کنسلی را در این مدل یکسان فرض کردیم.

۲. تاثیر واریانس بر تعداد بیش‌فروشی:

سوال اینجاست اگر واریانس در توزیع نرخ کنسلی، بیشتر بود، چه اتفاقی می‌افتاد؟ خوب است نگاهی کلی به توزیع نرخ کنسلی بیندازیم. توضیح دادیم که این نرخ تا حد خوبی با توزیع نرمال نیز توصیف می‌شود، برای جلوگیری از پیچیدگی، صرفاً توزیع‌های نرمال متناظر را رسم می‌کنیم.

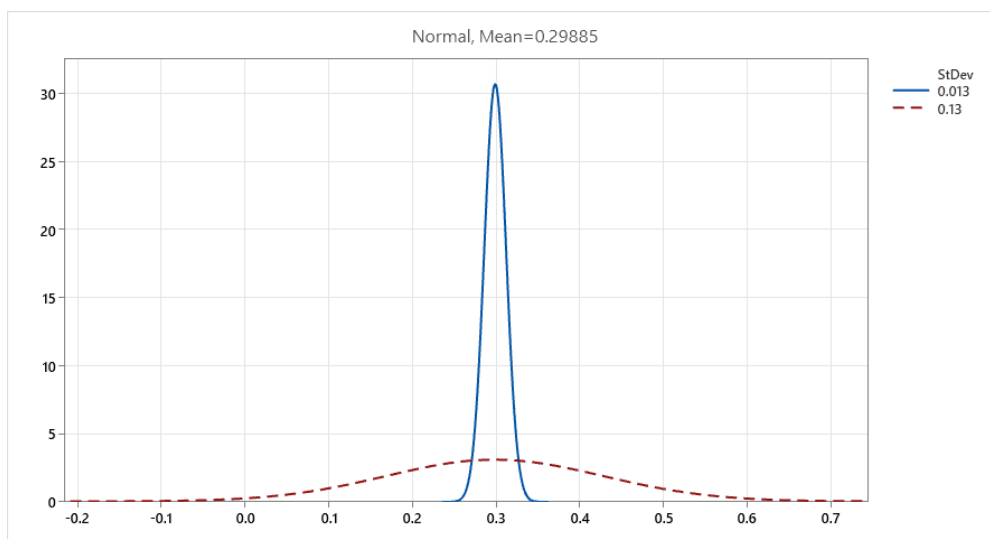


تصویر ۵: توزیع نرمال کنسلی‌ها با واریانس و میانگین واقعی داده‌ها



تصویر ۶: توزیع نرمال کنسلی‌ها با واریانس ۱۰۰ برابر داده‌های واقعی

به عنوان مثال با ۱۰ برابر نمودن انحراف معیار (تصویر ۶)، نرخ بهینه ۰٫۰۹ می‌شود، در نتیجه میزان بیش‌فروشی به ۲۲۷ اتاق کاهش می‌یابد. به عبارتی هرچه پراکندگی نرخ کنسلی بیشتر باشد، ریسک میزان معینی از بیش‌فروشی افزایش می‌یابد. شکل پایین به دلیل کشیدن هر دو توزیع در یک پلات، واضح‌تر این مفهوم را بیان می‌کند.



۳. قانون حد مرکزی در داده‌های نرخ کنسلی روزانه:

در تصویر شماره یک، توزیع دوجمله‌ای برای ۲۵۲۸ آزمایش، یعنی تمام اتاق‌های هتل رسم شده است، همچنین تصویر ۵ نیز، تخمینی نرمال از توزیع واقعی داده‌هاست. غیر از شکل توزیع، دو توزیع مقدار بهینه بیش‌فروشی برای ۲۵۲۸ اتاق را، تنها ۱۳ عدد متفاوت از یکدیگر گزارش می‌کنند. یعنی توزیع دوجمله‌ای نرخ کنسلی، در تعداد آزمایش‌های بالا، به توزیع نرمال میل کرده است.

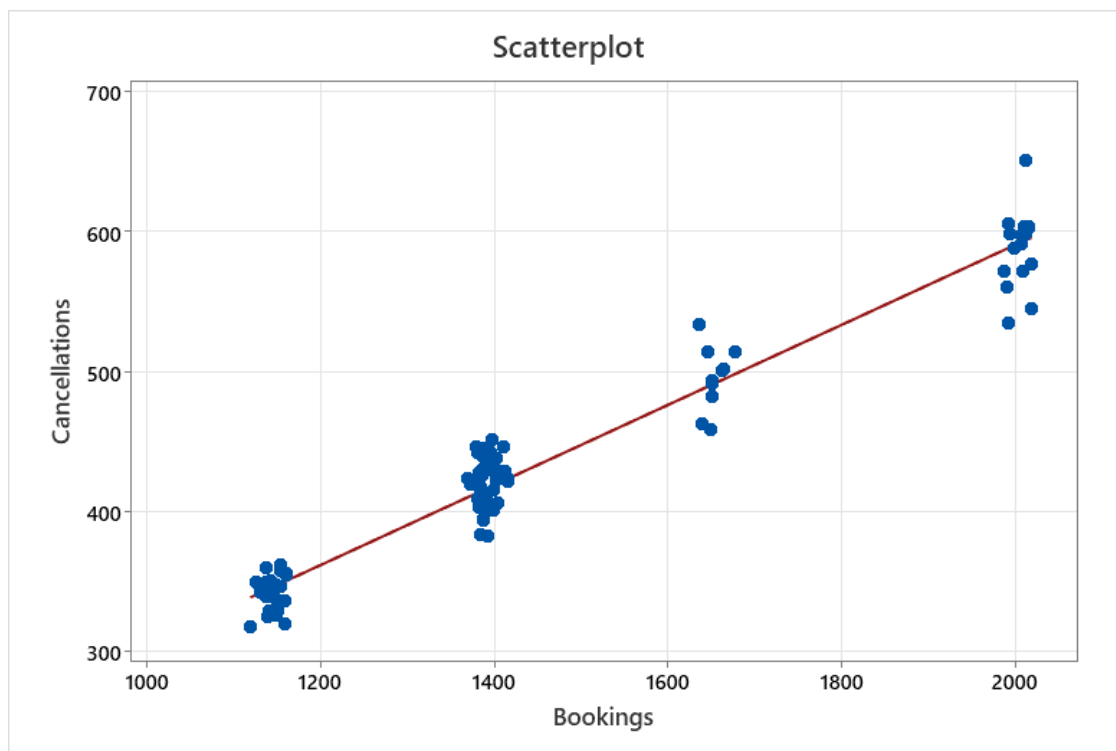
استفاده از مدل‌های یادگیری ماشین برای پیش‌بینی بیش‌فروشی

یکی دیگر از راه‌های پیش‌بینی تعداد بیش‌فروشی مناسب، استفاده از الگوریتم‌های یادگیری ماشین برای استخراج الگوهای نهفته در داده‌های کنسلی است.^۸ در این مدل‌ها، دیگر خبری از یکسان فرض کردن احتمال کنسل شدن رزروها و دیگر مفروضات سختگیرانه نیست و پیش‌بینی‌ها کاملاً داده‌محور است.

استفاده از رگرسیون خطی برای پیش‌بینی میزان بیش‌فروشی مناسب

در توزیع نرخ کنسلی‌ها دیدیم که این نرخ واریانس بالایی نداشته و حول میانگین آن متمرکز است. این درحالی است که تعداد کنسلی‌ها به شدت پراکنده بوده و نوسان زیادی دارد. نگاهی به توزیع توأم تعداد رزروها و تعداد کنسلی‌ها کاملاً موضوع را روشن می‌کند.

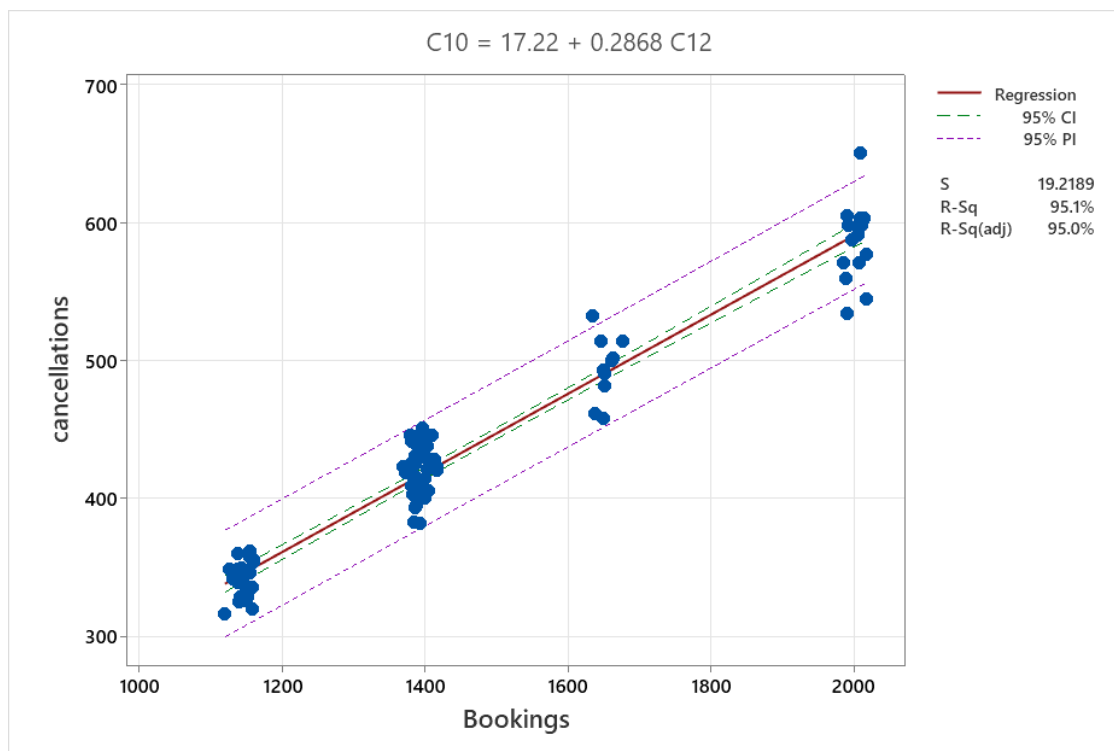
^۸ Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior, Nuno Antonio, Ana de Almeida and Luis Nunes



تصویر ۷: نمودار پراکندگی تعداد کل رزرو ها و تعداد کنسلی ها

نمودار بالا رابطه‌ای کاملاً خطی بین تعداد رزرو های روزانه و تعداد کنسلی ها در همان روز را نشان می‌دهد، حال میتوانیم با استفاده از تحلیل رگرسیون خطی، با تکیه بر داده‌های قبلی، پیش‌بینی کنیم که برای هتل ما با ۲۵۲۸ ظرفیت، چند کنسلی اتفاق خواهد افتاد. جالب اینجاست که این هتل در داده‌های خود تا به حال روزی با تعداد رزرو ۲۵۲۸ اتاق بیشتر نداشته است، پس مقایسه پیش‌بینی آن با توزیع های احتمالاتی و هم‌خوانی داشتن یا نداشتن آنها با یکدیگر، معیار خوبی برای ارزیابی مقادیر بهینه بدست آمده برای بیش‌فروشی خواهد بود.

مدل ذیل، با دقت ۹۵ درصد (PI: Predicted Interval)، حداقل ۷۰۱ اتاق را برای بیش‌فروشی در ۲۵۲۸ اتاق پیشنهاد داده است.



تصویر ۸

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
19.2189	95.06%	95.00%	94.78%

Prediction

Fit	SE Fit	95% CI	95% PI
742.143	7.60981	(727.025, 757.262)	(701.078, 783.209) XX

مدل	میزان بیش‌فروشی پیشنهادی
Binomial	716
Empirical-Normal	703
Linear Regression	701

جدول ۳: مقایسه تعداد بیش‌فروشی‌های بهینه با مدل‌های مختلف

دیگر توضیحات مورد انتظار گزارش

آماده سازی داده ها | Data preparation

در داده های این هتل، مقادیر پرت یا نیاز به اصلاح دیده نشد و داده نیاز به پیش پردازش خاصی نداشت. تنها Rating حدود ۵۸ درصد از سطر ها خالی بودند که با فیلتری ساده حذف شدند.^۹

^۹ فایل اکسل پیوست، شیت preprocessing

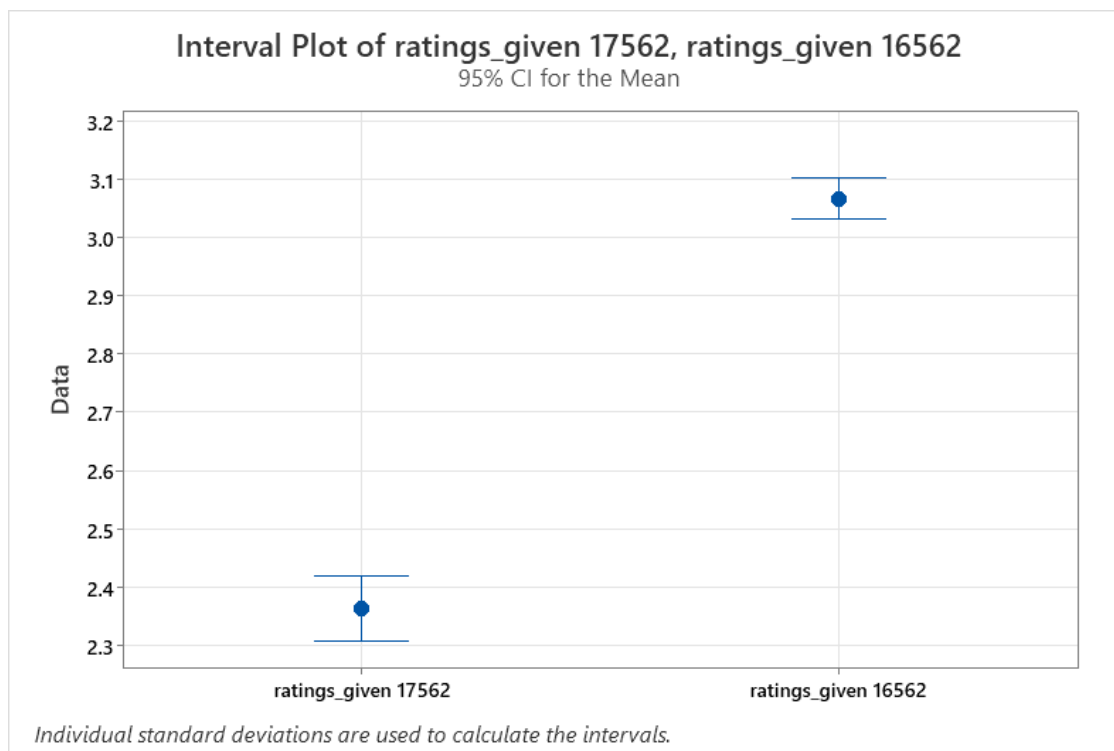
توزیع های شرطی | Conditional Distributions

Property_id	Average of ratings_given	Var of ratings_given
16558	4.251	1.116
16559	4.325	1.104
16560	3.014	0.668
16561	4.277	1.142
16562	3.067	0.659
16563	4.270	1.126
17558	3.051	0.636
17559	4.321	1.077
17560	3.035	0.602
17561	4.298	1.138
17562	2.364	1.151
17563	4.286	1.187
17564	2.295	1.052
18558	3.070	0.640
18559	2.328	1.120
18560	4.264	1.144
18561	4.248	1.253
18562	4.304	1.124
18563	3.071	0.648
19558	2.363	1.094
19559	3.048	0.621
19560	4.282	1.151
19561	3.079	0.639
19562	4.277	1.194
19563	3.019	0.671
Grand Total	3.619	1.525

جدول ۴

میانگین امتیازهای داده شده برای هتل های مختلف تفاوت زیادی می کند و این نشان می دهد که شعب مختلف تاثیر مستقیم و مستقلى بر میانگین کل امتیاز داده شده به هتل دارند^{۱۰}. تصویر پایین نیز دو هتل را از منظر امتیاز مسافران دقیق تر مقایسه می کند.

^{۱۰} فایل اکسل پیوست، شیت Hotel_By_Rate



تصویر ۹

بررسی بصری همبستگی بین تقاضا و امتیازدهی

همانطور که اشاره شد، در این گزارش شاخصی به اسم Demand تعریف شد که نسبت تمام رزروها به ظرفیت هتل را نشان می‌دهد. یعنی به ازای هر اتاق چقدر تقاضا وجود دارد. برای هتل‌های مختلف، بین میانگین امتیاز داده شده و میانگین تقاضای روزانه، رابطه‌ای قوی وجود دارد. تقاضا برای هتل‌هایی که میانگین امتیاز بیشتری دارند، بیشتر است.

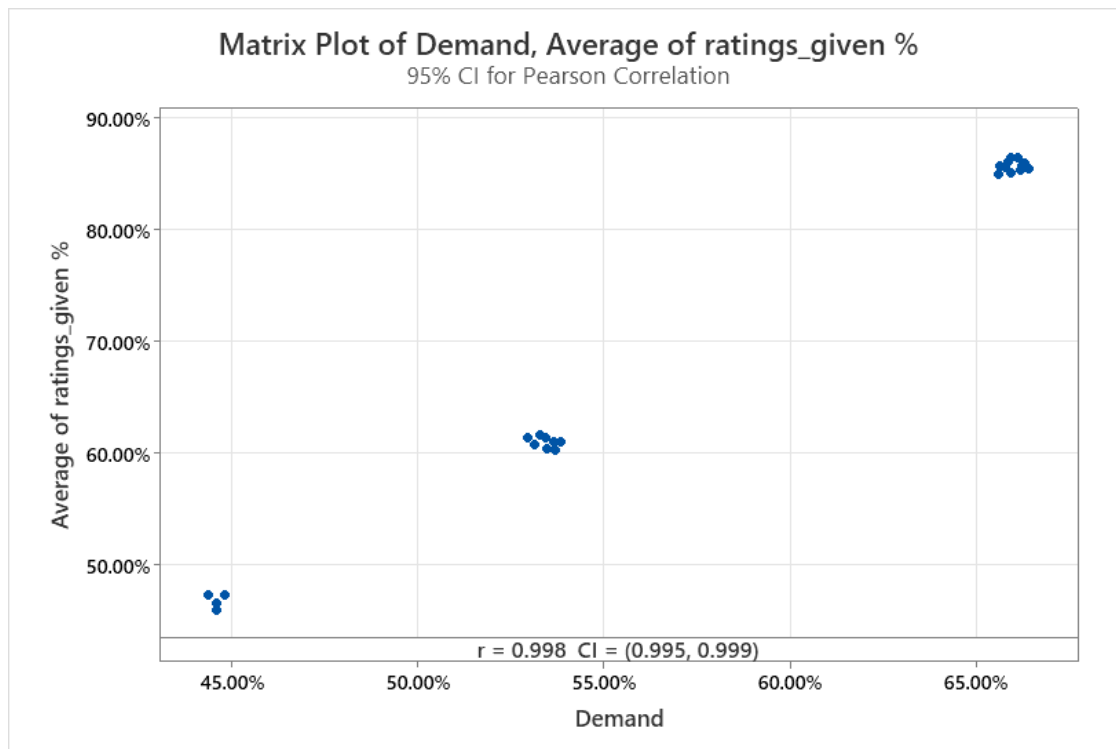
برای ماتریس همبستگی داریم:

$$\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 0.02430627 & 0.01296608 \\ 0.01296608 & 0.00694886 \end{pmatrix}$$

$$\sigma_X = \sqrt{0.02430627}, \quad \sigma_Y = \sqrt{0.00694886}$$

$$\rho_{XY} = \frac{0.01296608}{\sigma_X \sigma_Y} \approx 0.998$$



منابع

- فصل چهارم کتاب the theory and practice of Revenue management با عنوان Overbooking از انتشارات springer بسیار الهام بخش بود.
- Introduction to probability, CRC Press, Stanford University