بخش اول – معرفی مجموعه داده

- تجارت الکترونیک یک بخش تجاری بسیار بزرگ است که به خریداران امکان دسترسی به انواع کالاها و خدمات را با چند کلیک می دهد. بسیاری از پلتفرم های خرید محبوب مانند آمازون یا علی بابا هر سال میلیون ها تراکنش را پردازش می کنند. در سال های اخیر به علت بیماری کرونا خرید آنلاین به شدت افزایش یافته است از این رو بازار خرید آنلاین بسیار رقابتی شده است و برای پلتفرم های خرید آنلاین قوی و نوآورانه مهم است. یک راه ممکن برای افزایش معاملات خرید آنلاین، درک و پاسخگویی به رفتار خریداران آنلاین است. با توجه به داده های خرید آنلاین کافی و تکنیک های یادگیری ماشینی، **می توان قصد خرید بازدیدکنندگان وب سایت را تعیین کرد** پس در این تمرین می خواهیم با کاربرد الگوریتم های ماشین لرنینگ در Marketing Analytics (تجزیه و تحلیل بازاریابی) که یکی از مباحث مطرح در سال های اخیر است، بیشتر آشنا شویم.
- دیتاست داده شده در فایل زیپ، شامل اطلاعات مختلف مربوط به رفتار مشتری در وب سایت های خرید آنلاین کالا می باشد که به ما کمک می کند تا تجزیه و تحلیل بازاریابی را انجام دهیم و KPI ها و معیارهای مربوط به آن را درک کنیم. برای اطلاعات بیشتر راجع به به تجزیه و تحلیل بازاریابی و KPI، فیلم آنلاین جلسه ی ۴ (۱۵ آبان) را می توانید مشاهده کنید.
- این دیتاست، شامل بردارهای ویژگی متعلق به ۱۲۳۳۰ session است به این معنا که هر سطر رفتار یک مشتری را برای خرید کالا نشان می دهد و در کل ۱۰ ویژگی عددی و ۸ ویژگی طبقه بندی شده دارد. در جدول زیر یک توضیح کلی راجع به ویژگی های دیتاست، داده شده است.

جدول ۱: توضیح کلی ویژگی ها

توضیحات کلی	ویژگی ها
این ویژگی ها تعداد انواع مختلف صفحات بازدید شده توسط بازدید کننده در آن session و کل زمان صرف شده در هر یک از این دسته بندی صفحات را نشان می دهد .مقادیر این ویژگی‌ها از اطلاعات URL صفحاتی که کاربر بازدید می‌کند، گرفته می‌شود و زمانی که کاربر اقدامی را انجام می‌دهد، به‌طور هم‌زمان به‌روزرسانی می‌شود. (حرکت از صفحه ای به صفحه دیگر)	"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" , "Product Related Duration"
معیارهای اندازه گیری شده توسط "Google Analytics" برای هر صفحه در سایت تجارت الکترونیک را نشان می دهند. ویژگی "Bounce Rate " برای یک صفحه وب، به درصد بازدیدکنندگانی اشاره دارد که از آن صفحه وارد سایت می شوند و سپس (" Bounce") را بدون انجام هیچ گونه درخواست دیگری به سرور تجزیه و تحلیل در آن session، ترک می کنند. ویژگی «Exit Rate » درصد بازدیدکنندگان یک صفحه در وب سایت که از آن وب سایت به وب سایت دیگری خارج می شوند. ویژگی "Page Value" نشان دهنده میانگین مقدار یک صفحه وب است که کاربر قبل از انجام معامله تجارت الکترونیک از آن بازدید کرده است.	"Bounce Rate", "Exit Rate" , "Page Value"
نشان‌دهنده نزدیکی (closness) زمان بازدید از سایت به یک روز خاص (مثلاً روز مادر، روز ولنتاین) است که در آن session به احتمال زیاد با تراکنش نهایی می‌شوند .	"Special Day"
اطلاعاتی شامل ماه، نوع مرورگر، مذهب، نوع ترافیک، نوع بازیدکنندگان (جدید یا قدیمی) و اینکه ایا خرید در اخر هفته صورت گرفته است، به ما می دهد.	Month,Browser, Region, Traffic type, Visitor type(as returning or new visitor), ,Weekend
نشان دهنده ی این است که آیا بازدید کننده خریدی انجام داده است یا خیر. **ویژگی "Revenue " را به عنوان برچسب کلاس استفاده کنید.**	Revenue

بخش دوم – معرفی مساله مورد بررسی

- در ادامه با استفاده از مجموعه داده معرفی شده در قسمت قبل، قصد داریم یک مسئله طبقه بندی شده را در دنیای واقعی حل کنیم یعنی می خواهیم ببینیم با توجه به ویژگی های داده شده (رفتار مشتری در خرید)، مشتری خریدی انجام داده است یا خیر. با توجه به این که کارایی مدل نهایی ارائه شده اهمیت به سزایی در تشخیص دقیق دارد، روش های مختلف پیش پردازش، انتخاب ویژگی و طبقه بندی را مقایسه کنید و در نهایت بهترین پروسه(روش کاهش بعد در صورت لزوم و طبقه بندی) پردازش اطلاعات برای مجموعه داده را پیشنهاد دهید.
- مدل های مختلف را بر اساس متریک ها با یکدیگر مقایسه می کنند. برای اطلاعات بیشتر در مورد هر یک از متریک ها می توانید به این لینک مراجعه کنید.
- توجه کنید که لزومی ندارد از همه ی این متریک ها استفاده کنید. به انتخاب خودتان **۳** تا از بهترین متریک ها را انتخاب کنید و این ۳ متریک را **با** متریک Accuracy برای هر مدل گزارش کنید و دلیل بیاورید که چرا این ۳ تا متریک را انتخاب کردید.

بخش سوم – روش های پیش پردازش و انتخاب ویژگی

- همانطور که می دانید یکی از قسمت های مهم پیش بینی مدل، پیش پردازش است. پس تمامی پیش پردازش های ممکن را انجام دهید.
  - در صورت نیاز تمیزسازی داده انجام شود و دلایل آن توضیح داده شود. به طور مثال اگر قسمت‌هایی از داده از دست رفته است و شامل مقادیر nan هست با روش مناسب مقادیر nan جایگزین شود و دلیل انتخاب روش بیان گردد.
  - با توجه به این که در این مجموعه داده تعداد نمونه های دو کلاس نامتوازن است، چه روشی را برای مواجه با این مشکل انتخاب می کنید؟ (روش های متفاوتی مانند استفاده از متریک مناسب، ایجاد سمپل جدید برای داده یادگیری و ... را می توان استفاده کرد و یا با توجه به جنس مساله از عدم توازن چشم پوشی کرد.) به دلخواه روشی را انتخاب کنید و توجه داشته باشید انتخاب شما، مقایسه عملکرد طبقه بندها را در قسمت بعد تحت تاثیر قرار می دهد.
  - بررسی کنید ایا در این دیتاست داده شده، ما با نحسی ابعاد مواجه هستیم؟ اگر جواب شما بله یا خیر است با یکی از روش های کاهش بعد و پیاده سازی آن دلایل خود را **نشان دهید** و اصولا چرا از روش های کاهش بعد استفاده می کنند.
  - حال در ادامه برای این که کمی بیشتر با مفهوم Feature selection آشنا شویم، با استفاده از الگوریتم RandomForestRegressor بررسی کنید که کدام ۶ فیچر(ستون ها) بیشترین نقش را در ستون کلاس (Revenue) دارند؟
  - به نظر شما می توان با همین ۶ ویژگی در ادامه کار کنیم و ۱۱ ویژگی دیگر را برای ستون Revenue درنظر نگیریم؟ آیا وجود این ویژگی ها تاثیری در دقت مدل ها دارند؟

بخش چهارم – طبقه بندهای Discriminative

در این بخش می خواهیم از الگوریتم های Discriminative استفاده کنیم که سعی بر پیدا کردن مرزهای کلاس ها با استفاده از داده های یادگیری دارند و بر اساس این که نمونه جدید مشاهده شده در کدام طرف مرز قرار گرفته است طبقه بندی صورت می گیرد. در زیر چند نمونه از این طبقه بند ها معرفی شده است که باید میان آن ها مقایسه انجام شود و مدلی که بهترین عملکرد را دارد معین شود. **(از بهترین پارامترهای هر طبقه بند استفاده کنید).**

- دقت کنید که هدف اصلی آن است که میان گزینه های موجود روشی را بیابید که عملکرد بهتری را ارائه دهد. بنابراین باید با کوشش و خطا(همراه با استدلال و شهودی که در مورد مدل ها دارید) این مدل را پیدا کنید.
- برای سنجش عملکرد مدل خود ۳ معیار انتخابی خود **و** معیار Accuracy را برای الگوریتم های زیر گزارش کنید.
- **SVM**
- **Decision Tree**
- **KNN**
- **MLP**
- **Logistic Regression**

- نمودار Accuaracy بدست امده از طبقه بند های بالا را در یک نمودار رسم کنید. کدام طبقه بند بهترین عملکرد را از لحاظ متریک Accuaracy دارد؟
- آیا می توانید برای این دیتاست مدل های دیگری را ارائه دهید که دقتی بالاتر از ۵ مدل داده شده بدهد؟ **(امتیازی)**

نکات تحویل

- مهلت ارسال این تمرین تا پایان روز جمعه ۲۴ دی ماه خواهد بود.
- انجام این تمرین به صورت یک نفره می‌باشد.
- در رسم نمودارها به فرمت، سایزها، برچسب محورها، تمیزی و مواردی که برای رسم نمودار آموختید، توجه شود. همچنین نمودار ها و جداول را به صورت واضح و همراه با زیرنویس و بالانویس در گزارش خود بیاورید.
- لطفا هر گونه فرض در حل سوالات را در گزارش خود ذکر کنید.
- برای هر کد که در فایل نهایی ضمیمه می‌کنید، گزارش بنویسید. کدهای ضمیمه شده بدون گزارش مربوطه نمره‌ای نخواهند داشت. (این گزارش‌ها تنها معیار تفکیک کد شما و کدهای موجود در منابع مختلف مانند اینترنت خواهند بود.)
- لطفا یک فایل زیپ شامل pdf گزارش، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمائید.

HW7_[Lastname]_[StudentNumber].zip

- در صورت وجود سوال و یا ابهام می‌توانید از طریق رایانامه زیر با دستیار آموزشی در ارتباط باشید:

z.habibzadeh213@gmail.com