دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

درس یادگیری ماشین

دکتر عباس حسینی

پروژه‌ی پایانی

زمان انتشار: ۱۳ آذر ۱۴۰۰

زمان تحویل: ۱۵ بهمن ۱۴۰۰

۱- مقدمه: معرفی فضای مسئله

امروزه مسئله‌ی پیش‌بینی واکنش کاربران به تبلیغات استفاده‌های تجاری گسترده‌ای دارد و سرمایه‌گذاری بسیاری روی آن می‌شود. این که پیش‌بینی کنیم کاربران به هر تبلیغ ما چه واکنشی نشان خواهند داد باعث خواهد شد تا ما هزینه‌ی تبلیغات را کاهش دهیم یا بهره‌وریمان از هر تبلیغ را افزایش دهیم. در این پروژه قصد داریم تا به یکی از مسائل اساسی این حوزه حمله کنیم.

۲- شرح مسئله: سوال دقیقا چیه؟

قصد داریم تا با در دست داشتن ویژگی‌هایی از یک تبلیغ اینترنتی با استفاده از شبکه‌های عصبی عمیق **پیش‌بینی کنیم که آیا کلیک شدن روی یک محصول، در نهایت منجر به خریداری شدن آن می‌شود یا نه.** دقت کنید که می‌دانیم هر تبلیغ نمایش داده لزوما کلیک شده و قصد داریم تا بین تبلیغات کلیک شده، خریداری شدن یا نشدن، که اصل سود فروشنده در آن است، را در تشخیص دهیم.

در واقع ورودی مدل ما یک سطر از ویژگی‌های یک تبلیغ اینترنتی است و خروجی مورد نظر پیش‌بینی مدل ما از خریده شدن یا نشدن آن محصول پس از کلیک شدن روی تبلیغ خاصی از آن است.

۳- دیتاست چیه؟

در این پروژه از دادگان criteo live data استفاده می‌کنیم که که اطلاعات مربوط به تبلیغات اجناس مختلف طی ۹۰ روز در آن ذخیره شده اند. در هر سطر این دادگان، اطلاعاتی در مورد یک کلیک انجام شده بر روی این تبلیغات ذخیره شده است. این اطلاعات شامل منجر شدن آن کلیک به خرید محصول، اختلاف زمانی بین کلیک کردن روی تبلیغ و خریده شدن محصول توسط کاربر و خصوصیات محصول تبلیغ شده و خود کاربر هستند. این اطلاعات به صورت مفصل در بخش بعد توضیح داده شده‌اند.

ستون‌های دادگان

نام ستون	معنی ستون
sale	آیا یک کلیک منجر به خرید شده است یا خیر (به چشم برچسب داده ببینیدش)
SalesAmountInEuro	درآمد ناشی از فروش یک محصول. توجه کنید که این مقدار ممکن است متفاوت از قیمت محصول باشد. در صورتی که کلیک منجر به خرید نشده باشد، مقدار این ستون برابر **منفی یک** خواهد بود.
time_delay_for_conversion	زمان بین کلیک روی تبلیغ و خریده شدن محصول نمایش داده شده توسط کاربر در اینجا هم مقدار متناظر با کلیک‌هایی که منجر به خرید نشده‌اند **منفی یک** خواهد بود. اگرچه در اکثر مواقع، فاصله‌ی زمانی بین خرید و کلیک کمتر از یک روز و ۱۰ ساعت است، این مقدار ممکن است تا ۹۰ روز باشد.
click_timestamp	زمان انجام شدن یک کلیک. دیتاست بر اساس این مقدار مرتب شده است.
nb_clicks_1week	تعداد کلیک‌هایی که یک تبلیغ در هفته‌ی آخر بازه‌ی ۹۰ روزه دریافت کرده است.
product_price	قیمت محصول
product_age_group	گروه سنی که خریدار اصلی محصول مربوط به یک تبلیغ هستند.
device_type	نوع وسیله‌ای که کاربر با آن کلیک را انجام داده.
audience_id	این ویژگی به دلیل حفظ حریم خصوصی کاربران مخفی شده.
product_gender	جنسیت افرادی که محصول برای استفاده‌ی آنها طراحی شده.
product_brand	سازنده‌ی محصول
product_category	هفت ستون مربوط به این لیبل هستند که دسته‌ای که محصول مربوط آن است را به صورت one-hot-encoding مشخص می‌کنند.
product_country	کشوری که محصول در آن به فروش می‌رسد.
product_id	شناسه هر محصول
product_title	نام محصول به صورت رمزنگاری شده.
partner_id	شناسه فروشنده محصول
user_id	شناسه‌ای یکتا برای نمایش هر کاربر

نکته: مقدار Missing Value برای تمام سطرها **منفی یک** است، به جز سطر click_timestamp که مقادیر ناموجود آن با **صفر** نمایش داده شده‌اند.

نکته: حجم دادگان هم برای این پروژه کاهش پیدا کرده و حدود ۱۰۰ هزار سطر داده برای training در نظر گرفته شده.

۴- نحوه‌ی ارزشیابی

گرفتن ۱۰۰ امتیاز از پروژه به معنی کل نمره‌ی پروژه است و با گرفتن تمام قسمت‌های امتیازی می‌توانید حداکثر تا ۱۲۰ نمره از پروژه دریافت کنید. در ابتدا لطفا یک بار کل مستند پروژه را بخوانید و سپس به بخش لینک‌های مفید مراجعه کنید و حتما لینک اول را با اولویت مطالعه کنید.

۴.۱- تحلیل اکتشافی داده[1] - (۱۵ نمره)

تمیز کردن داده، کشف ویژگی‌ها و الگوهای درون آن، ترسیم نمودارها و به طور کلی «تحلیل کردن داده‌ها» یکی از مهم‌ترین بخش‌های انجام پروژه‌های یادگیری ماشین است. در این قسمت از شما انتظار داریم تا جایی که به نظرتان مفید است بتوانید شواهدی از داده به دست بیاورید که بتواند شما را با استخراج ویژگی‌های مهم و طراحی مدل بهتر در مراحل بعدی کمک کند.

مهم است که بتوانید علاوه بر فراهم کردن شواهد برداشت خودتان را هم از آن‌ها بیان کرده و بگویید که این شواهد چه کمکی به شما در اخذ نتایج بهتر و شناخت داده کرده است. (لینک ۲ و ۳)

۴.۲- مهندسی ویژگی‌ها[2] - (۱۰ نمره)

پس از بررسی کردن داده و به دست آوردن شواهد و کاستی‌های موجود در داده، مانند تعداد سطرهای خالی زیاد در یک ستون، باید بتوانید با در نظر گرفتن اطلاعاتی که از قسمت قبل به دست آورده‌اید دادگان اولیه را اصلاح کرده و آن را برای ورودی دادن به الگوریتم «ماشین‌فهم‌تر» کنید.

در این مرحله شما باید تصمیم بگیرید که چه ستون‌هایی را باید از داده حذف کنید (چرا؟)، خانه‌های بدون مقدار را با چه مقادیر دیگری مقداردهی کنید، ستون‌ها را چگونه نرمالایز کنید و... (لینک ۴)

---

[1] Exploratory Data Analysis (EDA)
[2] Feature Engineering

۴.۳- مسیر منتهی به مدل نهایی - (۱۵ نمره + ۵ امتیازی)

در نهایت قرار است که یک مدل نهایی در دست داشته باشیم و از آن استفاده کنیم، ولی در مسیر رسیدن به مدل نهایی لازم است تا دو کار انجام شود: ۱- تست کردن مدل‌ها با معماری‌ها و الگوریتم‌های مختلف ۲- tune کردن هرکدام از مدل‌ها و بررسی نتایج آن‌ها

به همین خاطر از شما می‌خواهیم تا علاوه بر ارائه‌ی مدل نهایی ارائه شده، مسیری که برای رسیدن به آن طی کرده‌اید را گزارش کرده و در مورد تصمیم‌گیری‌هایی که در این مسیر انجام داده‌اید استدلال کنید. لازم است تا حداقل دو مدل مختلف یادگیری عمیق را در این بخش تست کنید و همچنین مسیر tune کردن آن‌ها را توضیح دهید.

امتیازی ۵ نمره: تست کردن یک مدل غیر از شبکه‌عصبی و مقایسه‌ی عملکرد آن

نکته: می‌توانید برای شروع به مطالعه از مدل‌های پیشنهادی در انتهای مستند پروژه شروع کنید.

۴.۴- نتایج مدل نهایی - (۲۵ نمره + ۸ امتیازی)

در این مرحله لازم است تا مدل یادگیری عمیقی که در نهایت انتخاب می‌شود و یادگیری روی آن صورت می‌گیرد مورد بررسی قرار بگیرد.

در این مرحله از شما انتظار داریم تا:

- ۱۰ نمره: دلیل انتخاب این مدل را بیان کنید و نحوه‌ی کارکرد آن را به طور خلاصه توضیح دهید.
  - روند تغییر عملکرد مدل در حین یادگیری را ارائه کرده و متریک‌های مفیدی برای بررسی مدل ارائه و گزارش کنید.
  - ضعف‌های مدل خود را بررسی کنید و بررسی کنید که نقاط ضعف و قوت مدل شما در چه سناریوهایی رخ می‌دهد.
- ۱۵ نمره: متریک اصلی‌ای که باید آن را بیشینه کنید F-score[3] است که تیم تدریس هنگام تحویل پروژه با داده‌ی تست این مقدار را با تست کردن مدل شما محاسبه می‌کنید.
  - چارک اول و دوم کلاس ۱۵ نمره‌ی کامل را از این بخش می‌گیرند.
  - چارک سوم و چهارم به ترتیب ۱۰ و ۵ نمره از این بخش می‌گیرند.
    - در صورتی که بیش از ۵ واریانس از میانگین کلاس فاصله نداشته باشند. یعنی عملکرد مدلشان خیلی پرت نباشد!

[3] https://en.wikipedia.org/wiki/F-score

- ۸ نمره امتیازی:
  - میزان اهمیتی که مدل به هر کدام از ویژگی‌های ورودی می‌دهد و تفسیر چگونگی تصمیم‌گیری مدل را انجام دهید (لینک ۵) (۳ امتیاز)
  - به ازای هر نیم واریانس بیشتر بودن از میانگین F-score کلاس یک نمره امتیازی دریافت می‌کنید. (حداکثر ۵ امتیاز)

۴.۵- دیپلوی‌منت - (۲۰ نمره + ۱۲ نمره امتیازی)

برای این که یک پروژه‌ی یادگیری ماشین بتواند توسط کاربرهای مختلف استفاده شود می‌بایست در فضایی دیپلوی شود و بتوان با آن توسط یک پروتکل صحبت کرد و با ورودی دادن به آن، از آن خروجی دریافت کرد. یکی از مهم‌ترین بخش‌های این پروژه استفاده کردن از MLFlow برای دیپلوی کردن آن است.در این مورد لازم است تا بخش‌های مختلف پروژه را به کانتینر[4] تبدیل کنید (۱۰ نمره) و با ساختن یک پایپ‌لاین[5] از این کانتینرها (۱۰ نمره) بتوانید آن را روی کامپیوتر شخصی خود دیپلوی کنید (۱۲ نمره).

برای آشنا شدن با اهمیت MLOps کافی است به نیازمندی‌های شغلی ML Engineer شرکت‌های مختلف مانند اسپاتیفای و آمازون توجه کنید! توصیه می‌شود که لینک‌های مفیدی که در این رابطه در پایین پروژه گنجانده شده را در ابتدای پروژه مطالعه کرده و حتما پروژه را از ابتدا با در نظر داشتن دیپلویمنت نهایی آن پیش ببرید.

برای آشنایی اولیه با MLFlow به این لینک مراجعه کنید. همچنین برای چگونگی دیپلوی کردن مدل یادگیری عمیق می‌بایست از این مستند بهره بگیرید. دقت کنید که اکثر مستندهای موجود در این رابطه در نهایت مدل را روی فضای ابری‌ای مانند AWS یا Google Cloud دیپلوی می‌کنند، ولی این بخش از شما خواسته نشده و مدل را باید روی سیستم شخصی خودتان دیپلوی کنید.

۴.۶- گزارش - (۱۰ نمره)

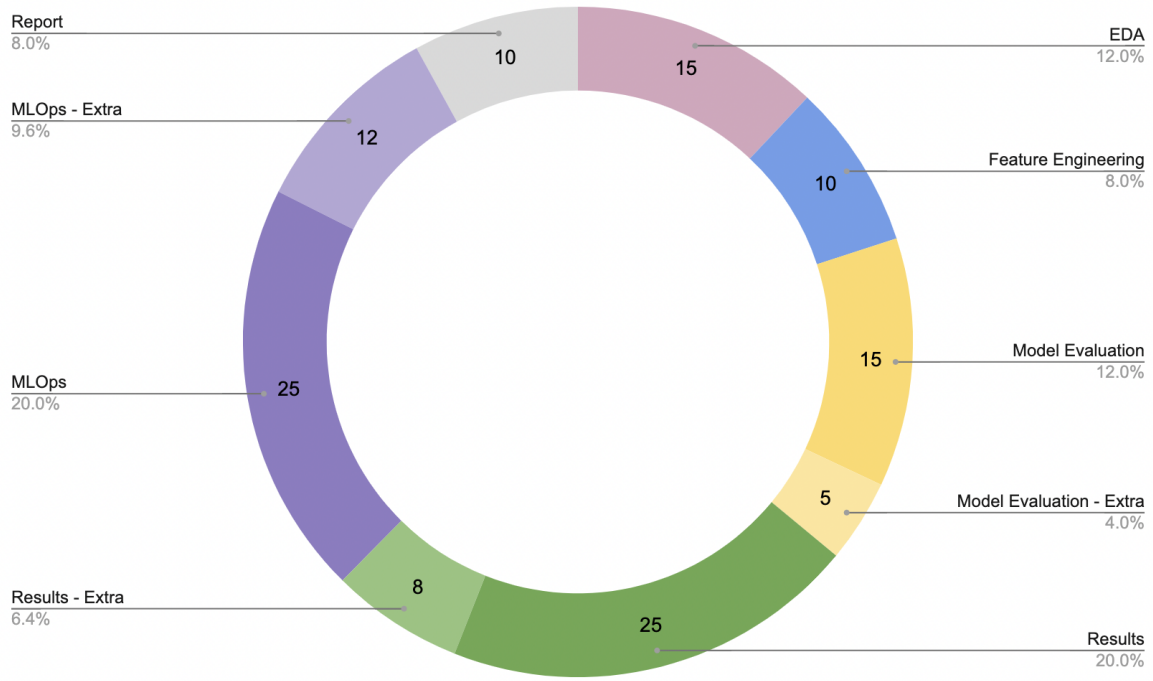
ارائه کردن مسیر و نتایج پروژه یکی از مهم‌ترین مهارت‌های نرمی است که یک متخصص یادگیری ماشین باید بتواند آن را به خوبی انجام دهد، چنان که توانایی data storytelling یکی از مهم‌ترین ویژگی‌های یک دانشمند داده در سال ۲۰۲۱ شمرده شده است. لذا از شما انتظار داریم تا مستند مناسبی از پروژه‌ای که انجام می‌دهد تهیه کنید. (لینک ۶)

---

[4] container
[5] pipeline

شما در این مسئله باید بتوانید به صورت ساده و خلاصه مسیری که در روند حل مسئله طی کرده‌اید و چالش‌هایی که با آن روبرو شده‌اید را توضیح دهید و تصمیماتی که در حل این چالش‌ها گرفته‌اید را توجیه کنید. در فرآیند این توضیح هر چه از مصورسازی‌ها و مثال‌های مناسب‌تری استفاده کرده باشید ارزش‌مندتر است.

مهم است که مقاله‌ها و منابعی که در طی مسیر حل مسئله از آن استفاده می‌کنید را ذکر کنید و به یک دیگر با به اشتراک گذاشتن منابع مفید در جهت انجام پروژه‌ی بهتر کمک کنید. لطفاً وقت کافی برای نوشتن گزارش اختصاص دهید و به منظور از دست نرفتن جزئیات این بخش را موازی کارهای دیگر پیش ببرید.



شکل ۱: توزیع بخش‌های نمرات پروژه با مجموع ۱۲۵ - ۲۵ نمره‌ی امتیازی

۵- چند نکته

- این پروژه به منظور تقویت یادگیری خارج از مباحث درس شما کمی از مباحث اصلی درس خارج است. این مورد به عمد و به قصد تقویت «یادگیری به منظور استفاده‌ی آنی» وجود دارد و انتظار نداریم که کیفیت و سادگی انجام آن توسط شما مانند مباحث تمارین مرتبط درس باشد.
- همانطور که بیان شد، یکی از اهداف این پروژه تقویت مهارت یادگیری شما بسته به نیاز مسئله است؛ بنابراین جستجو در منابع مختلف و انجام مطالعات و دیدن مثال‌های مشابه بسیار مورد استقبال قرار خواهد گرفت. از شما می‌خواهیم تا **منابعی که مورد مطالعه قرار می‌دهید را در گزارش خود بیاورید**.
- در مورد زبان برنامه‌نویسی و ابزارهای مورد استفاده **هیچ محدودیتی** وجود ندارد. اگر چه استفاده از زبان پایتون توصیه می‌شود.
- پروژه به صورت گروهی و در قالب **گروه‌های دو نفره** است.
  - آپلود شدن پروژه توسط یکی از اعضا کافی است.
  - نام و شماره‌دانشجویی افراد را در گزارش ذکر کنید.
  - وظایفی که هر کدام از اعضای پروژه داشته‌اند را نیز در گزارش بیاورید.
- خروجی‌های مورد نیاز پروژه: ۱- کد ۲- مستند توضیح
  - استفاده از خروجی jupyter notebook به دلیل اینکه خروجی‌های مورد نیاز را به صورت یک‌پارچه قابل ارائه می‌کند توصیه می‌شود.
  - اگر از ژوپیتر استفاده می‌کنید می‌توانید مستندات و کد را در قالب همان یک فایل تحویل دهید.
- دقت کنید که تمامی خروجی‌های پروژه به عنوان دارایی معنوی[6] هر عضو گروه تلقی خواهد شد و اگر موردی نتیجه‌ی کار شما در پروژه نیست، لطفا آن را ذکر کنید.
- مشکلات و سوالاتتان را، از هر جنس، در **پیاتزای درس** بپرسید.

[6] Intellectual Property

۶- لینک‌های مفید

پیشنهاد می‌کنیم، پیش از شروع پروژه به لینک‌های زیر توجه کنید.

۱- مراحل انجام پروژه‌ی یادگیری ماشین

۲- چرا EDA کنیم؟

۳- چگونه EDA کنیم؟

۴- چگونه feature engineering کنیم؟

۵- اهمیت data storytelling

۶- در مورد چگونگی و چرایی تفسیرپذیری

۷- MLOps چیست؟

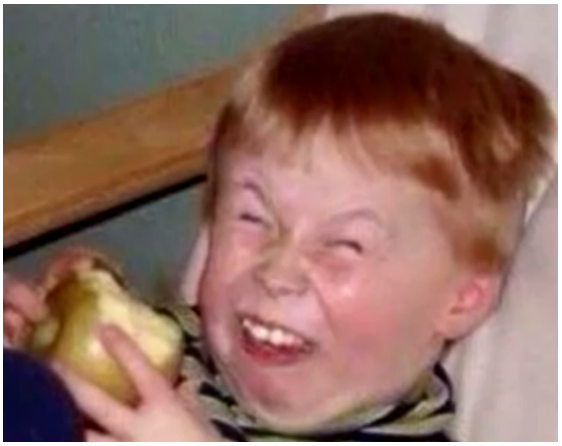
۸- در مورد MLOps و کاربرد آن

۹- مدل‌های پایه‌ی پیشنهادی:

۹-۱- Deep Factorizartion Machine

۹-۲- Wide and Deep NN's

۹-۳- مدل کلاسیک XGBoost



امیدواریم از این پروژه‌ی کوتاه و ساده لذت ببرید.

موفق باشید.