رویداد ملی چالش داده

۱۵ اردیبهشت لغایت ۱۵ تیر ۱۳۹۶

جامعه آزاد کلان داده ها

کارگروه فرهنگ سازی حوزه کلان داده با مشارکت گروه ها و جوامع کاربری فعال در کشور **(کاهگل)**

نگارش ۰,۶

فرهنگ سازی حوزه کلان داده با مشارکت گروه ها و جوامع کاربری فعال در کشور

اعضای کارگروه

وزارت ارتباطات و فناوری اطلاعات

پژوهشگاه ارتباطات و فناوری اطلاعات

شورای عالی فضای مجازی

سازمان فناوری اطلاعات

جامعه آزاد رایانش ابری ایران

کارگروه کلان داده دانشگاه صنعتی شریف

مرجع هادوپ ایران

دیده بان کلان داده ها

مرکز تحقیقات پردازش های فوق سریع

مرکز تحقیقات رایانش ابری دانشگاه صنعتی امیرکبیر

حامیان صنعتی اولین رویداد چالش داده

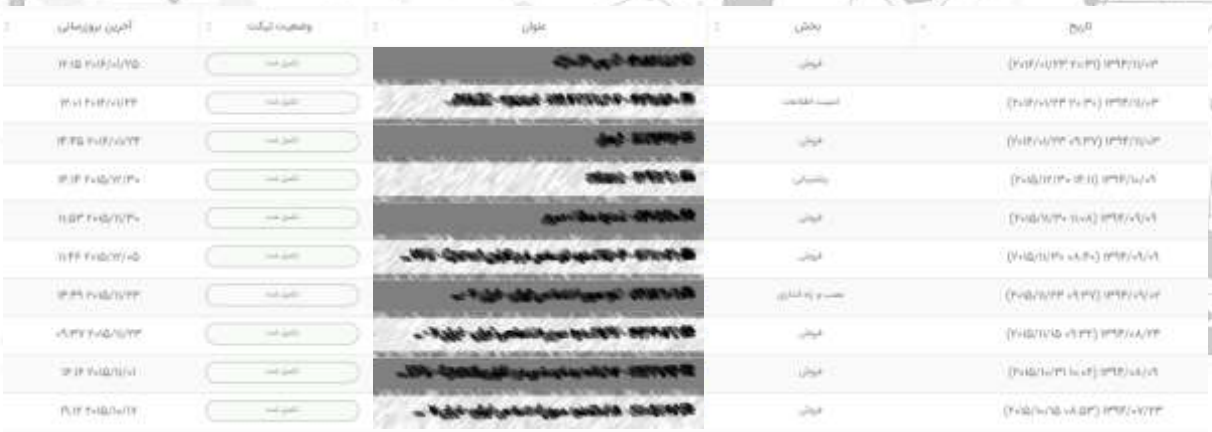
طرح مساله

با گسترش آی تی به عنوان زیرساختی استراتژیک در تمام بخش های جامعه، ایجاد مدلی برای پیش بینی مشکلات و جلوگیری از آن ها هر روز اهمیت بیشتری پیدا می کند. چرا که به حداقل رساندن این مشکلات می تواند باعث بهبود کارایی و جلوگیری از اتلاف زمان و انرژی در تمام مخاطبان شود. شرکت های ارائه دهنده "خدمات میزبانی وب و سرور" بخشی از تامین کنندگان این زیرساخت هستند که وظیفه ی پیش بینی و تامین نیازهای کاربران خود را به عهده دارند. با یک مدل مناسب می توان نارضایتی کاربر از خدمات را در مراحل اولیه شناسایی و راهکاری برای رفع مشکلات وی ارائه نمود. بدیهی است در این مدل باید رفتار کاربران در نظر گرفته شود.

یکی از مهمترین خدمات این شرکت ها سرورهای اختصاصی برای راه اندازی کلود خصوصی و سرویس‌های کلود عمومی است که بنا به نیاز مشتریان به آنها ارائه می شود. هزینه سرورها بنا به انتخاب کاربر در سررسیدهای ماهیانه، سه ماه، شش ماهه و یک ساله پرداخت می شود. با توجه به اینکه مشکلات منجر به از دسترس خارج شدن سرور در این سرویس ها و در نتیجه نارضایتی مشتریان منجر به چالش های زیادی می شود و از سوی دیگر لغو یا انتقال سرورها بین سرویس دهندگان مختلف نیز برای مشتریان با دردسر همراه است. اطلاع داشتن از مشتریانی که احتمال دارد سرویس خود را لغو کنند و اطلاعات خود را به سرویس دهنده‌ی دیگری منتقل نمایند به شرکت‌های میزبانی این امکان را میدهد تا با یافتن راهکار مناسب و پیشنهاد تغییر در سرویس کنونی مشتری از جابجایی اطلاعات جلوگیری نمایند. اینکار باعث جلوگیری از اتلاف زمان برای جابجایی سرویس می گردد که در نهایت به افزایش بهره‌وری در صنعت کشور منتهی می‌شود. بطور مثال یک مشتری که دارای ۳۰۰ وب‌سایت در ابر اختصاصی خود است در هنگام جابجایی اطلاعات با حداقل ۲۴ساعت وقفه در سایت‌ها مواجه خواهد شد. در صورتی که هر کدام از وب سایتها توسط ۳نفر اداره شوند، انتقال اطلاعات برای این مشتری بیش از ۶۶۰۰ساعت کاری از کار افتادگی و زیان خواهد داشت و این زیان با پیش بینی ما قابل پیگیری می‌باشد.

بر همین اساس از شما خواسته خواهد شد تا بر مبنای اطلاعات مشتریان در یک دوره زمانی، احتمال لغو و یا تمدید سرویس وی را پیش بینی کنید. ورودی مساله تعداد تیکت های کاربر در مورد یک سرویس خاص و اطلاعات فاکتورهای پرداختی است. انتظار داریم بتوان بر اساس ترکیبی از وضعیت پرداخت های مشتری(بر اساس فاکتورها) و چالشهای فنی (بر اساس تیکت ها) به مدلی از میزان رضایت یا توان مالی وی (براساس تاخیر در پرداخت) سرویس رسید.

این داده ها بخشی از داده های مشتریان شرکت ایران سرور به عنوان یکی از بزرگترین ارایه دهندگان خدمات میزبانی در ایران می باشد. نمونه داده ارائه شده در فایل آموزش مربوط به فاکتورها و تیکت های یک ماه مشتریان است. تیکت راه ارتباطی مشتریان با کارشناسان در پورتال ایران سرور است که از طریق آن سوالات و مشکلات مختلف خود را مطرح می نمایند.



داده‌های ورودی

۱- هر سرویس با مشخصات یکتایی شناخته می شود که جزئیات آن در جدول **CloudService** آمده است:

CloudService

ServiceID	ProductID	ServerID	UserID	ServiceOrderDate	InvoicePeriod	NextDueDate	ServiceStatus	RegistrationDate
24417	923	0	13109	12/04/2014	Monthly	28/02/2017	Cancelled	09/04/2014
25907	792	135	14206	12/08/2014	Monthly	08/03/2017	Active	29/07/2014
38067	1076	220	22216	12/12/2016	Monthly	13/03/2017	Suspended	09/12/2016
36537	1010	0	22729	21/09/2016	Monthly	19/02/2017	Active	20/09/2016
14103	792	76	11303	02/12/2012	Monthly	19/03/2017	Active	02/12/2012
39262	895	220	24723	13/02/2017	Monthly	16/04/2017	Active	13/02/2017
24956	791	87	15357	31/05/2014	Quarterly	09/01/2017	Active	28/05/2014
26806	894	145	17014	27/10/2014	Monthly	06/03/2017	Suspended	24/10/2014
31812	791	110	18058	05/12/2015	Monthly	28/02/2017	Active	02/12/2015
33504	894	228	20854	12/03/2016	Monthly	15/03/2017	Cancelled	09/03/2016
37912	894	128	23620	04/12/2016	Monthly	26/02/2017	Active	01/12/2016

- ServiceID: شناسه یکتا که نشان دهنده‌ی سرویس مشتری است.
- ProductID: شناسه نوع سرویس (یکی از پلن های کلود عمومی یا خصوصی) که نشان دهنده مشخصات پلن خریداری شده توسط مشتری است.
- ServerID: شناسه سرور (برای کلود خصوصی ۰ و برای کلودعمومی نشانگر شناسه NODE یا سرور میزبان)
- UserID: شناسه یکتا مربوط به مشتری

- RegistrationDate: تاریخ عضویت مشتری در پورتال
- ServiceOrderDate: تاریخ ایجاد سرویس
- InvoicePeriod: دوره زمانی سررسید فاکتور (براساس انتخاب مشتری یک/سه/شش یا دوازده ماهه)
- NextDueDate: برای سرویس های فعال نشاندهنده تاریخ سررسید بعدی و برای سرویسهای لغوشده نشانگر تاریخ آخرین پرداخت و زمان لغو سرویس است.
- ServiceStatus: وضعیت فعال/ غیرفعال بودن سرویس:
  - Active: فعال
  - Terminate/Cancel: هر دو نشان‌دهنده لغو سرویس هستند.
  - Suspend: معلق شدن سرویس که به دلیل تاخیر در پرداخت فاکتور به وجود می آید. یک وضعیت موقت که پس از مدتی به فعال یا لغو تغییر خواهد نمود.

سرورهای اختصاصی با ServerID=0 شناخته می شوند. طرح های مختلف آنها با ProductID های متفاوت از یکدیگر متمایز می گردند و در نتیجه سرویسهای باProductID یکسان مشخصات فنی یکسانی دارند.
سرورهای مجازی علاوه بر اینکه (مانند سرورهای اختصاصی) با ProductID بر اساس ویژگیهای فنی از یکدیگر تفکیک میشوند، بر اساس ServerID هم قابل بررسی هستند؛ ServerID یکسان نشان دهنده سرور میزبان مشترک است که ممکن است برای مشتریان مختلف سرویسهای مختلف چالشهای یکسانی ایجاد کند.

۲- برای هر سرویس مشتری در دوره زمانی های مشخص فاکتور صادر می شود. این فاکتور شامل اطلاعات زیر است:

Invoice

serviceID	UserID	CurrentDueDate	PaidDate	PriceType	discount	invoiceStatus
15124	6748	01/04/2017	01/03/2017	A	1	Paid
16228	7560	09/05/2017	09/03/2017	A	1	Paid
25235	13494	14/05/2017	14/03/2017	A	1	Paid
24102	13874	25/05/2017	25/02/2017	A	0	Unpaid
7484	5410	10/05/2017	10/03/2017	A	0	Paid
26120	15351	14/05/2017	14/03/2017	A	1	Paid
15523	8413	18/07/2017	18/03/2017	A	0	Paid
16894	9640	01/05/2017	01/03/2017	A	0	Paid
20602	10414	07/05/2017	07/03/2017	A	0	Paid

- ServiceID: شناسه یکتا که نشان دهنده‌ی سرویس مشتری و کلید ارتباطی با جدول CloudService است.
- CurrentDueDate:تاریخ سررسید فاکتور در ماه جاری
- PaidDate: تاریخ پرداخت فاکتور
- PriceType: دسته بندی مبلغ فاکتور (A برای ارزانترین و E گرانترین)
- InvoiceStatus: وضعیت پرداخت فاکتور:
  - Paid: فاکتور در زمان ثبت شده در PaidDate پرداخت شده است.
  - Unpaid: فاکتور پرداخت نشده است.
  - Cancel,Refunded: فاکتور لغو شده و یا مبلغ به حساب کاربر برگشت داده شده است.
- Discount: اعمال تخفیف در مبلغ فاکتور

در این مجموعه داده تمامی فاکتورهایی که در سررسیدهای مشخص برای مشتری ایجاد شده است گردآوری شده اند. در مجموعه داده CloudService مشخصه ای به نام InvoicePeriod دوره پرداخت مشتری را مشخص می کند، هرچند این مقدار آخرین وضعیت دوره پرداخت را نگهداری می کند و ممکن است مشتری در طی زمان دوره پرداخت خود را تغییر دهد. با بررسی زمان صدور فاکتورهای مختلف برای یک سرویس می توان به زمان پرداخت ها پی برد. همچنین اختلاف بین CurrentDueDate و PaidDate می تواند پرداخت پیش یا پس از موعد مشتری را نشان دهد. در صورتی که روی فاکتور به هر دلیلی تخفیف صادر شده باشد، مقدار فیلد Discount برای آن یک می شود. شما می توانید میزان تاثیر تخفیف در رضایتمندی مشتری را بررسی نمایید.

برای اینکه بتوانیم با دقت بیشتری رفتار مشتری را پیش بینی نماییم، نیاز به بررسی تجربیات وی در این سیستم داریم. از این رو علاوه بر فاکتورهای مربوط به سرویس های ابری (سرورهای مجازی و اختصاصی)، فاکتورهای مربوط به سایر سرویسهای وی را نیز در این جدول ارائه نموده‌ایم. تفاوت این فاکتورها در فیلد ServiceID است که برای سرویسهای متفرقه فاقد مقدار می‌باشد. این سرویسها شامل دامنه ها و هاست های اشتراکی است. ویژگی مهم این سرویسهای متفرقه در برابر سرویس های ابری، دوره پرداخت آن ها است که عموما به صورت سالانه می باشد. ضمن اینکه این سرویسها در مقایسه با سرویس های ابری مقرون به صرفه تر هستند. با این وجود شما باید علاوه بر تحلیل دقیق فاکتورهای مربوط به سرویسهای ابری از بین تمام مجموعه داده، ارتباط بین ثبت، سفارش و پرداخت فاکتورهای ابری و سرویسهای متفرقه برای مشتری را نیز مورد بررسی قرار دهید.

۳- مشتری می تواند سوالات و مشکلات خود را در تیکت با پشتیبانی ایران سرور در میان بگذارد. کارشناسان سوال وی را پاسخ می دهند و تا زمانی که مشکل برطرف شود این سوال و جواب در همان تیکت ادامه می یابد. بر همین اساس در مجموعه داده سوم اطلاعات تیکت های ایجاد شده برای هر سرویس داده شده است.

Ticket

TicketID	ServiceID	UserID	TicketTime	FirstReplyTime	LastReplyTime	DepartmentID	RepliesCount	PersoneIID	ResponseRate	QC_Check
204259	25460	14495	21/02/2017 09:56	21/02/2017 10:20	21/02/2017 10:20	5	2	50	NULL	0
204262	25689	14967	21/02/2017 10:11	21/02/2017 10:22	21/02/2017 10:42	2	3	57	NULL	0
204271	26217	4502	21/02/2017 10:48	21/02/2017 10:52	21/02/2017 10:52	4	2	48	NULL	0
204272	26216	4502	21/02/2017 10:48	NULL	NULL	4	1	NULL	NULL	0
204285	15920	538	21/02/2017 11:34	21/02/2017 11:35	21/02/2017 11:38	1	3	40	NULL	0
204292	15920	538	21/02/2017 11:52	21/02/2017 12:15	21/02/2017 13:15	5	4	50|50	NULL	0
204316	26136	15361	21/02/2017 12:45	21/02/2017 13:17	21/02/2017 14:49	5	6	50|50|50	50=10	1
204317	26136	15361	21/02/2017 12:46	21/02/2017 16:21	25/02/2017 10:53	6	8	59|38|31|59	59=10,38=9,31=9	1
204363	16292	8186	21/02/2017 16:25	21/02/2017 16:28	21/02/2017 16:28	2	2	56	NULL	0

- TicketID: شناسه تیکت یکتا ایجاد شده برای سرویس
- TicketTime: زمان ایجاد تیکت توسط مشتری
- FirstReplyTime: زمان اولین پاسخ توسط کارشناس
- LastReplyTime: زمان آخرین پاسخ در تیکت (حل شدن موضوع)
- RepliesCount: تعداد سوال و جواب های رد و بدل شده در تیکت
- DepartmentID: شناسه دپارتمان مربوطه در ایران سرور
- PersonelID: شناسه کارشناس پاسخگو در تیکت
- ResponseRate: امتیاز اعطا شده به هر یک از کارشناسان پاسخگو در تیکت توسط مشتری (برای کارشناس a و b امتیاز i و j که عددی بین ۱ تا ۱۰ است به این صورت نمایش داده می شود: "a=i |b=j")
- QC_Check: اولویت پیگیری توسط واحد کنترل کیفیت

مانند فاکتورها، تیکت ها هم برای تمامی سرویسهای فعال مشتری اعم از سرویسهای ابری و سرویسهای دامنه و هاست اشتراکی ارائه شده اند. رابطه بین تیکت های یک مشتری برای سرویسهای مختلفش و رضایت وی از مجموع خدمات ایران سرور چالش جالبی است که با این اطلاعات قابل بررسی است.

تیکت می تواند توسط مشتری یا توسط کارشناسان ایجاد شده باشد و ممکن است بنا به نیاز توسط کارشناسان شیفت ها و دپارتمان های مختلف پاسخ داده شود. در صورتی که تیکت از طرف ایران سرور و برای اطلاع رسانی به کاربر ایجاد شده باشد ممکن است پاسخی دریافت نکند و مقدار متغیر RepliesCount برای آن یک باشد. هیچ تیکتی بدون پاسخ کارشناسان ایران سرور بسته نمی شود.

کاربر می تواند پس از بسته شدن تیکت به پاسخگویی هر کارشناس امتیاز دهد. هرچند همه مشتریان به نظرسنجی پاسخ نمی دهند. علاوه بر این شما می توانید بین شناسه کاربران مختلف و دپارتمان ها نیز به دنبال رابطه باشید.

از آنجایی که پشتیبانی مشتریان توسط واحد کنترل کیفیت بررسی می شود، برخی تیکت های خاص به صورت ویژه بررسی می شوند. این تیکت ها با برچسب QC_check مشخص شده اند. ممکن است برای این برچسب بین تیکت های یک مشتری، یک پلن خاص سرور مجازی یا اختصاصی ارتباط وجود داشته باشد.

serviceID در این سه مجموعه داده مشترک است. هر کاربر می تواند چندین سرویس فعال داشته باشد که برای آن ها به صورت مجزا فاکتور صادر می شود و امکان ایجاد تیکت اختصاصی دارند.

خروجی نهایی و روش ارزیابی

در این چالش شما باید با استفاده از مدل پیش بینی خود، بر اساس داده های یک بازه زمانی مشخص احتمال لغو و یا تمدید سرویس توسط مشتری را در سررسید جاری تخمین بزنید. به عبارتی ورودی برنامه دادگان مربوط به تیکت ها و فاکتورهای سرویس های مجموعه ای از مشتریان است و خروجی آن باید **احتمال لغو سرویس** در سررسید زمانی جاری باشد. برنامه شما در نهایت با برنامه ای که دیگر دوستان می نویسند رقابت خواهد کرد.

نحوه رقابت به این صورت است که به ازای یک فایل ورودی، میزان اختلاف خروجی برنامه ها با داده واقعی بر اساس روش (ROC)* محاسبه می‌شود و هر کدام که اختلاف کمتری داشت برنده مسابقه خواهد بود. برای این کار به ازای سطوح آستانه از ۵۰ درصد تا ۱۰۰ درصد، مشخصه های عملکردی سیستم به شرح زیر محاسبه می‌شود:

- TP (True Positive): پیش بینی **لغو سرویس** که درست طبقه بندی شده است.
- FP (Flase Positive): پیش بینی **لغو سرویس** که اشتباه طبقه بندی شده است.
- TN (True Negative): پیش بینی **عدم لغو سرویس** که درست طبقه بندی شده است.
- FN (False Negative): پیش بینی **عدم لغو سرویس** که اشتباه طبقه بندی شده است.

به ازای هر سطح آستانه، میزان پیش بینی صحیح و میزان اشتباهات بر اساس رابطه Mean Utility محاسبه می‌شود:

$$\textbf{Mean Utility}=0.6TP+0.4TN-0.4FP-0.6FN$$

در نهایت میانگین رابطه فوق به ازای سطوح آستانه ۵۰ تا ۱۰۰ درصد بر اساس رابطه زیر بدست می آید. هر مدلی که مقدار رابطه زیر در آن بیشتر باشد برنده رقابت خواهد بود. (t مقدار آستانه احتمال لغو سرویس می باشد)

$$\frac{1}{n}\sum_{t=50}^{100} MeanUtility(T=t)$$

* https://en.wikipedia.org/wiki/Receiver_operating_characteristic