گزارش پروژه ۴

حسام رمضانیان

۸۱۰۱۰۰۲۴۸

سوال۱)

درنوت بوک موجود است

سوال۲)

درنوت بوک موجود است

سوال۳)

باتوجه به نمودار هبستگی رسم شده در نوت بوک

MntCoffee,MntMeatProducts,Incomeبیشترین همبستگی را دارند

سوال۴)

در نوت بوک موجود است

سوال۵)

نمودار ها در نوت بوک موجود است

سوال۶)

فیچر هایی که مقادیر نامعتبر داشتند را با حذف ردیف ها اصلاح کردم و داده های پرت را از مجموعه داده ها حذف کردم در نوت بوک قابل مشاهده است

سوال ۷)

یکی از روش های مورد استفاده برای پر کردن داده های miss استفاده از یک مقدار ثابت میباشد که سریع و راحت است اما به علت اینکه این داده ها یکسان هستند میتواند یک bias بر روی داده ها ایجاد کند

روش دیگر استفاده از از مقادیر ردیف های قبل و بعد است که با توجه به نوع داده های ما که وابستگی خاصی نسبت به هم ندارند این روش مناسب نیست

یک روش استفاده از linear interpolation است که بهتر از روش قبل عمل می‌کند اما مشابه روش قبل در صورتی که داده های مجاور وابستگی نداشته باشند این روش به درستی عمل نمیکند

سوال۸)

Income = 223

Mntcoffee = 205

NumWebVisitsMonth = 200

MntGoldProds = 13

در نوت بوک miss ها برطرف شده است

سوال۹)

Normalize : این روش مقادیر را به مقیاس ۰ تا ۱ در می‌اورد که باعث میشود مقیاس ها به یک فرمت باشند تا در زمان train برای مدل اشکال ایجاد نشود

$$Norm(X) = (X - min(X)) / (max(X) - min(X))$$

Standardizing : این روش مقیاس مقادیر را به گونه ای تغییر میدهد که میانگین ۰ و واریانس ۱ شود

$$Standard(X) = (X - mean(X)) / (Std(X))$$

علت استفاده از این روش ها این است که مدل های هوش مصنوعی برای train شدن و به دست اوردن نتیاج بهتر نیاز دارند تا داده ها در یک مقیاس یکسان باشد تا روابط میان داده ها را به خوبی پیدا کنند

سوال ۱۰)

به طور معمول از دو روش برای تبدیل این داده ها استفاده میشود

1) برای هر مقدار در فیچر یک فیچر جدید اضافه کنیم که مقادیر باینری ۰ و ۱ را میپذیرد

2) آن ها را براساس یک ترتیب خاص متناظر با مقادیر عددی حسابی قرار دهیم

در صورتی که همه مقادیر موجود در فیچر معنی دار باشند و قصد حذف کردن فیچر را نداشته باشیم برای تمامی مقادیر این پیش پردازش ضروری است

سوال ۱۱)

بله ستون هایی که همبستگی پایینی با ستون هدف دارند را میتوان از داده ها حذف کرد زیرا داده های مفیدی برای train کردن مدل نیستند

سوال۱۲)

به طور معمول نسبت داده ها برای train بین ۷۰ تا ۸۰ درصد میباشد و سایر داده ها به منظور test مورد استفاده قرار میگیرند

یکی از روش ها به این صورت است که داده ها را به صورت رندوم بین testو train پخش کنیم

روش دیگر به طور معمول برای داده هایی که بر اساس زمان مرتب شدن استفاده میشود به این صورت که بخش های ابتدایی را برای train و بخش های انتهایی را برای تست استفاده میکنیم

سوال۱۳)

این دسته از داده ها برای ارزیابی هایپر پارامتر های مدل استفاده میشود که ایا مدل نسبت به قبل بهبود یافته یا خیر ، و بررسی اینکه ایا نیاز به توقف زود هنگام داریم یا نه باتوجه با اینکه این داده ها متفاوت از داده های testو train است میتوان به خوبی با استفاده از ان ها این پیشرفت را بررسی کرد

سوال۱۴)

در این روش داده ها را به k قسمت یا fold تقسیم میکنیم سپس k تکرار بر روی داده ها انجام میدهم که در هر مرحله یک بخش به عنوان validation و سایر بخش ها را به عنوان train استفاده میکنیم این روش معمولا برای دیتاست های کوچک استفاده میشود دقت شود که در هر تکرار یک بخش جدید به عنوان validation مورد استفاده قرار میگیرد پس از پایان این k تکرار میانگین ارزیابی این تکرار ها به عنوان نتیجه نهایی مورد استفاده قرار میگیرد

سوال۱۵)در این بخش میخواهیم یک معادله خط پیدا کنیم که متغیر مستقل ان یکی از فیچر ها و متغیر وابسته داده های ستون هدف باشد که بیشترن شباهت را با داده های ما داشته باشد به این منظور میخواهیم میزان خطا را کاهش دهیم پس باید معادله از RSS مشتق گرفته و با دو معادله ۲ مجهول به دست امده ضرایب بهینه را به دست اوریم

سوال۱۶)

MntCooffe بهترین فیچر است زیرا بیشترین همبستگی را با ستون هدف دارد.(در جدول رسم شده در نوت بوک قابل مشاهده است)

سوال۱۷)

RSS: مربع خطا پیشبینی مدل نسبت به مقدار واقعی در ستون هدف میباشد که معیاری است که سعی در بهینه کردن ان داشتیم

$$RSS = \Sigma(y - \hat{y})2$$

MSE: میانگین مقدار RSSمیباشد (n : تعداد نمونه ها)

$$MSE = (\Sigma(y - \hat{y})2) / n$$

RMSE : ریشه دوم MSE است که مقادیر کمتر بهترند و ۰ نشان‌دهنده عدم وجود خطاست.

$$RMSE = (\Sigma(y - \hat{y})2 / n)^(1/2)$$

R2 Score : برخلاف RMSE که در مقادیر مثبت مرزی ندارد R2 score دارای مرز در مقادیر مثبت میباشد و بزرگتر از ۱ نمیشود اما در مقادیر منفی مرزی ندارد.R2 یا ضریب تعیین با مقایسه مجموع مربعات اختلاف بین مقادیر واقعی و پیش‌بینی شده متغیر وابسته با مجموع کل مربعات اختلاف بین مقادیر واقعی و میانگین متغیر وابسته محاسبه می‌شود.

R2 = 1 - (SSres / SStot)

سوال۱۸)

RMSE در صورتی که مدل عملکرد بهتری داشته باشد به صفر نزدیکتر است و R2 به یک مقدار صفر R2یعنی پیشبینی تصادفی و مقادیر منفی به معنی پیشبینی بدتر از حالت تصادفی است برای ۴ فیچر که همبستگی بیشتری داشتند مدل را train کردم و مقدار این مقادیر در نوت بوک موجود است

سوال۱۹)

طبق نتایج به دست امده بیشترین دقت برای decisionTree پس از انLogstic Regression و کمترین دقت برای k-nearest- Neighbours نتایجدر نوت بوک موجود است

سوال۲۰)

در نوت بوک موجود است

سوال۲۱)

Overfitting : یعنی مدل بیش از حد به داده‌های آموزشی وابسته شده و نمی‌تواند روی داده‌های جدید عملکرد خوبی داشته باشد. این معمولا زمانی رخ می‌دهد که مدل خیلی پیچیده باشد و سعی کند نویزها و جزئیات غیرضروری داده‌های آموزشی را یاد بگیرد.

Underfitting : یعنی مدل نتوانسته روابط و الگوهای مهم در داده‌ها را یاد بگیرد. این معمولا زمانی اتفاق می‌افتد که مدل خیلی ساده باشد و انعطاف‌پذیری کافی برای مدل‌سازی داده‌ها را نداشته باشد.

سوال۲۲)

با حذف مرحله نرمالایز کردن از پیش پردازش نتیجه نهایی تمامی مدل ها دچار اندکی افت میشود

با حذف مرحله حذف داده های پرت نیز مدل نتیجه بدتری داشت زیرا این داده ها در مرحله train قدرت پیشبینی مدل را کاهش میدهند

با حذف مرحله حذف داده های بی معنی مثل kidhome با مقدار منفی نیز درصورتی که این داده ها در بخش test قرار بگیرند به شدت کاهش میابد

سوال۲۳)

در نوت بوک رسم شده است

سوال۲۴)

n_estimators : تعداد درخت هایی است که مدل train را با آن ها انجام میدهد که باعث طولانی شدن فرایند trainمیشود اما واریانس را کاهش میدهد

max_depth :عمق درخت هایی که ایجاد میشوند را مشخص میکند که با مقدار دهی مناسب ان میتوان از overfitting جلوگیری میکند

با توجه به نمودار های رسم شده در نوت بوک مقدار n_estimators تا حدود ۱۰۰-۲۰۰ باعث بحبود دقت مدل میشود اما پس از ان overfittingرخ میدهد

همچنین max_depth تا حدود عدد ۵-۱۰ باعث بحبود دقت مدل میشود اما پس از ان overfittingرخ میدهد

سوال۲۵)

bias :بررسی میکند که یک مدل چقدر به الگویی واقعی موجود در دیتا ها نزدیک شده است مقدار زیاد ان به معنی ساده سازی بیش از حد یا underfitting است

Variance : به معنی این است که مدل تا چه میزان به داده های train حساس شده است درواقع بالا بودن این مقدار به معنی overfitting میباشد

مدل Decision tree واریانس بالایی دارند و به راحتی دچار overfitting میشود در صورتی که در random forst با ایجاد درخت های متعدد سعی میشود تا بالانس خوبی میان بایاس و واریانس ایجاد شود

سوال۲۶)

اضافه کردن مقادیر کوچکی از نویز تصادفی به هر داده‌ی خاص می‌تواند شناسایی افراد را از روی داده ها دشوارتر کند، اما همچنان امکان train کردن مدل ها بر اساس این مجموعه‌ی داده‌ها ممکن است.

سوال۲۷)

نوییز لاپلاس از توزیع لاپلاس جهت ایجاد نویز استفاده میکند که باعث میشود داده ها بیشتر تغییر کنند و حریم خصوصی بیشتر حفظ شود اما باعث میشود که داده دچار اشکال شوند در مقابل نویز نمایی از توزیع نمایی استفاده میکند که یعنی حریم خصوصی کمتر حفظ میشود اما داده ها سالم تر هستند

سوال۲۸)

از نویز لاپلاس استفاده شد که باعث مقداری کاهش در دقت مدل ها شد مقادیر در نوت بوک موجود است

سوال۲۹)

این یک روش یادگیری مجموعه‌ای است که پیش‌بینی‌های چندین مدل ضعیف را ترکیب می‌کند تا یک پیش‌بینی قوی‌تر تولید کند.

به این صورت عمل میکند که ابتدا یک مدل با یکی از الگوریتم ها مثل decisionTree میسازیم و اموزش میدهیم که به ان weak learner گفته میشود سپس خطای مدل را روی ٔاده های train محاسبه میکنیم سپس یک weak learner جدید ایجاد میکنیم تا این خطا را پیشبینی کند و خطا مدل را بهبود ببخشد سپس مدل به مجموعه مدل ها اضافه میشود و نتیجه ان توسط میانگیری وزنی ترکیب میشود این کار را تا زمانی که به یک تعداد تکرار مشخص یا کم تر شدن تابع زیان از یک مقدار خاص ادامه میدهیم مدل نهایی میانگین وزنی تمام این weak learner ها را برمیگرداند

۳۰سوال)

XGBoost مانند بوستینگ گرادیان از مجموعه‌ای از درختان تصمیم به عنوان مدل‌های پایه استفاده می‌کند. هر درخت بر روی باقی‌مانده‌ها یا خطاهای درخت‌های قبلی آموزش داده می‌شود. بنابراین اولین درخت بر روی داده‌های آموزش اصلی آموزش داده می‌شود. دومین درخت بر روی باقی‌مانده‌های درخت اول آموزش داده می‌شود و به همین ترتیب تا زمانی که متوقف شود ادامه دارد.هنگام آموزش یک درخت، XGBoost از الگوریتم greedyبرای تقسیم داده‌ها در هر گره استفاده می‌کند و بهترین تقسیم که باعث بهبود حداکثری تابع هدف می‌شود را پیدا می‌کند. معیارهای متداول تقسیم شامل کاهش واریانس برای رگرسیون و افزایش اطلاعات برای طبقه‌بندی هستند.هر درخت از اشتباهات درخت‌های قبلاً آموزش داده شده می‌آموزد. پیش‌بینی‌های تمام درخت‌ها از طریق میانگین‌گیری وزنی با هم ترکیب می‌شوند تا پیش‌بینی نهایی انجام شود. بنابراین هر درخت متوالی سعی می‌کند درخت‌های قبلی را اصلاح و بهبود دهد. هرچه تعداد درخت‌ها بیشتر باشد، مدل دقیق‌تر می‌شود. XGBoost همچنین دارای پارامترهای منظم‌سازی است تا اورفیت شدن را با افزایش تعداد درخت‌ها کنترل کند.

سوال۳۱) در نوت بوک موجود است