به نام خداوند بخشنده ی مهربان

عنوان پژوهش و گزارش:

(Airbnb)

نگارش:

نگار درویشی

فروردین ۱۴۰۰

چکیده

از سال ۲۰۰۸ ؛ (ار بی ان بی) به میهمانان و میزبانان کمک کرده است تا به روشی منحصر به فرد و شخصی تر سفر کنند. این شرکت از یک تشک بادی برای اجاره به همکاری جهانی, با ارزش بیش از ۳۰ میلیارد دلار رسید.

Airbnb یک بازار آنلاین است که افرادی که می خواهند خانه های خود را اجاره دهند با افرادی که به دنبال محل اقامت در آن محل هستند ، ارتباط برقرار می کنند. در حال حاضر بیش از ۱۰۰۰۰۰ شهر و ۲۲۰ کشور در سراسر جهان را پوشش می دهد. نام این شرکت از "تشک بادی B&B" گرفته شده است.

برای میزبانان ، شرکت در Airbnb راهی برای کسب درآمد از دارایی آنها است ، اما با این خطر که مهمان به آن آسیب برساند. برای میهمانان ، مزیت می تواند اقامتگاه های نسبتاً ارزان باشد ، اما با این خطر که ملک آنچنان که لیست به نظر می رسد جذاب نخواهد بود.

ما سعی داریم در اینجا به تجزیه و تحلیل اطلاعات و داده هایی از این موضوع به کمک استنتاج های آماری و داده ها بپردازیم.

فهرست مطالب

مقدمه

از هزاران لیست در شهرهای مختلف ، (ار بی ان بی) به یک غرق بزرگ اطلاعات تبدیل شده است. داده های ارائه شده توسط صاحبان خانه اغلب بزرگ ، نامرتب و در عین حال فوق العاده مفید هستند. هدف این پروژه استخراج دانش از این مجموعه داده ها با استفاده از تکنیک ها و روش های متداول در علم داده ها است.

همانطور که بیشتر صاحبان خانه املاک خود را روی سیستم عامل قرار می دهند ، (ار بی ان بی) قادر است قیمت های مناسب را برای لیست ها براساس مدل های یادگیری ماشین آموزش داده شده در مجموعه های بزرگ داده پیشنهاد دهد.

با هدف پیش بینی قیمت های لیست در شهر نیویورک ، به صاحبان خانه اجازه می دهد تا ملک خود را به طور مناسب قیمت گذاری کنند. به طور خاص ، ما به دنبال پاسخ به این سوال هستیم: (ار بی ان بی) با توجه به مجموعه ای از ویژگی های لیست ، چه قیمت هایی را باید به میزبان خود پیشنهاد دهد؟

این سوال مهم است زیرا با در دسترس قرار دادن اطلاعات بیشتر ،تجربه برای میزبان و مشتری را بهبود می بخشد.

(ار بی ان بی) از ابتدای تأسیس خود در سال ۲۰۰۸ و با افزایش تعداد اجاره های ذکر شده ، هر ساله رشد شهابی را شاهد بوده است.

(ار بی ان بی) با موفقیت هرچه بیشتر ، صنعت مهمان نوازی سنتی را مختل کرده است ، نه تنها مسافرانی که به دنبال جنجال و جذابیت هستند بلکه مسافران تجاری نیز به عنوان ارائه دهنده اصلی محل اقامت خود به (ار بی ان بی) متوسل می شوند.

شهر نیویورک با بیش از ۵۲۰۰۰ لیست از نوامبر ۲۰۱۸ یکی از گرمترین بازارهای (ار بی ان بی) بوده است. این بدان معنی است که بیش از ۴۰ خانه در هر کیلومتر مربع اجاره داده شده است. در

NYC در Airbnb! شاید بتوان موفقیت (ار بی ان بی)

در نیویورک را ناشی از نرخ بالای هتل ها دانست که عمدتا ناشی از قیمت های گزاف اجاره در شهر است.

استفاده از سیستم (ار بی ان بی) مزایای بسیاری دارد؛ از قبیل :

انتخاب گسترده:

میزبانان انواع مختلفی از املاک – اتاق های یک نفره ، مجموعه ای از اتاق ها ، آپارتمان ها ، قایق های تفریحی، قایق های خانه ای ، کل خانه ها ، حتی یک قلعه را در (ار بی ان بی) ذکر کرده اند.

لیست های رایگان:

میزبان برای ذکر مشخصات خود نیازی به پرداخت هزینه ندارد. لیست ها می توانند شامل توضیحات نوشتاری ، عکس های زیرنویس شده و نمایه کاربری باشند که میهمانان بالقوه می توانند کمی درباره میزبان ها بدانند.

میزبان ها می توانند قیمت خود را تعیین کنند:

این که هر شارژ در هر شب ، هفته یا هر ماه چه هزینه ای دارد ، به عهده هر میزبان است.

جستجوی قابل برنامه ریزی:

مهمانان می توانند پایگاه داده (ار بی ان بی) را جستجو کنند – نه تنها براساس تاریخ و مکان ، بلکه براساس قیمت ، نوع ملک ، امکانات و زبان میزبان.همچنین می توانند کلمات کلیدی ("نزدیک به لوور") را اضافه کنند تا جستجوی خود را بیشتر محدود کنند.

خدمات اضافی:

در سال های اخیر (ار بی ان بی) پیشنهادات خود را گسترش داده و شامل تجربیات و رستوران ها است. علاوه بر لیستی از مکانهای اقامتی موجود برای تاریخی که قصد سفر دارند ، افرادی که براساس مکان جستجو می کنند لیستی از تجربیات مانند کلاسها و گشت و گذار را که توسط میزبانان محلی (ار بی ان بی) ارائه می شود ، مشاهده می کنند.

محافظت از مهمانان و میزبانان:

به عنوان محافظتی برای میهمانان ، (ار بی ان بی) قبل از انتشار وجه به میزبان ، هزینه مهمان را به مدت ۲۴ ساعت پس از ورود به سیستم نگه می دارد.

هدف و فرض

هدف از مدل های پیش بینی روشن است , ویژگی های لیست X را

به مدل وارد می کنیم. ، قیمت لیست Y

را تولید می کنیم. با این حال ، ما نمی خواهیم همه قیمت های لیست

را پیش بینی کنیم. ما فقط می خواهیم قیمتY را که برای مهمان منطقی و قابل قبول است پیش بینی کنیم ، بنابراین مدل ما می تواند پیشنهاد قیمت مطمئن را به میزبان ارائه دهد.از آنجا که ما سوابق معاملات لیست Airbnb را نداریم. ما فرض می کنیم که همه لیست ها معاملات موفقی نداشته اند ، یعنی لیست های جدیدی که مشتری نداشته اند. بیشتر فرض می کنیم لیست هایی که توسط میهمانان مرتباً مرور می شوند منطقی تر و مورد قبول میهمان هستند. بنابراین ، ما لیستی را فیلتر کردیم که هیچ نمره بررسی ندارد. این لیست ها ممکن است خصوصیات نامعقولی داشته باشند که مانع از

اجاره مهمانان می شوند ، از جمله قیمت بسیار بالا یا عدم امنیت.

شرح داده ها

این مجموعه داده شامل سه جدول اصلی است:

listings– داده های لیست. برخی از ویژگی های مورد استفاده در تجزیه و تحلیل عبارتند از قیمت (پیوسته) ، طول جغرافیایی (پیوسته) ، عرض جغرافیایی (پیوسته) ، نوع لیست (دسته بندی) ، سوپرهاست (دسته بندی) ، محله (دسته بندی) ، رتبه بندی (پیوسته) از جمله سایر ویژگی ها.

بررسی ها – ویژگی های کلیدی شامل تاریخ (زمان داده) ، لیست_گذاری (گسسته) ، بررسیگر (گسسته) و نظر (متنی) است.

تقویم – جزئیات مربوط به رزرو برای سال آینده را با ذکر لیست ارائه می دهد. در کل چهار ویژگی شامل list_id (گسسته) ، تاریخ (زمان قرارگیری) ، موجود (طبقه ای) و قیمت (مداوم).

ویژگی های دیگر

از جمله این ویژگی ها می توان به لیست قیمت ، نوع اتاق ، امکانات رفاهی و موقعیت مکانی و غیره اشاره کرد. سایر ویژگی ها شامل اطلاعات میزبان ، چیدمان اتاق ، امکانات رفاهی ارائه شده ، خط مشی ، قیمت لیست و امتیازات بررسی می باشد.

بسیاری از ویژگی ها غیر عددی هستند ، یعنی بولین، دسته بندی و متنی هستند. در حالی که اکثر ویژگی ها را می توان به مقادیر عددی تبدیل کرد ، سایر موارد به صورت بولین ,دسته بندی و متنی باقی مانده و با ابزارهای رگرسیون مناسب درمان می شوند.

امکانات

این لیست ها شامل ویژگی های عددی و غیر عددی است. به عنوان مثال ، امکانات یک لیست ، به عنوان مثال تلویزیون و وای فای در یک آرایه وجود دارد. از آنجا که بسیاری از امکانات در بسیاری از خواص مشترک است ، ما ابتدا همه امکانات را در مجموعه ای از ۳۵ امکانات بی نظیر در همه لیست ها خلاصه کردیم. سپس هر لیست را با ۳۵ ویژگی گسترش می دهیم ، که هر یک نشان دهنده یک راحتی منحصر به فرد است که به صفر یا نادرست رسیده است. بعد ، ما از طریق هر لیست ادغام می شویم و مجموعه امکانات آن را بررسی می کنیم.

در ۳۵ ستون اضافی ، اگر همان ویژگی در آرایه یافت شود ، ویژگی مربوطه را روی یک یا true قرار می دهیم. ما این توالی را از طریق لیست های اضافی تکرار می کنیم و هر ورودی را به یک تنظیم می کنیم تا یک راحتی را نشان دهد. در نتیجه ، تمام امکانات موجود در مجموعه ما با اضافه شدن ۳۵ ستون جدید رمزگذاری می شوند. ما اکنون ستون اصلی آرایه ها را حذف می کنیم ، و ستون های رمزگذاری شده را می گذاریم تا وجود یک امکانات در لیست را نشان دهیم.

انواع تختخواب

فقط ۵ نوع تختخواب برای هر لیست وجود دارد که شامل Airbed ، Couch، Futon ؛ مبل Pull-out و Real Bed می باشد.

ما در اینجا از همان روش رمزگذاری استفاده می کنیم ، جایی که هر نوع با ۱ نشان داده می شود. در نتیجه ، ما پنج ستون جدید از ویژگی های رمزگذاری شده اضافه کردیم. سرانجام ، ما ویژگی های بسیاری با مقادیر بولین داریم که ویژگی های لیست و میزبان را نشان می دهد. به عنوان مثال ،آیا این یک لیست ورود سریع یا با رزرو سریع ۲۴ ساعته است.

یک نگاه سریع به داده ها نشان می دهد که موارد زیر وجود دارد:

در کل ۵۰،۹۶۸ لیست منحصر به فرد در NYC, اولین اجاره در نیویورک در آوریل ۲۰۰۸ در هارلم ، منهتن انجام شد.

از آن زمان تاکنون بیش از ۱ میلیون بار توسط مهمانان نوشته شده است.

قیمت یک لیست از ۱۰ دلار در هر شب تا ۱۰،۰۰۰ دلار متغیر است.

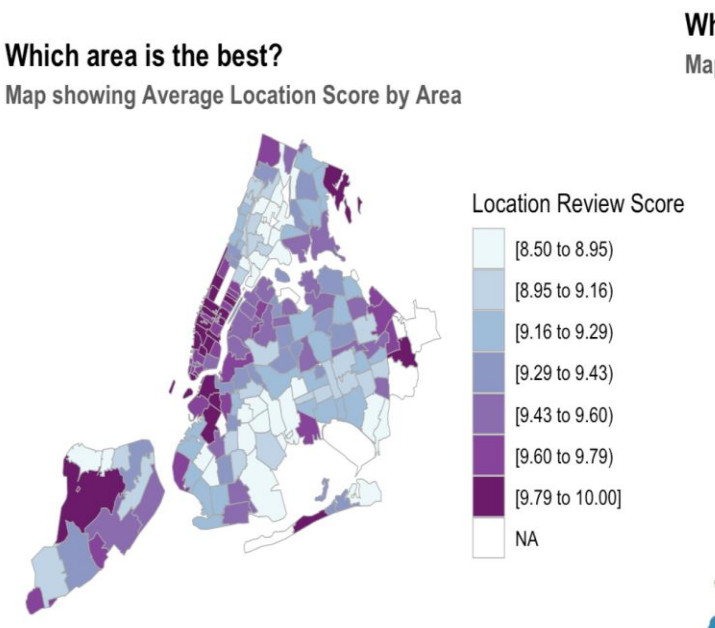
این لیست با قیمت ۱۰،۰۰۰ دلار در Greenpoint ، بروکلین قرار دارد. آستوریا ، کوئینز و بالا وست ساید ، منهتن.

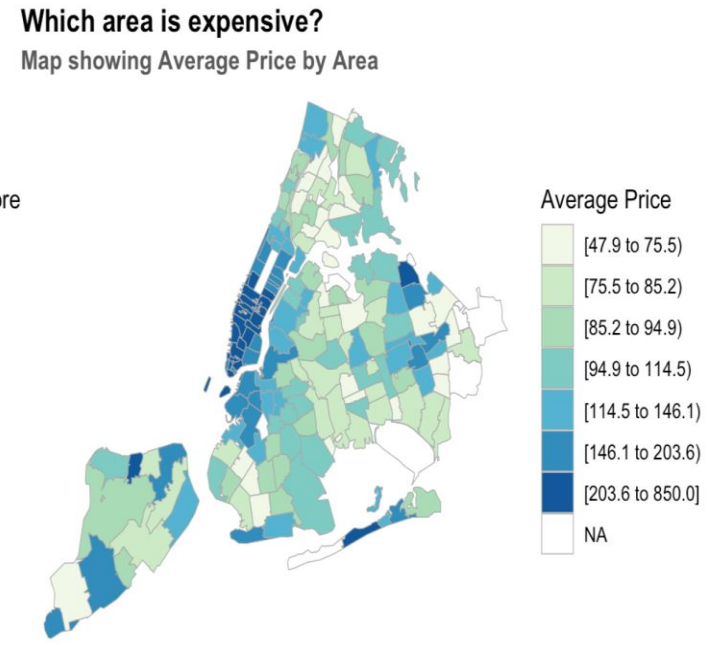
محبوبیت AIRBNB در نیویورک چگونه است؟

مشابه تعداد میزبان ها ، تعداد لیست های منحصر به فرد دریافت کننده به طور پیوسته در طول سالها افزایش یافته است ، که نشان دهنده افزایش تصاعدی تقاضا برای اجاره Airbnb است.

درک صحنه املاک NYC

کاربران Airbnb اقامت خود را بر اساس مکان ، تمیزی و پارامترهای دیگر ارزیابی می کنند. در اینجا با داده های نمره مکان کار می کنیم. دیدن میانگین امتیازات برای هر محله جالب خواهد بود. نمرات مکان باید نشان دهنده جذابیت محله باشد. محله هایی که دارای امتیاز بالایی هستند ، از اتصال بهتری (ایستگاه های مترو) برخوردار خواهند بود و به نقاط داغ شهر نزدیکتر هستند (تایمز اسکوئر ، امپایر استیت ، وال





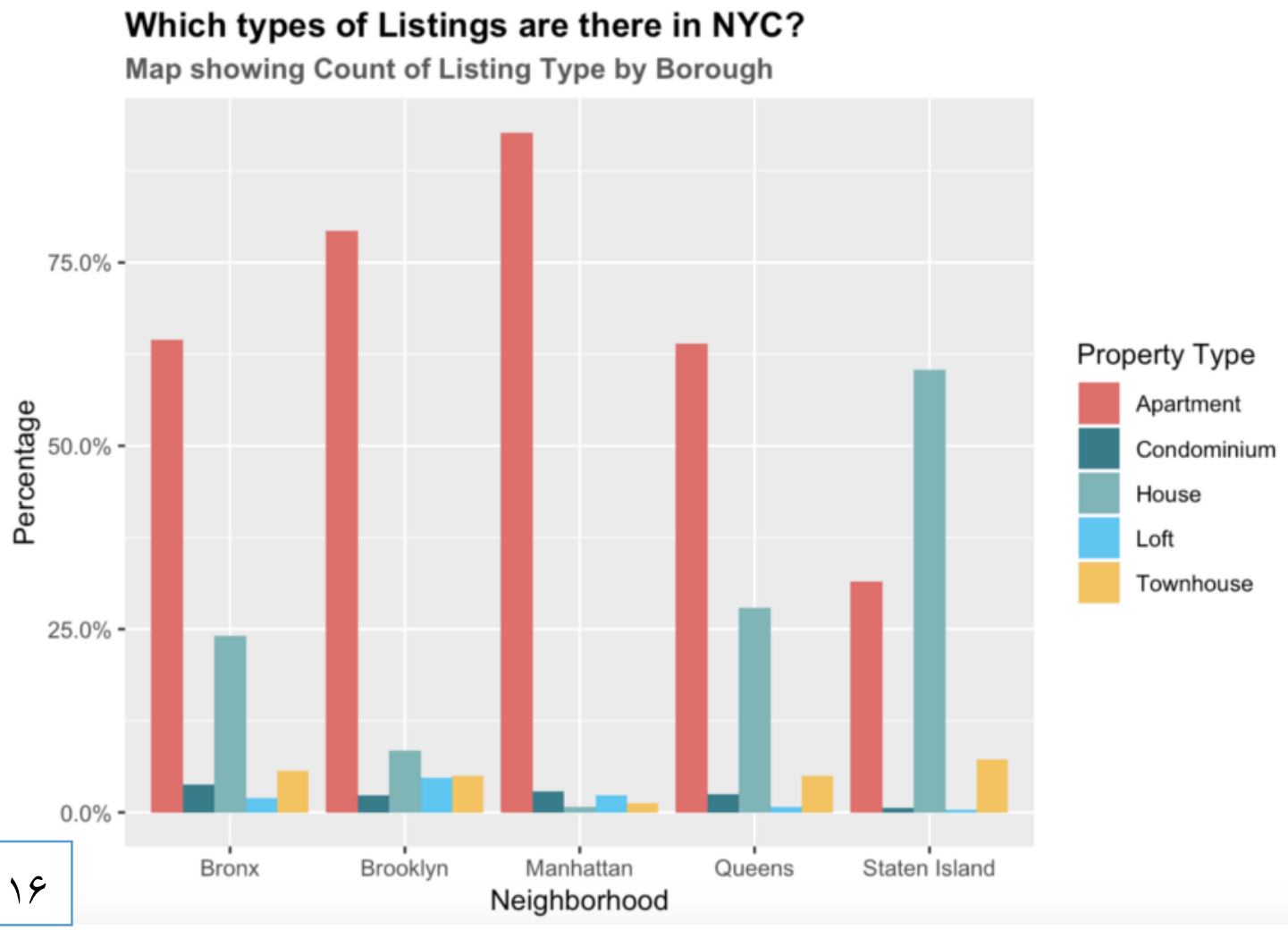
منهتن بالاترین امتیازات موقعیت مکانی را برای منطقه مرکز شهر (esp زیر پارک مرکزی)دریافت می کند. در جزیره استاتن ، مناطق نزدیک به پارک ایالتی بالاترین امتیاز مکان را دارند. محله های بروکلین نزدیک به منهتن دارای رتبه بندی مکان بالاتری هستند. با نگاهی به سیستم مترو NY در بروکلین ، جالب است که مشاهده کنیم مناطق دارای درجه بالایی با حضور خط مترو مطابقت دارند. در مورد برانکس که خطوط مترو رفت و آمد نمی کنند نیز همین مسئله است.

هزینه های لیست تا حد زیادی با نمرات مکان مطابقت دارد. مکانهای دارای امتیاز بالا نیز گرانترین مکانها هستند. بدیهی است که مکان دارای امتیاز بالا نیز گران خواهد بود (تقاضا در مقابل عرضه)

با این حال جالب است که چند نکته دور از ذهن را تشخیص دهیم: ۱)یافتن رتبه بندی بالا – مناطق کم اجاره (بهترین هر دو جهان): منطقه ایالت پارک در جزیره استاتن (که در نمودار قبلی به آن پرداخته شده است) یکی از چنین مناطقی است که با وجود داشتن بالاترین رتبه بندی مکان ، اجاره بها نسبتاً کم است. یکی دیگر از نقاط خوب واقع در شمال شرقی بروکلین است.

۱) یافتن رتبه پایین – مناطق با اجاره بالا (بدترین وضعیت هر دو جهان): منطقه Elm Parkدر استاتن آیلند دارای اجاره های نامتناسب بالا و در عین حال امتیازات بسیار کمی است. چنین مکانهای دیگری را می توان در مناطق شمال برانکس یافت.

اکنون بیایید انواع لیست هایی را که در NYC وجود دارد کشف کنیم. در زیر نموداری آورده شده است که توزیع انواع مختلف لیست توسط بخش های مختلف را نشان می دهد.



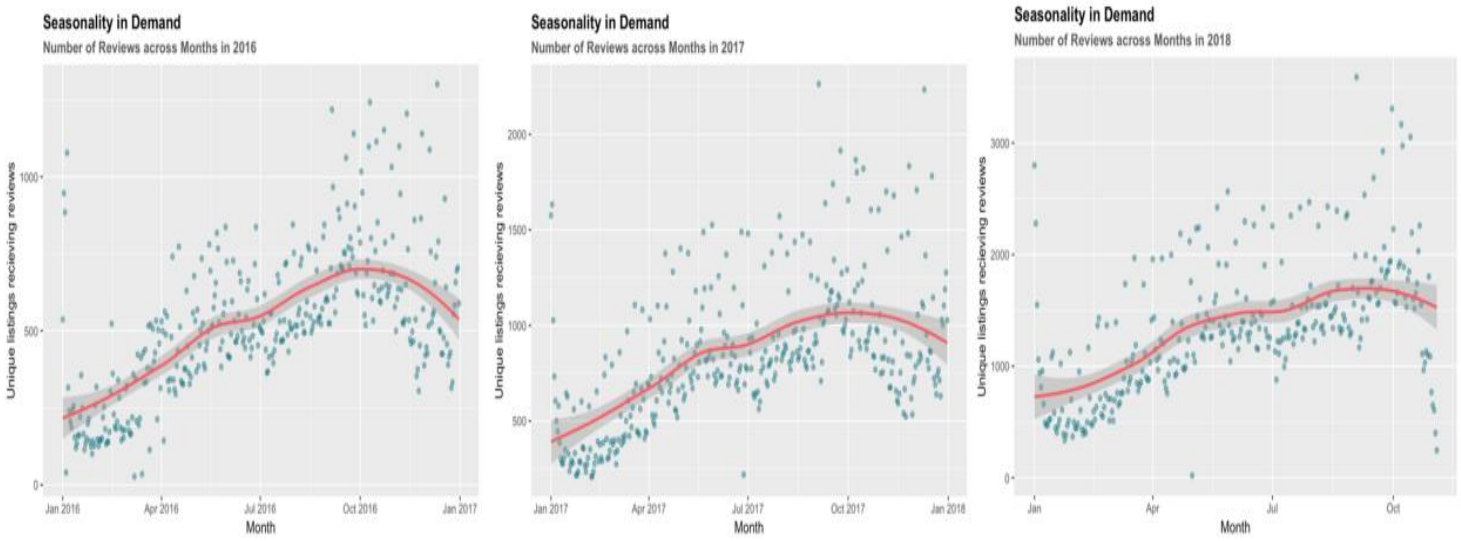
لیست های سبک آپارتمان در هر چهار محله به جز جزیره استاتن بالاترین تعداد را دارد. جزیره استاتن دارای ویژگی های "House" بیشتر از "آپارتمان ها" است. این به نظر شهودی می رسد ، زیرا جزیره استاتن دارای جمعیت کم است ، بنابراین "فضای" بیشتری در مقایسه با سایر بخش ها دارد.

تجزیه و تحلیل تقاضا و قیمت گذاری:

در این بخش ، تجزیه و تحلیل تقاضا و قیمت را برای اجاره های Airbnb انجام می دهیم. برای درک فصلی بودن تقاضا را در طول سالهای تأسیس Airbnb در سال ۲۰۰۸ و ماه های سال بررسی خواهیم کرد.

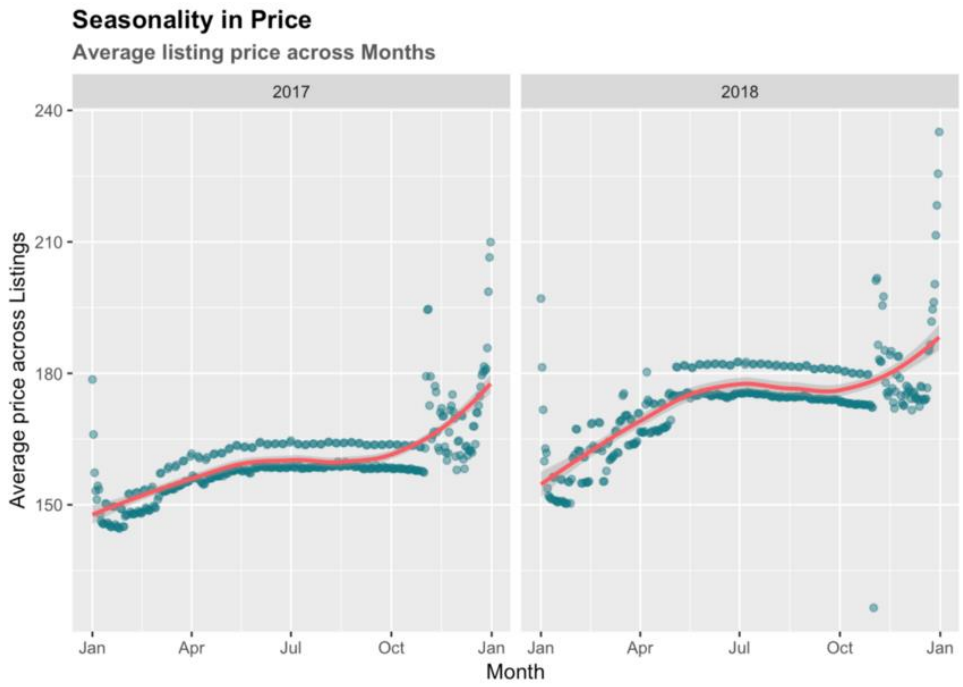
همانطور که قبلاً ذکر شد ، به دلیل در دسترس نبودن اطلاعات رزرو ، از تعداد بازبینی ها به عنوان پروکسی تقاضای اجاره Airbnb استفاده خواهیم کرد. فرض این است که تعداد بررسی ها با تقاضای اجاره مطابق ادعای Airbnb مبنی بر بررسی ۵۰٪ از مهمانان برای اقامت خود مطابقت دارد. علاوه بر این ، مهمانان باید ظرف ۲ هفته از اقامت خود ، یک بررسی ارائه دهند ، از این رو تعداد بررسی ها می تواند تقاضای خوبی را برای یک دوره خاص ارائه دهد.

از طریق نمودار "Airbnb چقدر محبوب است؟" که قبلا نشان داده شده است (برای سهولت مراجعه مجدداً آن را در زیر آورده ایم) ، می توان الگوی فصلی را در تعداد بررسی / تقاضا مشاهده کرد. هر ساله اوج و افت تقاضا وجود دارد که نشان می دهد ماه های خاصی در مقایسه با ماه های دیگر شلوغ تر است.



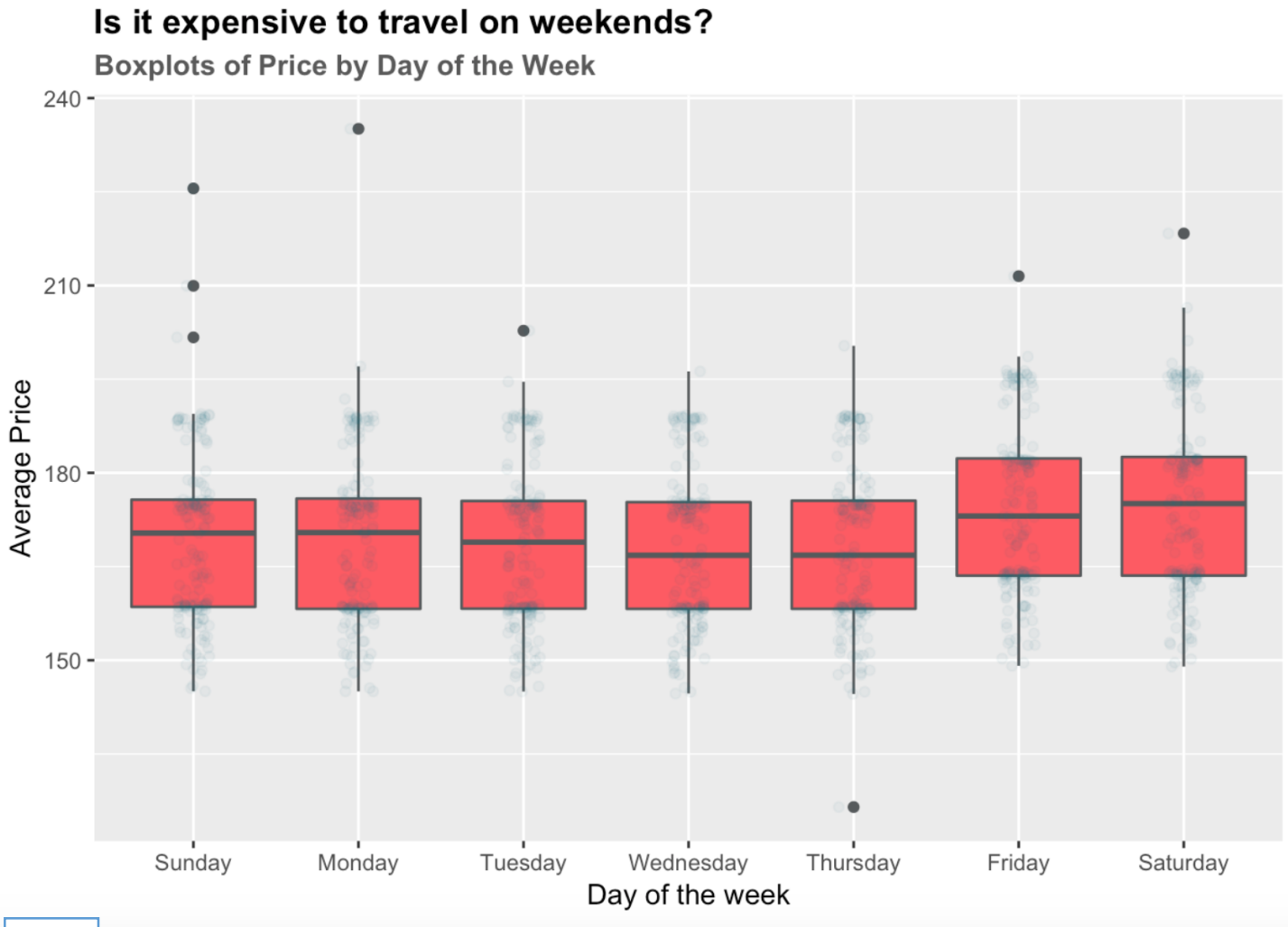
بررسی این موضوع در سطح نقطه ای نشان می دهد که تقاضا در ژانویه کمترین است و تا اکتبر افزایش می یابد ، تا زمانی که شروع به کاهش می کند تا پایان سال.

آیا فصلی در قیمت اجاره وجود دارد؟ بیایید نگاهی به میانگین قیمت روزانه لیست ها در طول سال ها داشته باشیم.



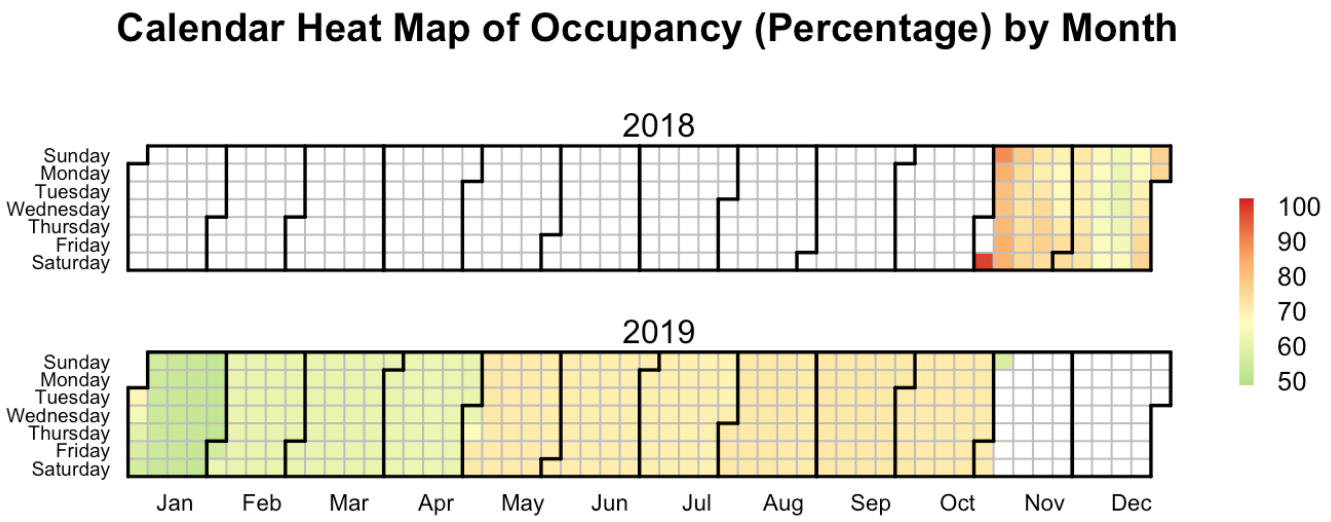
با پیشرفت در طول سال و اوج گیری در دسامبر ، میانگین قیمت ها در بین لیست ها افزایش می یابد. این الگوی مشابه تعداد بررسی ها / تقاضا است به جز در ماه های نوامبر و دسامبر ، که در آن تعداد بررسی ها (نشانگر تقاضا) شروع به کاهش می کند.

همچنین می توانیم دو نمودار از نمودارها را مشاهده کنیم که نشان می دهد قیمت های متوسط در روزهای خاص در مقایسه با روزهای دیگر بالاتر بود. در ادامه ، برای درک این پدیده ، یک جعبه قیمت متوسط قیمت در روز هفته ترسیم می کنم.



همانطور که می بینیم ، جمعه و شنبه ها در مقایسه با سایر روزهای هفته گران ترند ، شاید به دلیل تقاضای بیشتر برای اسکان.

با مطالعه چگونگی وضعیت اشغال برای سال آینده ، با استفاده از داده های تقویم جدول ، درصد اشغال سال بعد را می فهمیم.



می توان استنباط کرد که ژانویه ساکت ترین ماه است و با پیشرفت در سال میزان اشغال بیشتر می شود. این امر با نتایج حاصل از تجزیه و تحلیل تعداد بررسی ها (نشانگر تقاضا) که روند افزایشی در طول سال را نشان می دهد ، پیوند می خورد.

تجزیه و تحلیل نظرات مشتری

این مجموعه داده تعداد زیادی داده را در اختیار ما قرار می دهد ، اما هیچ چیز به اندازه بررسی / بازخورد مشتری نزدیک به مشتری نیست. اگر به درستی استخراج شود ، آنها می توانند چیزهای زیادی در مورد طرز فکر مشتری ، انتظاراتشان و میزان برآورده شدن آنها به ما بگویند. برای منطقی بودن نتیجه نهایی ، داده های متن مرور نیاز به تمیز کردن زیادی دارند – به عنوان مثا, کاما ,درصد و غیره باید حذف شوند .تجزیه و تحلیل ابر کلمه روندهای جالبی را نشان می دهد. به نظر می رسد مکان مهم باشد ، زیرا کلمات "همسایگی" ، "مکان" ، "منطقه" به طور برجسته در ابر کلمه وجود دارد. گزینه های حمل و نقل مانند "مترو" ، "پیاده روی" نیز مورد اشاره مکرر است. کلماتی مانند "آشپزخانه" به ما می گویند که بسیاری از افراد ترجیح می دهند آشپزی کنند تا بیرون غذا بخورند. در دسترس بودن "رستوران ها" با ذکر نام نیز نزدیک است. محل استحمام و تختخواب ، همانطور که انتظار می رود ، اگر در شرایط مطلوب نباشد ، می تواند معامله کننده روشنی باشد. کلمه "میزبان" بسیار مورد توجه است. نشان دهنده نقش مهمی است که میزبان در شکل گیری تجربه Airbnb بازی می کند.

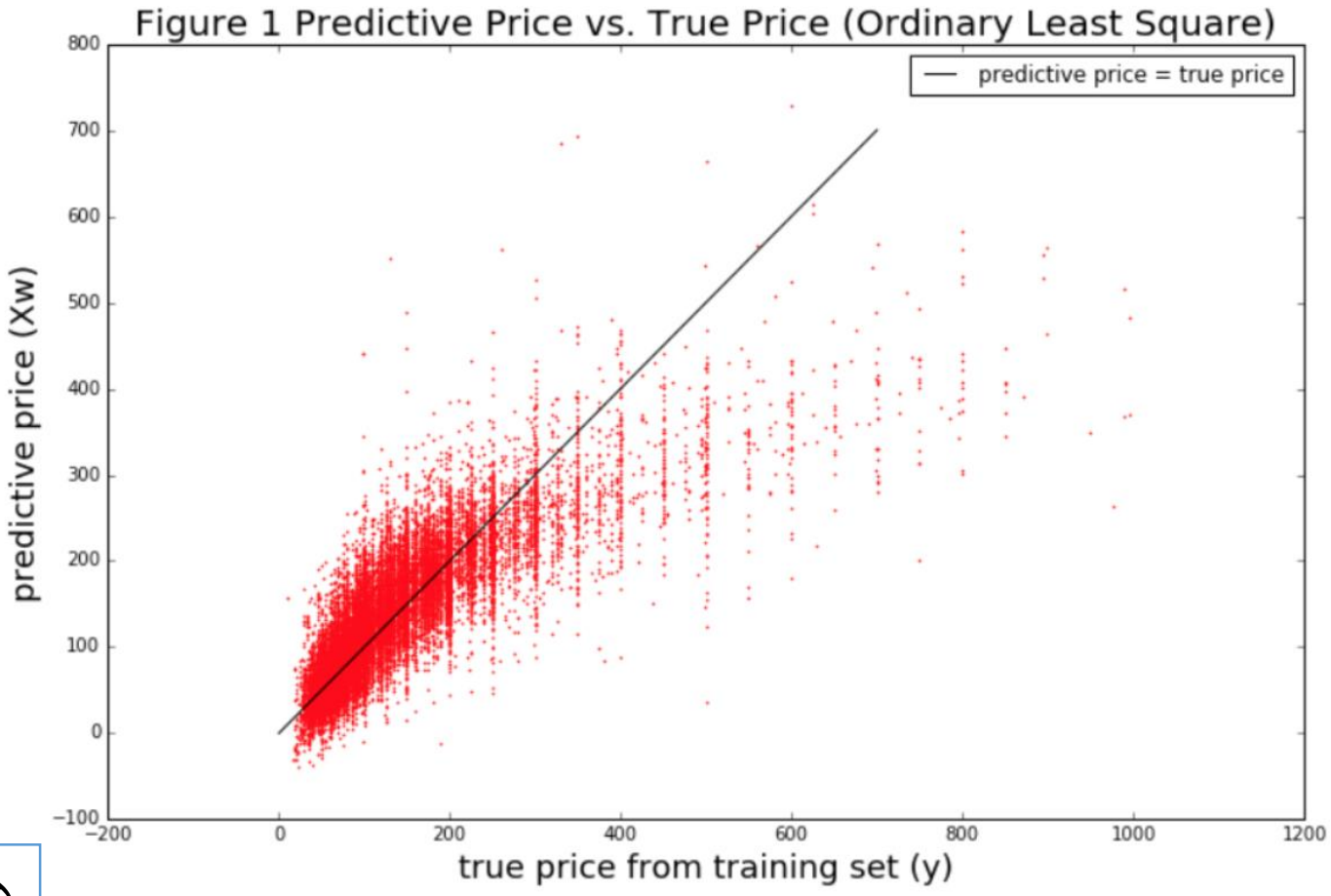
بردارهای کلمه راهی موثر برای کشف نزدیکترین کلمات به عبارات جستجوی خاص فراهم می کنند. با استفاده از داده های بررسی ، یک فضای برداری برای ساختن ابر کلمه ای از کلمات مشابه برای استخراج ایجاد کردیم.

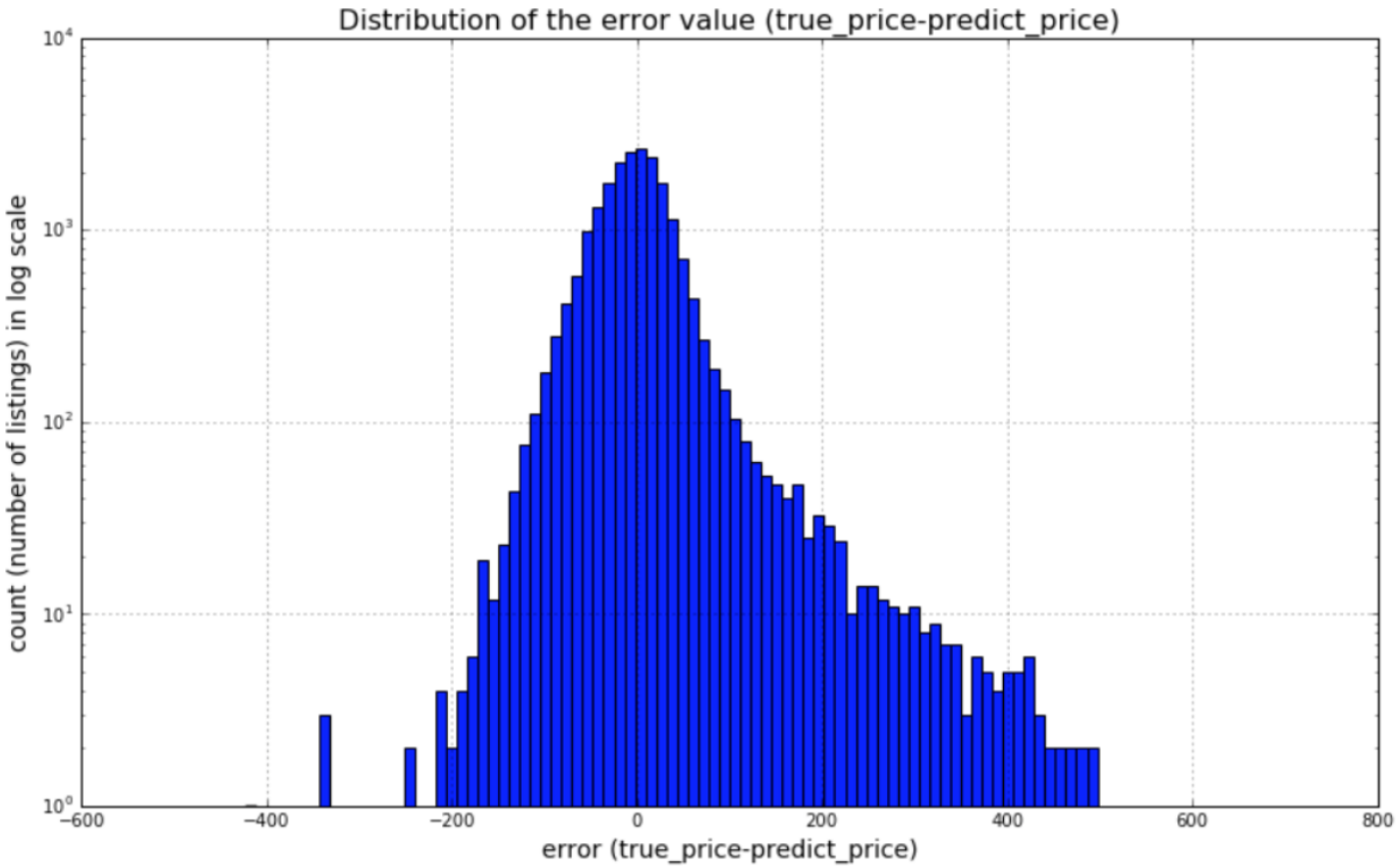
اولین ابر کلمه برای کلمه "ناراحت کننده" است. کلمات مشابه "ناراحت کننده" معمولاً آنهایی هستند که همراه با آن مرتباً اتفاق می افتند ، یعنی دلایل ناراحتی. ابر کلمه این را نشان می دهد – به کلماتی مانند "تنگ" ، "شلوغ" ، "کوچک" ، "گرفتگی" و "به هم ریخته" توجه کنید که نشان می دهد کمبود فضا یکی از رایج ترین شکایات است. "گرم" ، "مرطوب" و "سرد" از موارد معمول دما هستند. محیط "گرد و خاک" ، "کثیف" مردم را به نوشتن بازخورد منفی سوق می دهد. بسیاری احساس "عصبی" ، "ناامن" و "استرس" می کنند. به وضوح یک پرچم قرمز برای مستاجران آینده است.

به همین ترتیب ، با پرسش از کلمه کلیدی "راحت" ، امیدواریم مواردی را که منجر به یک تجربه مثبت شده است ، ببینیم. واژه هایی مانند "ساکت" ، "قابل پیاده روی" ، "تمیز" ، "بدون لک" و غیره به طور برجسته , برجسته می شوند ، که باز هم اهمیت محیط ، مکان و تمیزی را نشان می دهد. "میزبانان" و "ارتباطات" مفید به راحتی منجر می شوند. تمیز بودن و اندازه تختخواب ها تأثیر تعیین کننده ای دارند.

مدل اول: رگرسیون خطی

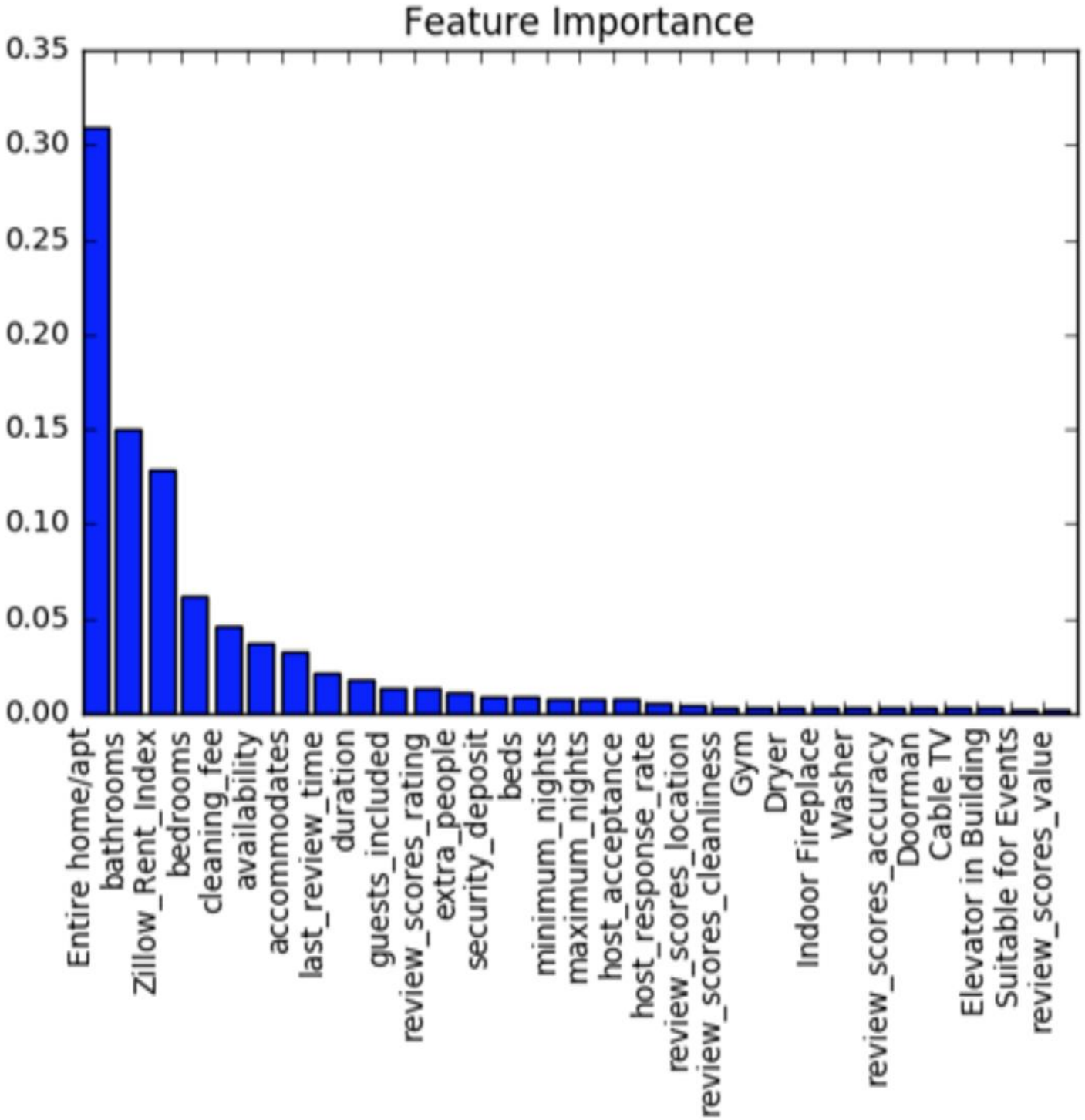
برای شروع ، ما یک مدل خطی با استفاده از رگرسیون حداقل مربع معمولی روی مجموعه داده های لیست X برای پیش بینی قیمت لیست Y ایجاد می کنیم. متاسفانه مدل های خطی عملکرد خوبی ندارند. شکل ۱ نشان می دهد که وقتی قیمت واقعی بالا باشد ، مدل قیمت را پیش بینی نمی کند. می توانیم ببینیم که توزیع خطا در شکل ۲ به جهت مثبت متمایل شده است ، این بدان معنی است که پیش بینی های بیشتری با خطای مثبت نسبت به خطای منفی داریم.





یافتن ویژگی های مهم: جنگل تصادفی

ما می خواهیم از مهمترین ویژگی ها در کمک به تفاوت قیمت بین خانه ها درک بصری بگیریم. جنگل تصادفی ابزاری بسیار ضروری است که به ما کمک می کند تا اهمیت این ویژگی ها را بررسی کنیم.


Feature Importance
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Entire home/apt
bathrooms
Zillow_Rent_Index
bedrooms
cleaning_fee
availability
accommodates
last_review_time
duration
guests_included
review_scores_rating
extra_people
security_deposit
beds
minimum_nights
maximum_nights
host_acceptance
host_response_rate
review_scores_location
review_scores_cleanliness
Gym
Dryer
Indoor Fireplace
Washer
review_scores_accuracy
Doorman
Cable TV
Elevator in Building
Suitable for Events
review_scores_value

ما از اهمیت تولید شده توسط جنگل های تصادفی به عنوان وزن هر ویژگی استفاده می کنیم که به قیمت گذاری خانه (ار بی ان بی) کمک می کند.

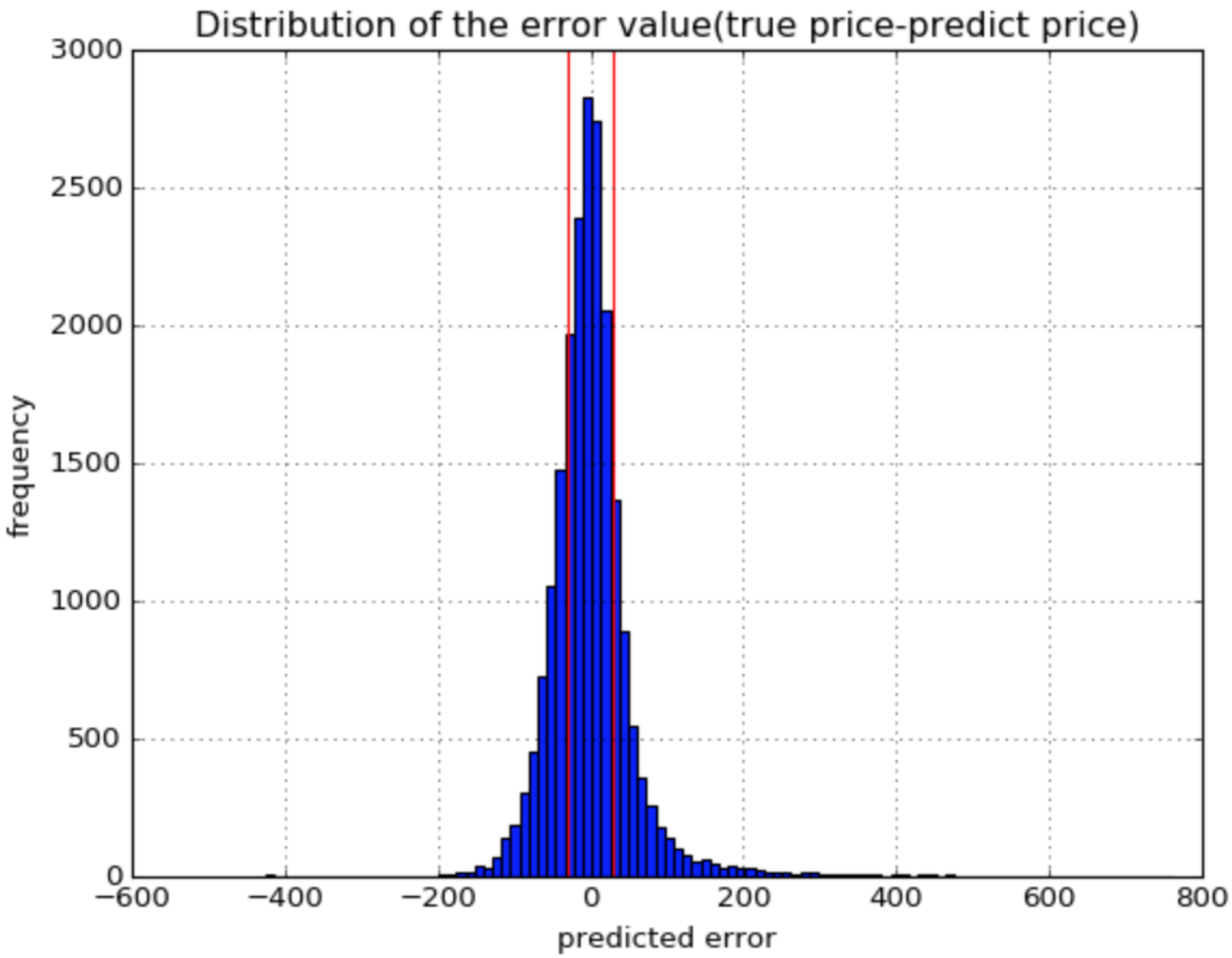
بهینه سازی مدل

پس از گرفتن نتیجه جنگل تصادفی ، بعد ویژگی ها را کاهش دادیم. ۲۰ عامل مهم در نتیجه در مدل رگرسیون خطی گنجانده خواهد شد. کل مجموعه داده نیز به طور تصادفی به مجموعه آموزش و مجموعه آزمون تقسیم شد که هر کدام به ترتیب ۸۰٪ و ۲۰٪ ورودی ها را شامل می شد.

با استفاده از تابع از دست دادن درجه دوم با یک تنظیم کننده هنجار ، مدل خطی ما می تواند وزن ویژگی های مختلف را تحت تأثیر, کمتر تحت فشار قرار دهد.Scikit-learn این عملکرد را برای ما فراهم می کند. برای بدست آوردن بهترین لامبدا مناسب ، اعتبارسنجی k را برابر با k=5 اجرا کردیم.

طبقه بندی داده های پرت

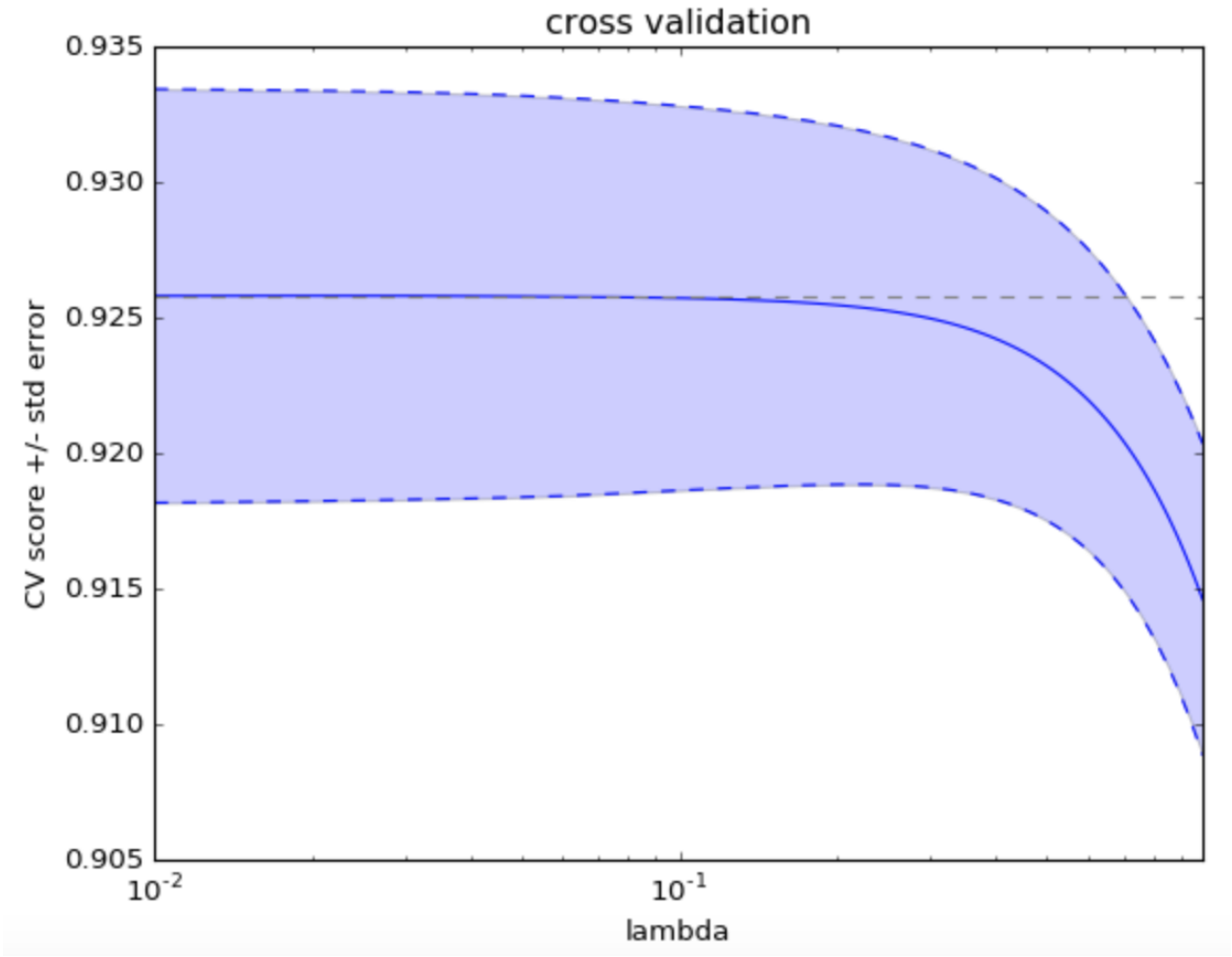
برای غلبه بر مشکل عدم کفایت که در بالا ذکر شد ، تصمیم گرفتیم طبقه بندی کنیم تا نشان دهد آیا مدل ما قیمت را به خوبی پیش بینی می کند یا خیر. مدل ما تمایل دارد لیست ها را با قیمت های بالاتر پیش بینی کند و لیست ها را با قیمت های پایین بیش از حد پیش بینی کند. پیش بینی آن لیست ها دشوار است زیرا ممکن است دارای برخی ویژگی های منحصر به فرد باشند که ما از مجموعه داده های خود استخراج نمی کنیم ، مانند شرح متن یا تصویر. بنابراین ، ما می خواهیم آن لیست ها را با خطای بزرگ منفی یا مثبت نادیده بگیریم ، و آنها را به عنوان نقاط قیمتی در نظر می گیریم.ما نقاط داده را براساس خطای آنها بین قیمت های پیش بینی شده و واقعی جدا می کنیم. نمودار زیر هیستوگرام خطاها را نشان می دهد که به دو قسمت تقسیم شده است. نقاط داخل منطقه محدود شده توسط دو خط در ۳۰ دلار از قیمت واقعی قرار دارند و در گروه "۱+" قرار می گیرند. لیست های خارج از منطقه با برچسب "-۱" ، از آنجا که خطاهای آنها بیش از ۳۰ دلار است.



سپس با استفاده از رگرسیون لاجستیک ، یک مدل طبقه بندی ایجاد می کنیم تا لیست ها را به عنوان داده های معتبر یا خارج از رده بندی طبقه بندی کنیم. با استفاده از این مدل طبقه بندی ، ما فقط مدلهای خطی را در لیست هایی که به عنوان داده های معتبر طبقه بندی می شوند ، اعمال خواهیم کرد.

مدل دوم: رگرسیون خطی روی داده های معتبر

سپس ، ما رگرسیون خطی را روی داده های معتبر (لیست های موجود در منطقه محدود) مجدداً انجام می دهیم ، این بدان معنی است که مدل نهایی ما فقط بر روی لیست های عادی تمرکز دارد. با استفاده از رگرسیون خطی و به دنبال آن اعتبار صلیبی k برابر ، ضریب هر ویژگی را بدست آوردیم و بهترین لامبدا را پیدا کردیم. شکل سمت چپ نمره واریانس مدلهای مختلف با لامبدا را نشان می دهد. منطقه آبی به معنی ۹۵٪ فاصله اطمینان از نمره واریانس است. نمره واریانس مقداری بین ۰ و ۱ است که وقتی پیش بینی دقیقاً با قیمت واقعی مطابقت داشته باشد ، ۱ است. شکل مناسب نشان می دهد که مدل رگرسیون خطی به روز شده منجر به پیش بینی خوب مجموعه آموزشی می شود.

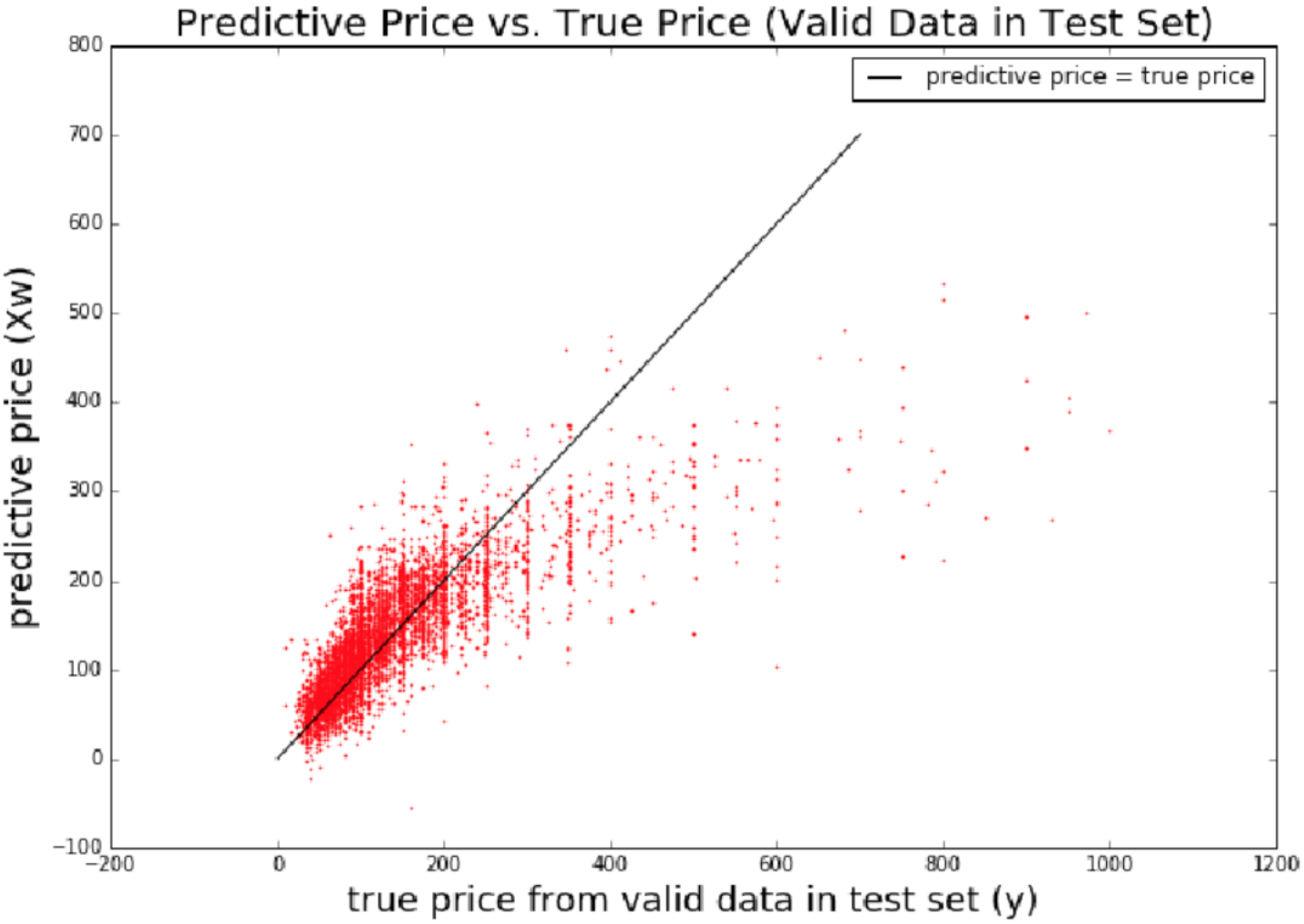

cross validation
0.935
0.930
0.925
0.920
0.915
0.910
0.905
CV score +/- std error
10^{-2}
10^{-1}
lambda

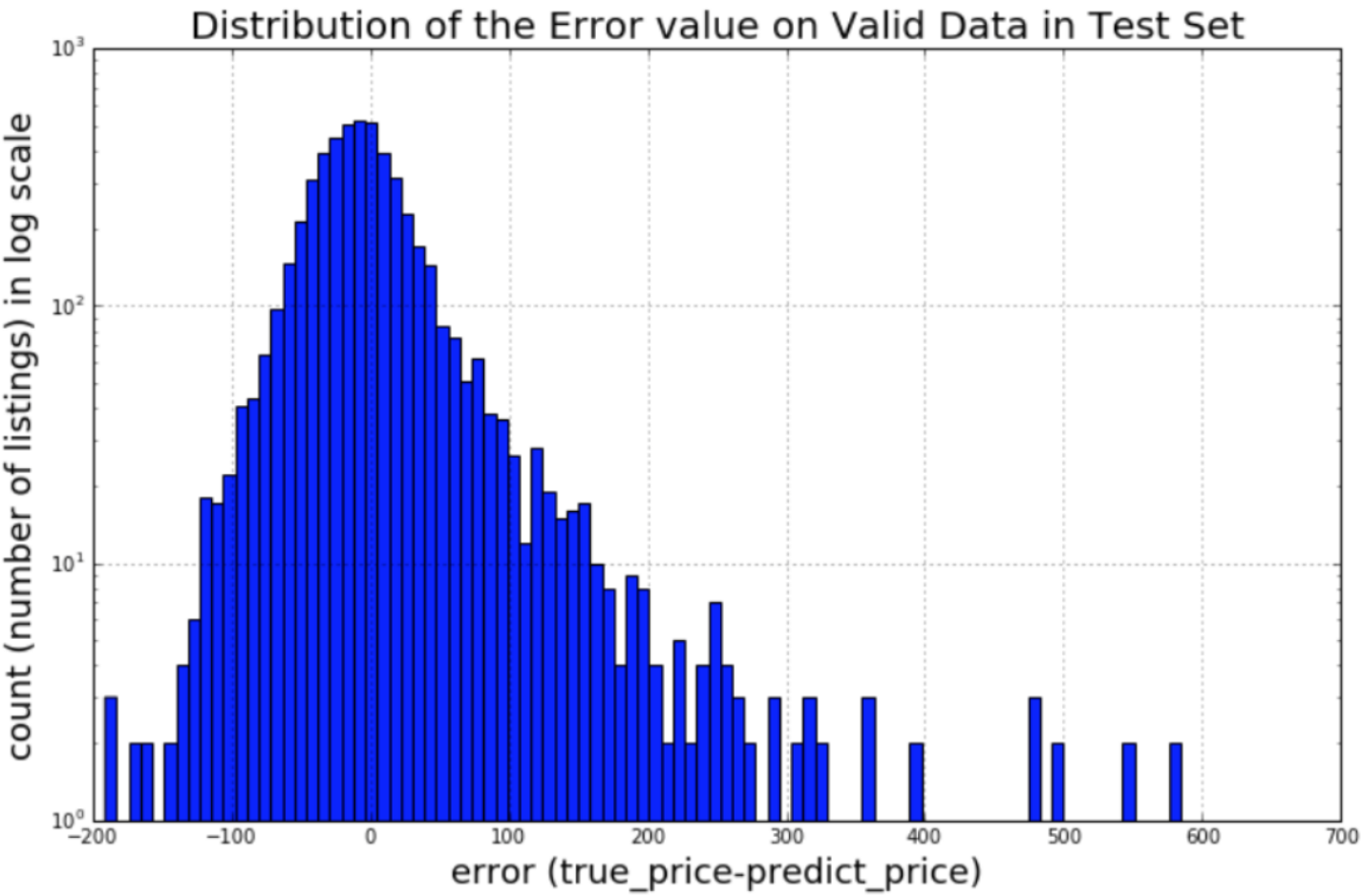

Predictive Price vs. True Price (Training Set)
predictive price = true price
800
700
600
500
400
300
200
100
0
−100
predictive price (Xw)
−100
0
100
200
300
400
500
600
700
800
true price from training set (y)

Feature Coefficient

Feature	Coefficient	Feature	Coefficient
Entire home/apt	50.3088	bathrooms	22.8411
cleaning_fee	0.3317	availability	17.4930
duration	-0.0004	guests_included	2.4008
security_deposit	0.0003	beds	-0.6477
host_acceptance	-2.1998	host_response	1.4236
Zillow_Rent_Index	0.0407	bedrooms	26.5680
accommodates	10.1101	last_review_time	0.0308
review_scores_rating	0.5396	extra_people	-0.0306
maximum_nights	0.0000	minimum_nights	-0.1793
review_scores_location	4.2416	verification_method	0.4665
interception	-221.9372		

بر اساس نتایجی که از مدل رگرسیون خطی به دست آوردیم ، ویژگی هایی از قبیل نوع اتاق ، تعداد حمام ها ، در دسترس بودن و تعداد مهمانان تأثیر مثبتی بر قیمت لیست دارد و سایر ویژگی ها مانند هزینه برای افراد اضافی و تعیین حداقل تعداد شب ها ، احتمال تعیین قیمت بالا را کاهش می دهد. سپس این مدل را روی مجموعه آزمایشات اعمال کنید. برای دیدن پیش بینی مجموعه آزمون ، ما ابتدا مجموعه آزمون را طبقه بندی می کنیم و نزدیک به ۹۰٪ داده های مجموعه آزمون معتبر قلمداد می شوند. حتی اگر ما فرآیند طبقه بندی ، مقدار دورترین قسمت را فیلتر کرده و در نتیجه محدوده پروژه خود را محدود کنیم ، ما همچنان پیش بینی قیمت لیست برای اکثر میزبان ها را داشتیم. سپس ، ما فقط مدل خطی را روی آن نقاط داده معتبر اعمال می کنیم. همانطور که در دو شکل زیر نشان داده شده است ، ما هنوز هم نمی توانیم مجموعه آزمون را به خوبی پیش بینی کنیم.


Predictive Price vs. True Price (Valid Data in Test Set)
predictive price = true price
predictive price (Xw)
true price from valid data in test set (y)
800
700
600
500
400
300
200
100
0
−100
−200
0
200
400
600
800
1000
1200

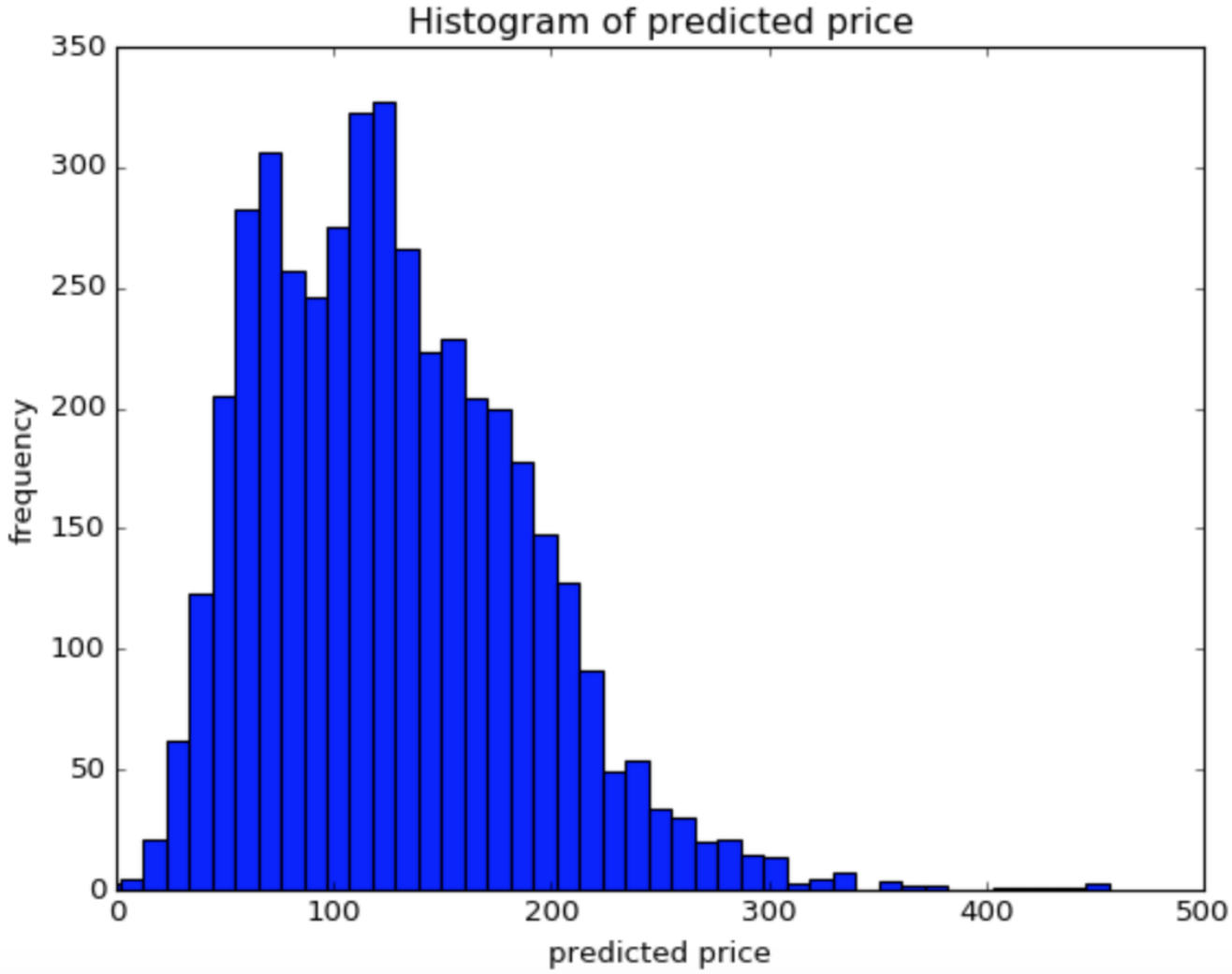

Distribution of the Error value on Valid Data in Test Set
count (number of listings) in log scale
error (true_price-predict_price)
10^3
10^2
10^1
10^0
−200
−100
0
100
200
300
400
500
600
700

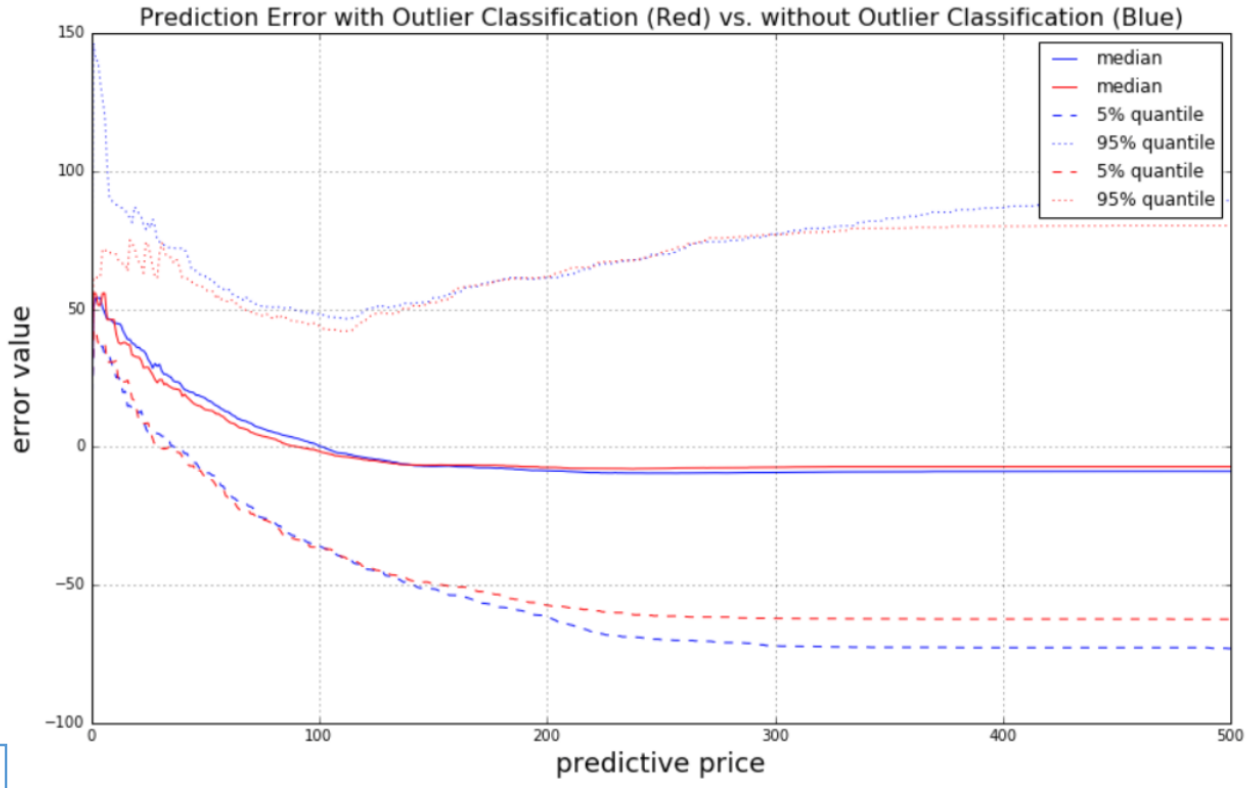
Confidenceمدل ما

پس از ساخت مدل های ما ، دانستن نحوه ارزیابی نتیجه یک مدل بسیار مهم است. روش زیر نحوه ارزیابی ما از میزان پیش بینی قیمت مدل از لیست Airbnb است. به این ترتیب ، ما قادر به مقایسه عملکرد بین مدل ها هستیم. از آنجا که هدف نهایی ما پیشنهاد قیمت به میزبان است ، یک مدل خوب باید خطای کوچک پیش بینی را ارائه دهد. از طرف دیگر ، یک خطای USD۳۰ برای توصیه قیمت USD۲۰۰ قابل توجه نیست ، اما ۳۰ USD ممکن است برای یک توصیه قیمت USD۸۰ یک خطای بسیار مهم باشد.

بنابراین ، ما تصمیم می گیریم ببینیم که قیمت پیش بینی شده با مقایسه خطاهای موجود در لیست ها با همان قیمت پیش بینی شده ، تا چه اندازه جدی از قیمت واقعی برخوردار است. همانطور که در شکل زیر نشان داده شده است ، به عنوان مثال ، در مورد لیست های با پیش بینی USD۲۰۰ ، خطای کمی ٪۵ – USD۷۰ است ، خطای کمی ٪۹۵ USD۵۰ است. در نتیجه ، با توجه به توصیه USD۲۰۰ قیمت ، ۹۰٪ اطمینان داریم که قیمت واقعی در بازه قیمتی

[۱۳۰ ، ۲۵۰] USD خواهد بود.


Histogram of predicted price
frequency
predicted price
0
50
100
150
200
250
300
350
0
100
200
300
400
500


Prediction Error with Outlier Classification (Red) vs. without Outlier Classification (Blue)
median
median
5% quantile
95% quantile
5% quantile
95% quantile
error value
predictive price
150
100
50
0
−50
−100
0
100
200
300
400
500

شکل صفحه ی قبل نشان می دهد که متداول ترین قیمت های لیست Airbnb در شهر نیویورک بین ۵۰ تا ۲۰۰ دلار است. شکل مناسب در بالا نشان می دهد که در این محدوده ، میانگین خطای مدل نزدیک به –۵ دلار بود. در بدترین حالت ، مقدار مطلق خطا می تواند ۴۵ دلار باشد که کمتر از ۴۰٪ قیمت پیش بینی شده بود. به طور خلاصه ، ما ۹۰٪ اطمینان داریم که خطا کمتر از ۴۰٪ قیمت پیش بینی کننده خواهد بود.

مدل سوم: الگوریتم نزدیکترین همسایگی ویژگی

روش دیگر برای پیش بینی قیمت یک لیست این است که لیست های مشابه را از مجموعه داده های نمونه انتخاب کنید ، مدل پیش بینی را فقط بر اساس لیست های مشابه به جای کل مجموعه داده های آموزش بسازید. بنابراین ، ما روشی را برای مقایسه ویژگی های یک لیست با لیست های دیگر در مجموعه داده توسعه می دهیم. قبل از محاسبه تشابه بین دو لیست ، باید مقادیر وزن را به هر ویژگی اختصاص دهیم.

وزن دهی صفت اهمیت هر صفت را با توجه به خروجی ارزیابی می کند ، ما وزن های مختلفی را به ویژگی های مختلف اختصاص می دهیم و بنابراین فاصله اقلیدسی هر یک از ویژگی ها را محاسبه می کنیم:

$$d\left(x, x^{(i)}\right) = \sum_{j=1}^{d} w_j \left(x_j - x_j^{(i)}\right)^2$$

d is the feature dimension, x_j is the j^{th}feature of x, $x_j^{(i)}$is the j^{th}feature of the sample $x^{(i)}$ in the pool set.

$$y_p^{(k)} = \frac{\sum_{i=1}^{G} y^{(i)} \exp\left(-d(x^{(k)}, x^{(i)})\right)}{\sum_{i=1}^{G} \exp\left(-d(x^{(k)}, x^{(i)})\right)}$$

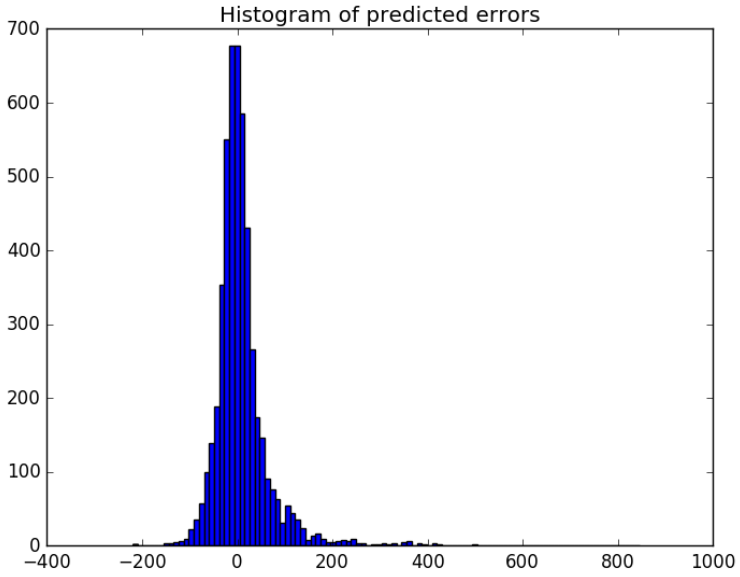
$x^{(k)}, y^{(k)}$ are the samples in the test set, $x^{(i)}, y^{(i)}$ are the samples in the pool set where the $d\left(x, x^{(i)}\right)$ is among the G smallest, $y_p^{(k)}$is the predicted value of $y^{(k)}$.

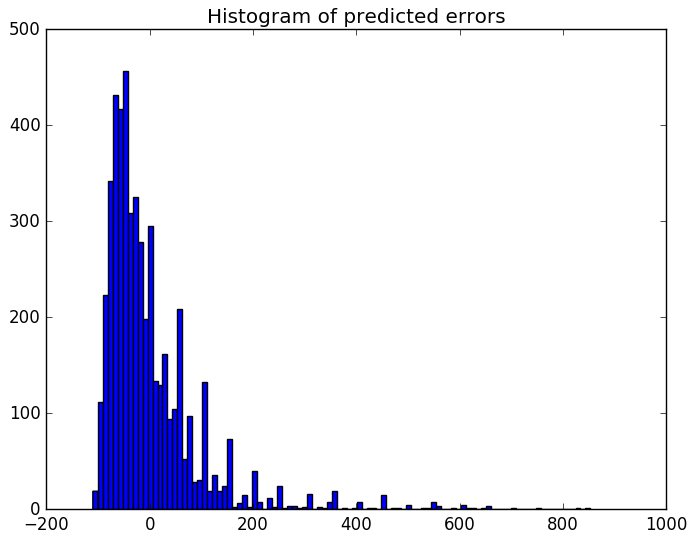
$$\text{Total } error = \sum_{k=1}^{T} \left(y^{(k)} - y_p^{(k)}\right)^2$$

ما ابتدا مقدار وزنی که توسط جنگل تصادفی تولید کرده ایم را امتحان می کنیم و هنگام توصیه قیمت ، قیمت هر خانه در استخر را در نظر می گیریم. از شکل می توان دریافت که قیمت پیش بینی شده بسیار بالاتر از قیمت واقعی است. ما میانگین و واریانس خطا را تحت قوانین مختلف فیلتر کردن مقایسه می کنیم. وقتی

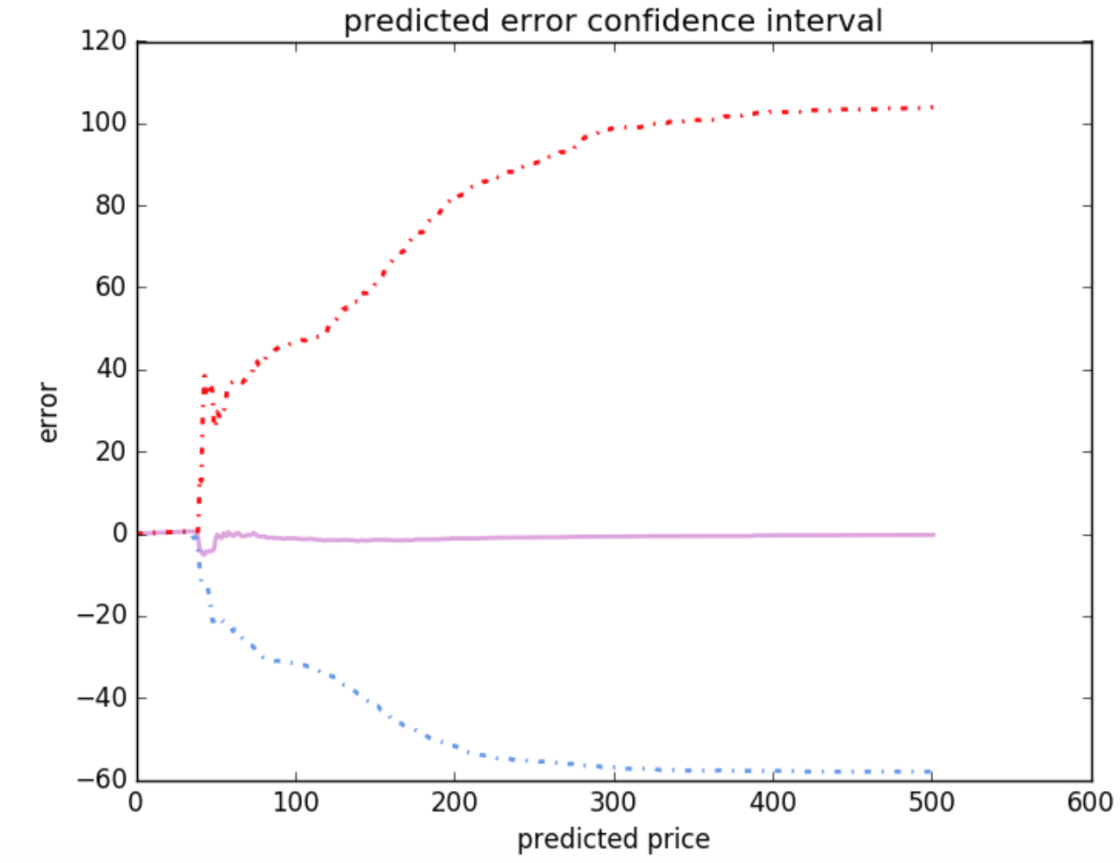
$$G = 10,2,5$$

را می گیریم ، میانه خطای مطلق بین قیمت پیش بینی شده و ارزش واقعی ۱۹,۱۵،۲۲,۰۲،۲۵,۰۱ است.



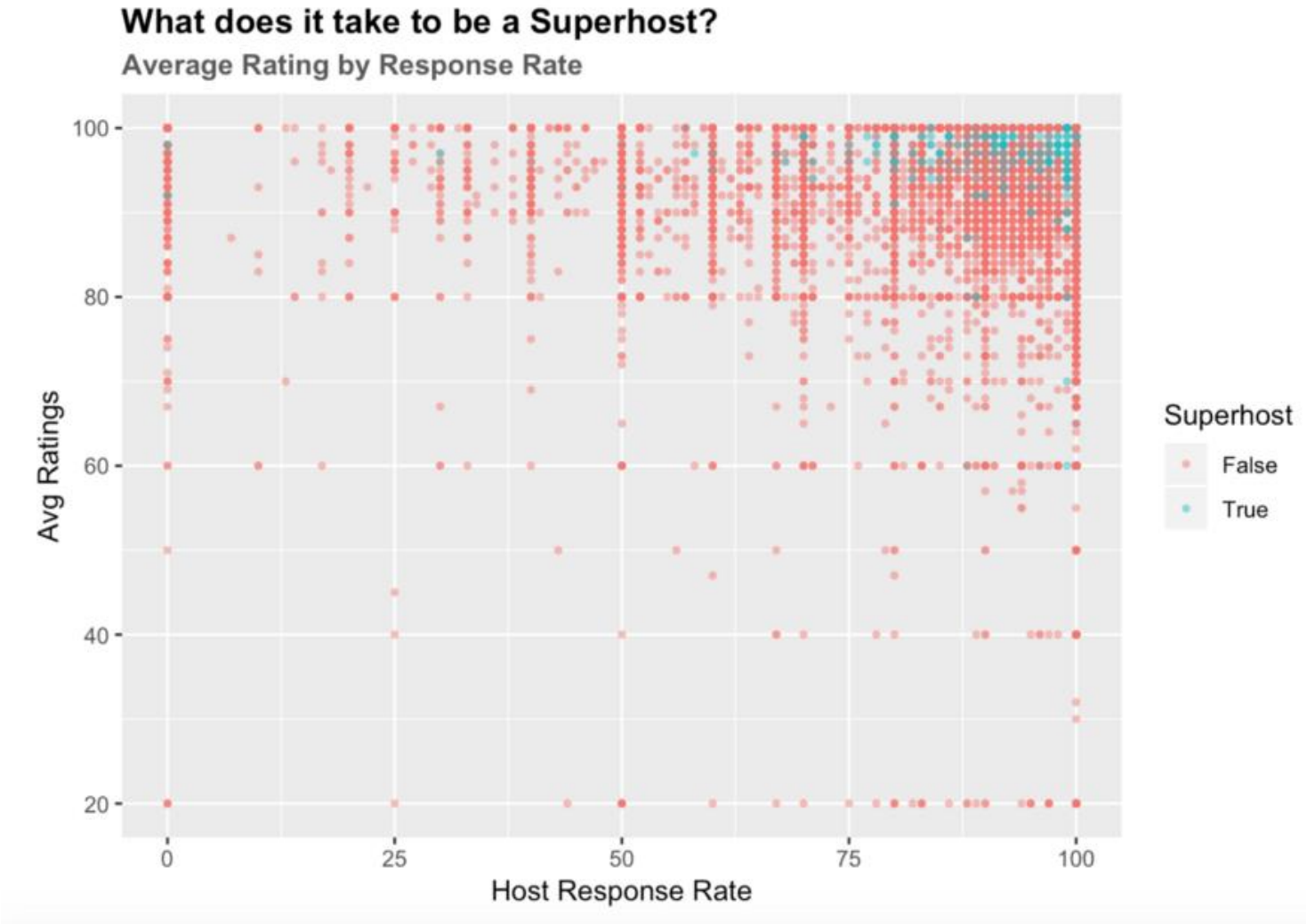


ما هیستوگرامهای بالا G = همه نمونه ها (شکل سمت چپ) و G = 10 (شکل سمت راست) را رسم می کنیم. شکل سمت چپ نشان می دهد که قیمت پیش بینی شده بسیار بالاتر از قیمت واقعی است. شکل درست نشان می دهد که اگر ۱۰ خانه اول را که شبیه خانه تازه وارد شده اند بگیریم و میانگین میانگین وزنی آنها را به عنوان قیمت پیش بینی شده بدست آوریم ، خطا بسیار نرم تر است. ما می توانیم اطمینان زیر را ارائه دهیم ، بسیار کم وجود نخواهد داشت:



همچنین می توانیم از روش گرادیان پروکسیمال برای یافتن مقدار وزنی بهینه که مجموع خطا را به حداقل می رساند ، استفاده کنیم.

تجزیه و تحلیل آنچه برای تبدیل شدن به یک Superhost لازم است:

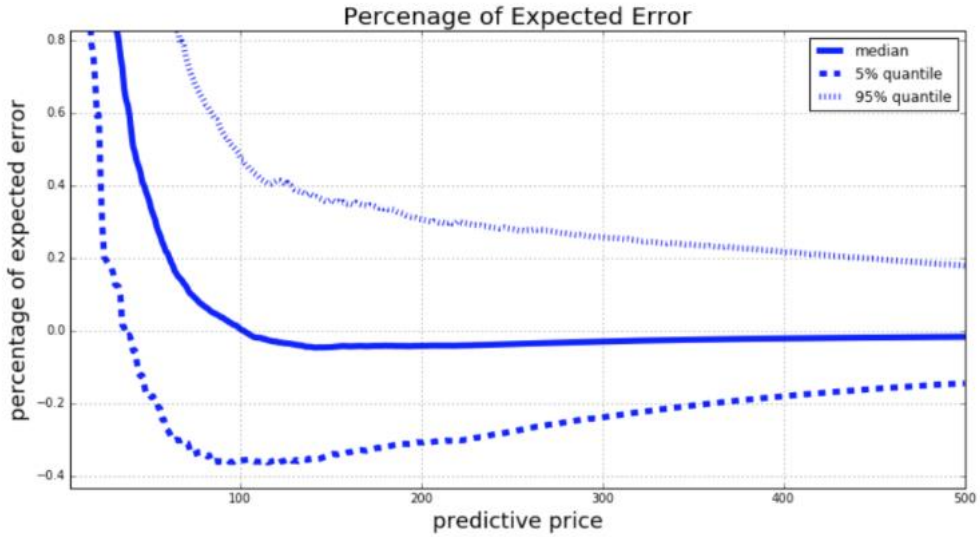


Airbnb عنوان "Superhost" را به بخش کوچکی از میزبانان قابل اعتماد خود اعطا می کند. این برنامه به عنوان یک برنامه تشویقی طراحی شده است که هم برای میزبان ، هم برای Airbnb و هم برای مشتریانشان یک برد محسوب می شود. این شرکت بزرگ به شکل رزروهای بالاتر کسب و کار بیشتری کسب می کند ، مشتری خدمات بهبود یافته دریافت و Airbnb مشتریان را راضی می کند.

اما برای Superhost بودن چه چیزی لازم است؟ Airbnbمجموعه ای از الزامات را دارد که برای یکی شدن باید برآورده شود. حفظ میزان بازبینی بالای ۵۰٪ ، میزان پاسخ دهی بیش از ۹۰٪ و غیره , منطقه , رتبه بندی ، همچنین می توانیم چند میزبان با نرخ پاسخ کمتر از ۷۵٪ (که ۹۰٪ + معیارهای تعیین شده توسط Airbnb را نقض می کند) مشاهده کنیم.

نتیجه

پس از مقایسه مقادیر خطا از سه مدل ، مدل دوم کمترین مقدار خطا را دارد. برای دانستن اینکه مدل سوم چگونه به تصمیم گیری کمک می کند ، بیشتر درصد تقسیم خطا را با تقسیم مقدار خطا بر قیمت پیش بینی محاسبه می کنیم. شکل زیر نشان می دهد که مدل ما برای قیمت پیش بینی شده بزرگتر از USD۱۰۰ ، با اطمینان ۹۰٪ روی خطای ۴۰٪ ، عملی است. برای قیمت پیش بینی شده کمتر از USD۱۰۰ ، مدل ما به راحتی قیمت را پیش بینی می کند. در نتیجه ، مدل ما اگر بیش از ۱۰۰ دلار باشد می تواند پیشنهاد قیمت میزبانهای Airbnb را ارائه دهد.



در مجموع ، Airbnbشاهد یک افزایش خارق العاده در شهر نیویورک است. در داخل Airbnb داده های مشابهی برای چندین شهر مهم دیگر در سراسر جهان میزبانی می شود و معتقیدم مقایسه الگوها و روندهای این شهرها بسیار جالب خواهد بود.

مراجع

_ کتاب آمار ریاضی و کاربردهای آن _ {جان فروند}

_ داده کاوی و مفاهیم _ {ژیاوی هان}

_ سایت Airbnb