مسابقه داده‌کاوی امیرکبیر

پیش‌بینی وفاداری مشتریان از روی تراکنش‌های بانکی

بررسی میزان وفاداری مشتریان بانک، یکی از مسائل مهم مطرح در سیستم بانکداری نوین محسوب می‌شود. امسال در مسابقه ملی داده‌کاوی امیرکبیر، قصد داریم این مساله را به چالش کشیده و با استفاده از راه‌حل‌های نوین عرصه داده‌کاوی و یادگیری ماشین به بهبود سرویس‌دهی بانک کمک کنیم.

سناریو

داده‌های مربوط به تراکنش‌های بانکی ۵ ماه مشتریان موجود است. این داده‌ها به ترتیب زمانی ذخیره نشده‌اند، ولی با استفاده از ویژگی‌های "transactionDate" و "transactionTime" می‌توان توالی زمانی تراکنش‌ها را به دست آورد.

موجودی هر مشتری بعد از انجام هر تراکنش موجود است. با این ویژگی می‌توان موجودی هر مشتری را در پایان روز محاسبه کرد. بنابراین می‌توان برای هر مشتری ("customerId") میانگین موجودی هر ماه به همراه انحراف معیار آن را به دست آورد. این دو متغییر را که متغیرهای هدف هستند "balanceStd" و "balanceAvg" می‌نامیم.

از شرکت کنندگان انتظار می‌رود با استفاده از تکنیک‌های داده‌کاوی و یادگیری ماشین مدل مناسبی بر اساس تراکنش‌های ۵ ماه اول کاربرها طراحی کنند و از آن مدل برای پیش‌بینی دقیق متغیرهای هدف در ماه بعدی استفاده کنند. متغیر های هدف برای هر کاربری که حداقل یک بار در ۵ ماه اول تراکنش داشته است، در بارگذاری جواب نهایی تیم مورد انتظار است.

داده‌ها

داده‌های گمنام‌سازی شده مربوط به تراکنش‌های واقعی یک بانک است که به همراه صورت مساله در سایت بارگذاری شده است. به دلیل حجم بالای داده های مسئله، داده ها به ۵ بخش(هر ماه یک بخش) تقسیم شده اند. در ادامه چند نکته در مورد فایل‌های ضمیمه شده آورده شده است:

۱. هر سطر از داده‌ها نشانگر یک تراکنش بانکی است که با علامت‌های زیر جدا شده‌اند:

“CR”: carriage return (0xD), “LF”: line feed (0xA) or “CR” and “LF”

۲. خط اول هر فایل داده حاوی نام ویژگی‌ها است.
۳. در تمام سطرها، علامت "," ستون‌ها را از هم جدا می‌کند.
۴. علامت "." نشانگر جدا کننده اعشار است.

فایل "AUT DMC 2017 - Features.pdf" موجود در "task.zip" حاوی لیستی از تمام ویژگی‌ها و توضیحات مربوط به آن ویژگی‌ها است.

دقت کنید که چون داده‌ها مربوط به داده‌های واقعی ذخیره شده در بانک هستند امکان وجود داده‌ی پرت وجود دارد و تیم‌ها بنا به صلاحدید خود می‌توانند هر پردازشی روی آن‌ها انجام دهند.

بارگذاری فایل نهایی

تیم‌ها فرصت دارند تا قبل از پایان مسابقه (ساعت ۱۸:۳۰ اول آبان) یک فایل csv با فرمت زیر در صفحه تیم خود بارگذاری کنند. این فایل به عنوان پیش‌بینی‌های تیم شرکت کننده روی داده تست در نظر گرفته خواهد شد و با مقادیر واقعی مقایسه و امتیاز تیم محاسبه خواهد شد. امکان بارگذاری مجدد وجود دارد، ولی فقط آخرین بارگذاری به عنوان پاسخ نهایی تیم در نظر گرفته خواهد شد.

نام ستون	توضیحات	محدوده‌ی مقادیر
customerId	شماره مشتری	رشته
balanceAvg	میانگین موجودی حساب برای ماه ششم	اعداد اعشاری

balanceStd	انحراف معیار موجودی حساب برای ماه ششم با توجه به موجودی پایان روز	اعداد اعشاری

فایل نهایی باید دقیقا با فرمت بالا بوده و پیش‌بینی‌های انجام شده برای هر کاربر فقط یک بار در فایل آورده شده باشد. مثال زیر یک بارگذاری صحیح را نشان می‌دهد:

customerId,balanceAvg,balanceStd
guke_232,320.2,24.4
guke_245,-220.6,14.8
guke_322,1320.8,4.2
...

فرمت فایل نهایی قابل قبول یکی از فرمت‌های zip یا txt است. نام فایل به صورت <email>.zip قابل قبول است که به جای email آدرس ایمیل سرگروه باید قرار داده شود.

معیار ارزیابی

برای هر "customerId" ای که حداقل یک بار در داده های یادگیری ۵ ماه اول دیده شده اند، تیم ها موظفند دو مقدار "balanceAvg" به معنای میانگین موجودی ماه ششم و "balanceStd" به معنای انحراف معیار موجودی ماه ششم را پیش بینی کنند. ارزیابی با معیار OVL(Overalpping Coefficient) انجام خواهد پذیرفت. بدین صورت که از میانگین و انحراف معیار پیش بینی شده برای هر "customerId" یک توزیع نرمال با نام f_1 ساخته خواهد شد. همینکار برای مقادیر واقعی انجام شده و f_2 نامیده می‌شود. مساحت مشترک بین این دو توزیع به عنوان امتیاز کسب شده توسط تیم برای آن "customerId" در نظر گرفته خواهد شد(رابطه زیر). تیم برنده تیمیست که جمع امتیاز هایش بیشتر باشد.

$$OVL_i = \int \min[f1(x) \,.\, f2(x)]$$

در صورت تصمیم تیم داوری برای اعمال تغییر در هر قسمتی از توضیحات و قوانین، مراتب از طریق ایمیل به تیم‌ها اعلام می‌شود.