

پروژه

درس بزرگداده و سیاست گذاری

پیش از پاسخ سوالات به نکات زیر توجه کنید:

- پروژه درس به صورت انفرادی است و دارای دو بخش کلی کد نویسی و تحلیل می باشد.
- پاسخ شما باید شامل کدنویسی پایتون و یک گزارش از تحلیل کدنویسی خودتان و همچنین یک گزارش Academic با رعایت اصول گزارشنویسی که شامل تحلیل کد، مصورسازی دادهها، مدلها و هرگونه تحلیل دیگر میباشد.
- مدت زمان پاسخدهی حداکثر ۷ روز بوده و در صورت لزوم ۱ روز تمدید خواهد شد.

نحوه بارگذاری پاسخ پروژه

- در پاسخدهی به سوالات موجود هیچگونه محدودیتی وجود ندارد و میتواند از پکیج مناسب استفاده نمایید. (خارج درس کلاس نیز مشکلی نخواهد داشت)
- در صورت وجود هر گونه ابهام یا سوال از طریق گروه درس در پلتفرم تلگرام و یا از طریق ایمیل پیگیری نمایید. hamed_meraji@ie.sharif.edu
- برای تحویل فایل پاسخ خود لطفا تمامی فایلهایی که ممکن است در تصحیح برایتان موثر باشد را در یک فایل rar قرار دهید.
- فایل ارسالی حتما شامل نام، نام خانوادگی، شماره دانشجویی و همچنین نام فایل ارسالی به صورت: HW1_name_StudentNumber باشد.
- نمره تشویقی هر تمرین صرفا در همان تمرین لحاظ خواهد شد.

بررسی اجمالی:

Juko Analytics یک استارت‌آپ مشاوره تحلیلی خرده فروشی است که هدف آن کمک به زنجیره‌های خرده فروشی است تا استراتژی فروش و بازاریابی خود را با استفاده از بینش‌های مبتنی بر داده بهینه کنند.

قابلیت‌های کلیدی ارائه شده عبارتند از:

داشبورد تجزیه و تحلیل فروش: داشبوردهای تعاملی برای تجسم و نظارت بر داده‌های معاملات خرده فروشی در ابعاد مختلف مانند شعبه، ویژگی‌های مشتری، رشد فروش در طول زمان و غیره.

تجزیه و تحلیل پیشرفته: مدل‌سازی آماری و تجزیه و تحلیل what-if برای پیش‌بینی تقاضا، بهینه‌سازی قیمت‌گذاری، مدل‌سازی آمیخته بازاریابی بر اساس داده‌های معاملات گذشته.

تقسیم‌بندی و هدف‌گذاری: الگوهای تراکنش را تجزیه و تحلیل کنید تا بخش‌های مشتریان با ارزش بالا را شناسایی کرده و برای هدف قرار دهید.

مدیریت کمپین: کمپین‌های تبلیغاتی، طراحی کاتالوگ و اعلان‌های فشاری موبایل را با استفاده از داده‌ها برنامه‌ریزی کنید.

مزیت رقابتی:

این استارت‌آپ از یک تیم متخصص علم داده برای استخراج داده‌های تراکنش‌ها و ارائه تحلیل‌های عملی استفاده می‌کند که برنامه‌ریزی و تصمیم‌گیری‌های خرده‌فروشی را به‌طور قابل‌توجهی تضمین می‌کند. داشبوردهای سفارشی‌سازی شده پیشرفته به جای تکیه بر گزارش‌های ماهانه منسوخ شده، فروش ۷۸۲۴ را ارائه می‌دهند.

طرح‌کشی:

این استارت‌آپ مشاوره، زنجیره‌های خرده‌فروشی کوچک و متوسط را به عنوان مشتری در اختیار دارد. برای دسترسی به پیشنهادات تجزیه و تحلیل استاندارد و پروژه‌های مشاوره سفارشی، هزینه اشتراک می‌گیرد. از طریق یک مدل فریمیوم که تجزیه و تحلیل‌های اولیه را برای فروشگاه‌های منتخب ارائه می‌دهد، اعتبار ایجاد می‌کند.

پتانسیل رشد:

با ثبت نام مشتریان بیشتر، استارت‌آپ با عمودی کردن راه حل‌ها بر اساس دسته‌های عمده خرده‌فروشی گسترش می‌یابد تا بیشتر پیشنهادات خود را سفارشی کند. گسترش بین‌المللی همچنین یک مسیر رشد جذاب را ارائه می‌دهد. مدل‌های تحلیلی و ML با افزایش داده‌ها اصلاح‌تر می‌شوند.

در اینجا توضیح مختصری از هر یک از ویژگی ها یا برجسب ها در مجموعه داده آمده است:

شناسه فاکتور: یک شناسه منحصر به فرد برای هر فاکتور یا تراکنش.

شعبه: شعبه یا مکانی که معامله انجام شده است.

شهر: شهری که شعبه در آن قرار دارد.

نوع مشتری: نشان می دهد که مشتری مشتری دائمی است یا جدید.

جنسیت: جنسیت مشتری.

خط تولید: دسته یا نوع محصول خریداری شده.

قیمت واحد: قیمت یک واحد از محصول.

تعداد: تعداد واحدهای محصول خریداری شده.

مالیات ۵٪: مبلغ مالیات (۵٪ از کل هزینه) اعمال شده برای معامله.

مجموع: کل هزینه معامله با احتساب مالیات.

تاریخ: تاریخی که معامله انجام شده است.

زمان: زمانی از روز که معامله انجام شده است.

پرداخت: روش پرداخت مورد استفاده (به عنوان مثال، کارت اعتباری، پول نقد).

COGS (هزینه کالاهای فروخته شده): هزینه های مستقیم مربوط به تولید یا خرید محصولات فروخته شده.

درصد حاشیه ناخالص: درصد حاشیه سود برای معامله.

درآمد ناخالص: کل سود حاصل از معامله.

رتبه بندی: رتبه بندی رضایت مشتری یا بازخورد در مورد معامله.

- ساختار داده (۱۵ نمره)

- ۱- انواع هر یک از داده‌های موجود را بیان کنید و برای هر دلیل خود را ذکر کنید. (۱۰)
- ۲- آیا مجموعه داده‌ای که در اختیار دارید، نیاز به پیش پردازش دارد؟ چرا؟ (۵)

- مصورسازی (۱۸۰ نمره)

- ۱- برای فاکتورهایی که ممکن است بتوانید توزیعی را بیان کنید یا ترسیم کنید، ترسیم کرده و به صورت کامل تشریح کنید. (۲۰)
- ۲- نمودار مناسبی ایجاد کنید که درصد تراکنش‌ها را بر اساس نوع پرداخت (نقد، کارت اعتباری، Ewallet) نشان دهد. (۵)
- ۳- یک نمودار مناسب ترسیم کنید که کل مقادیر فروخته شده برای هر خط محصول را نشان می‌دهد، آن را تحلیل کنید. (۵)
- ۴- یک هیستوگرام از قیمت‌های واحد برای تجزیه و تحلیل توزیع قیمت ایجاد کنید و نسبت به کدام فاکتور دیگر می‌توان قیمت را دسته بندی کنید. (۱۰)
- ۵- میانگین کل مبلغ فروش در هر ماه را در دوره ۳ ماهه محاسبه و ترسیم کنید. (۵)
- ۶- فروش بر اساس خط محصول را در طول زمان با استفاده از نموداری مناسب مصور کنید. (۵)
- ۷- نمودار مناسب گروه بندی شده ایجاد کنید تا میانگین کل فروش را بر اساس شهر و نوع مشتری مقایسه کند و تحلیل خود را بیان کنید. (۱۰)
- ۸- نمودار پراکندگی را با رتبه بندی در مقابل مقدار کل فروش ترسیم کنید تا همبستگی‌ها را بررسی کنید. (۱۵)
- ۹- نمودارهای جعبه‌ای ایجاد کنید تا به صورت بصری بررسی کنید که آیا تفاوت قیمت واحد بر اساس خط محصول وجود دارد یا خیر. (۱۰)
- ۱۰- گروه بندی بر اساس شاخه و ایجاد نمودارهای دایره‌ای برای مقایسه درصد بازدیدهای مشتریان جدید در مقابل اعضای عضو. (۱۰)
- ۱۱- برای مصورسازی مقادیر فروش بر اساس خط محصول در سراسر شاخه، یک نمودار مناسب ایجاد کنید. (۵)
- ۱۲- یک heatmap بصری ایجاد کنید که تغییرات مقدار کل فروش را در دوره ۳ ماهه نشان می‌دهد. (۱۰)
- ۱۳- کل فروش را در طول زمان، به تفکیک شاخه و با رنگی که نشان دهنده نوع پرداخت است، ترسیم کنید. (۱۰)
- ۱۴- برای مطالعه فصلی بودن روزانه و ماهانه تعداد معاملات، نمودار مناسبی تهیه کنید. (۵)
- ۱۵- از نمودارهای رگرسیون برای مطالعه همبستگی رتبه با مقدار کل فروش و درصد حاشیه ناخالص استفاده کنید. (۱۰)
- ۱۶- نمودار متناسبی ایجاد کنید تا میانگین فروش را بر اساس ساعت روز و روز هفته تجسم کنید. (۱۵)
- ۱۷- ترسیم نموداری دلخواه و متناسب برای به تصویر کشیدن روابط بین ویژگی‌های محصول مانند قیمت، حاشیه، کمیت و غیره. (۵ + ۱۰)
- ۱۸- از قابلیت‌های ترسیم pandas برای ایجاد برنامه نویسی مضرب کوچک برای نمایش زیرمجموعه‌های داده استفاده کنید. (۵)

- ۱۹- با توجه به ۱۸ بخش گذشته تحلیل شما و برداشت‌های شما در رابطه با این مجموعه داده چیست؟ آیا نیاز به جمع‌آوری داده‌های دیگری هست یا خیر؟ در صورت بله بودن پاسخ علت خود را بیان کنید؟ (۱۵)
- ۲۰- با توجه به داده‌ها آیا روزهای تعطیل قابل شناسایی است و آیا میزان خرید و فروش در روزهای تعطیل توزیع خاصی دارد؟ (**)

- مدل سازی (۳۰۰ نمره)

- ۱- برای مجموعه داده شده مدل KNN پیاده سازی کنید و آیا می‌توان K بهینه را بدست آورد؟ (۴۰)
- ۲- برای مجموعه گفته شده یک مدل درخت تصمیم، جنگل تصادفی و مدل رگرسیون خطی (در صورت امکان) پیاده سازی کنید. (۱۲۰)
- ۳- کدام یک از مدل‌های فوق عملکرد بهتری را دارد؟ کدام معیار را بررسی کردید؟ تمامی معیارهای ممکن برای ارزیابی هر یک از مدل‌ها را بیان کنید؟ (۵۰)
- ۴- در رابطه با روش‌هایی همانند رگرسیون لسو و ریج و پیاده سازی روی این مجموعه داده تحقیق کنید؟ (۲۰)
- ۵- در رابطه با مفاهیم feature selection در این مطالعه موردی تحقیق کنید و روش‌هایی را بیان کنید. (۲۰)
- ۶- در این مجموعه داده روش Apriori Method قابل بررسی است؟ اگر بله چرا؟ اگر خیر به چه داده‌های نیاز دارید؟ (۳۰)
- ۷- اگر بخواهیم از روش‌های خوشه‌بندی استفاده کنیم کدام یک از روش‌های خوشه‌بندی را استفاده می‌کنید؟ چرا؟ (۲۰) (**پیاده سازی)**

- مفهوم، تجزیه و تحلیل (۱۲۵ نمره)

- در هریک از سوالات زیر صرفاً تجزیه و تحلیل تشریحی خود را بیان کنید و نیازی به کد نویسی نیست.
- یادگیری عمیق: (هر سوال ۵ نمره)
- ۱- یک مدل RNN برای پیش‌بینی فروش مورد انتظار برای ماه آینده بر اساس داده‌های سری زمانی فروش قبلی آموزش دهید.
 - ۲- یک مدل طبقه‌بندی تصویر CNN بسازید تا به طور خودکار محصولات را از روی تصاویر آنها دسته‌بندی کند.
 - ۳- از تکنیک‌های تشخیص ناهنجاری مانند رمزگذارهای خودکار برای شناسایی تراکشن‌های غیرعادی یا تقلبی استفاده کنید.
 - ۴- از یادگیری تقویتی برای بهینه‌سازی پویا سیاست قیمت‌گذاری برای به حداکثر رساندن درآمد استفاده کنید.
 - ۵- یک مدل شبکه عصبی نموداری را برای پیش‌بینی احتمال خرید بسته‌های مختلف محصول آموزش دهید.

○ متن کاوی: (هر سوال ۵ نمره)

- ۱- بازخورد و نظرات مکتوب مشتری را برای شناسایی شکایات و پیشنهادات رایج تجزیه و تحلیل کنید.
- ۲- از تجزیه و تحلیل احساسات در ایمیل های مشتری در مورد مسائل و تجربه استفاده کنید.
- ۳- یک مدل موضوعی با استفاده از LDA در مورد درخواست های مشتری بسازید تا بر اساس انواع پرس و جو دسته بندی کنید.
- ۴- پیاده سازی تجزیه و تحلیل مبتنی بر جستجو برای استخراج اطلاعات کلیدی از فیلدهای متنی رایگان فاکتورهای خرده فروشی.
- ۵- از پردازش زبان طبیعی برای چت ربات هوشمند برای کمک به سوالات مشتریان استفاده کنید.

○ تجزیه و تحلیل (هر سوال ۵ نمره)

- ۱- سودآورترین بخش های مشتری را از الگوهای معاملات تاریخی شناسایی کنید.
- ۲- ارزیابی عملکرد در شعب خرده فروشی و پیشنهاد بهبود برای شعب فقیرتر.
- ۳- بهینه سازی نیازهای کارکنان در ساعات اوج فروشگاه بر اساس تجزیه و تحلیل زمان معامله.
- ۴- تنظیمات قیمت را برای محصولات مختلف بر اساس کشش تقاضا توصیه کنید.
- ۵- پیشنهاد مکان برای افتتاح فروشگاه های جدید با استفاده از تجزیه و تحلیل داده های جغرافیایی و جمعیتی.
- ۶- ارزیابی عملکرد کمپین های بازاریابی گذشته با تجزیه و تحلیل افزایش فروش ایجاد شده.
- ۷- ترافیک وب سایت و نیازهای زیرساخت ابری را بر اساس رشد پیش بینی شده فروش پیش بینی کنید.
- ۸- مدل های قیمت گذاری را ایجاد کنید تا به طور موثر پیشنهادات مشاوره و اشتراک را قیمت گذاری کنید.
- ۹- فرصت های فروش متقابل را بر اساس الگوهای رایج در بسته های محصول خریداری شده برجسته کنید.
- ۱۰- معیارهای عملکرد مانند حاشیه درصد در مقابل زنجیره های خرده فروشی رقیب را معیار قرار دهید.
- ۱۱- تجزیه و تحلیل هزینه های حمل و نقل بر اساس مکان برای شناسایی فرصت ها برای بهینه سازی لجستیک.
- ۱۲- ردیابی تأثیر عوامل خارجی مانند تعطیلات و آب و هوا بر رفتار خرید.
- ۱۳- برنامه استخدام را با مهارت های کلیدی مورد نیاز بر اساس استراتژی های رشد اطلاع دهید.
- ۱۴- ارزیابی امکان سنجی گسترش بین المللی با استفاده از داده های خرده فروشی جهانی.
- ۱۵- صرفه جویی در هزینه های بالقوه از معرفی اتوماسیون و هوش مصنوعی را کمی کنید.

(**): سوالات تشویقی و شامل نمره مازاد

اگر در سوال دو عدد جمع شده نمره اول، نمره خود سوال و نمره بعدی نمره تشویقی است.