گزارش کار پروژه درس داده کاوی - دکتر محمدپور

متین معزی - ۹۵۱۲۰۵۸

۲۴ خرداد ۱۳۹۹

سوال اول.

الف) انتخاب راه‌های مقابله با داده‌های گمشده (Missing Data) بسته به نوع داده‌ها، موضوع دیتاست و نظر تحلیل‌گر داده متفاوت است. به عبارت دیگر اینکه کدام روش بهترین روش است به محتوای دیتاست و موضوع آن و همچنین فرضیه‌های قبلی شخص تحلیل‌گر نسبت به داده وابسته است و نمی‌توان در مورد یک روش بطور قطعی نظر داد.
در این دیتاست در ستون‌های مساحت و جمعیت هر کدام یک داده و در ستون رشد جمعیت دو داده موجود نیستند به عبارت دیگر Missing Values می‌باشند. با توجه به موضوع دیتاست از بین روش‌های پیاده‌سازی شده روش k نزدیکترین همسایه بهترین روش است زیرا داده‌های جمع‌آوری شده برای ما با ارزش است لذا از دست دادن داده کار عاقلانه‌ای نیست همچنین پر کردن داده با یک عدد ثابت بدون توجه به سایر داده‌ها نیز کار منطقی نیست در روش KNN Imputaion بر اساس نزدیکترین همسایه‌های یک سطر، داده گمشده را جایگزین می‌کنیم در نتیجه روش KNN Imputation بهترین روش برای این دیتاست می‌باشد.

ب) در اکثر الگوریتم‌های یادگیری ماشین نیاز است تا بروی داده‌ها عملیات‌های ریاضی مانند عملیات‌های ماتریسی و جبری اعمال شوند که به وضوح انجام این کار روی داده‌های کیفی امکان پذیر نیست. یکی از انواع داده‌های کیفی که مرسوم است داده‌های دسته‌ای Categorical Data می‌باشد که مقدار این داده‌ها هر مقدار دلخواهی نیست بلکه محدوده‌ای از مقدارها را شامل می‌شود برای مثال در این دیتاست در ستون International Visitors مقادیر عضو مجموعه {A,B,C,D} می‌باشند همچنین کار کردن با داده‌های عددی نسبت به داده‌های کیفی قابل فهم‌تر و راحت تر است.
یکی از روش‌هایی که در آن دوری یا نزدیکی دو سطر یا جمع و تفریق دو سطر به خوبی انحام می‌شود ، روش One Hot Encoding می‌باشد که به تعداد دسته‌ها (یکی کمتر) ستون اضافه کرده و درآن مشخص می‌کنیم که هر سطر عضو کدام دسته است.

ج) مقیاس ستون‌های مساحت، جمعیت، رشد جمعیت و تعداد مبتلایان ویروس به ترتیب تقریبا برابر $10^6, 10^8, 10^0, 10^4$ می‌باشد که تفاوت زیادی با یکدیگر دارند. این تفاوت زیاد در مقیاس ویژگی‌های عددی باعث پایین آمدن کارایی الگوریتم‌های یادگیری ماشین و نادقیق شدن تحلیل‌ها و تفسیرهای نتیجه ارزیابی‌ها می‌شوند بنابراین باید روشی برای نزدیک کردن مقیاس‌های ویژگی‌ها انجام دهیم.
روش استاندارسازی Standardization برای ویژگی‌هایی مفید است که از توزیع نرمال پیروی کنند با توجه به نمودارها این روش در این دیتاست کاربردی ندارد. روش یکه‌سازی Normalization حساسیت داده‌ها را کم می‌کند به عبارت دیگر تمام سطرها با دقت زیادی نزدیک به یک می‌شوند که این امر باعث عدم تمایز بین سطرها می‌شود. درنتیجه روش Min\Max Scaler از روش‌های دیگر بهتر عمل می‌کند.

د) به این سوال نمی‌توان جواب قطعی داد. با توجه به نظر تحلیل‌گر داده و هدفش از تحلیل و همچنین دلیل دور افتادگی داده می‌توان آن را حذف کرد یا با روش دیگری آن را تحلیل کرد. ممکن است دور افتادگی داده از سایر داده‌ها ناشی از خطای مشاهده یا خطای جمع‌آوری داده باشد یا دور افتادگی طبیعی باشد یعنی مقادیر داده از هم فاصله زیادی داشته باشند.
از راه‌های جاگزین حذف کردن می‌توان به علامت دار کردن و یا ایجاد مقیاس جدید جهت نزدیک کردن داده‌ها اشاره کرد.

ه) ضرایب مدل خطی برای ستون‌های نرخ رشد (Population growth)، جمعیت (Total Population)، مساحت (.Area (sq (km) به ترتیب برابر $0.0026, -2.085 \times 10^{-5}, -9099.4398$ و ضریب ثابت برابر 6.232×10^4 می‌باشد.
مقدار زیاد RMSE نشان می‌دهد اختلاف زیادی بین مقدار واقعی و مقدار پیشبینی شده مدل وجود دارد.
در رگرسیون چند متغیره برای ارزیابی مدل به جای R-Squared از Adj. R-Squared استفاده می‌کنیم زیرا Adj. R-Squared انحراف متغیر پاسخ نسبت به ویژگی‌هایی را نشان می‌دهد که باعث بهتر شدن مدل می‌شود در واقع اگر متغیری اضافه کنیم که مدل را بهتر کند این مقدار اضافه می‌شود به عبارت دیگر کاهش Adj. R-Squared جریمه اضافه کردن متغیر جدید که مدل را بهتر نمی‌کند می‌باشد. ولی برخلاف Adj. R-Squared به ازای اضافه شدن ویژگی جدید مقدار R-Squared اضافه می‌شود. همچنین مقدار .Adj

R-Squared نشان می‌دهد مدل خطی مناسب برای برازش ویژگی‌ها به متغیر پاسخ نمی‌باشد.
با فرض $\alpha = 0.1$ و با توجه به p-value های ضرایب متغیرها فرض صفر رد می‌شود و نتیجه می‌گیریم فرض خطی بودن ویژگی‌ها نسبت به متغیر پاسخ نادرست است و رد می‌شود.

و)

$$a = -4.271 \times 10^{-13}$$

$$b = 0.0006$$

$$c = 5.103 \times 10^{-12}$$

سوال دوم.

الف) با توجه به نمودار pairplot در می‌یابیم متغیرهای طول عضویت (Length of Membership) و مقدار زمان سالانه (Yearly Amount Spent) با تقریب خوبی نسبت به هم خطی می‌باشند.
از روی نمودار heatmap مشخص است که متغیرهای زمان اپلیکیشن (Time on App) و مقدار زمان سالانه (Yearly Amount Spent) به یکدیگر وابسته‌اند اما نمی‌توان گفت که رابطه خطی رابطه خوبی برای این همبستگی می‌باشد.

ب) نتیجه برازش خط به صورت زیر می‌باشد:

	Coeffient
Avg. Session Length	25.9815
Time on App	38.5902
Time on Website	0.1904
Length of Membership	61.2791
Intercept: -1047.9328	

MSE و RMSE داده‌های آموزش به ترتیب برابر 106.85 و 10.33 می‌باشد که خطای کمی را نشان می‌دهد.
ابتدا فرض خطی بودن متغیر پاسخ نسبت به متغیرهای مستقل را بررسی می‌کنیم. با فرض $\alpha = 0.01$ و مشاهده p-value ویژگی‌ها در می‌یابیم که فرض H_0 برای تمام ویژگی‌ها به غیر از Time on Website رد می‌شود به عبارت دیگر فرض صفر بودن ضرایب ویژگی‌ها (به غیر از ویژگی ذکر شده) رد می‌شود. در نتیجه متغیر پاسخ نسبت به ویژگی‌هایی که p-value آن‌ها صفر است خطی می‌باشد.
معیار R-Squared میزان مناسب بودن خط برازش شده نسبت به داده‌ها را مشخص می‌کند و هرچه این عدد به یک نزدیکتر باشد نشانگر مناسب بودن خط برازش برای داده‌ها می‌باشد. همانطور که در سوال اول اشاره شد در رگرسیون چند متغیره از Adj. R-Sqaured استفاده می‌کنیم زیرا تنها تاثیر ویژگی‌هایی را بررسی می‌کند که خطی بودن نسبت به متغیر پاسخ را بهبود می‌بخشد بنابراین در اینجا مقدار Adj. R-Squared برابر 0.982 می‌باشد که اطمینان خوبی از مناسب بودن مدل به ما می‌دهد.
در ستون بازه اطمینان (ستون آخر) مشخص می‌شود که ویژگی با احتمال ۹۵٪ در چه بازه‌ای قرار می‌گیرد.

د) تخمین خطای تست که از K-fold Cross Validation به دست آمده از خطای داده‌های تست بیشتر است لذا با اطمینان Overfitting رخ نداده است. همچنین اختلاف خطای مدل و خطای تست زیاد نیست بنابراین می‌توانیم بگوییم Underfitting نیز رخ نداده است.
با توجه به خطاهای به دست آمده مدل پیچیدگی زیادی ندارد لذا واریانس آن کم است ولی خطای آن باعث زیاد بودن بایاس شده است در نتیجه این مدل واریانس کم ولی بایاس زیاد دارد.
با توجه به مقادیر Adj. R-Squared و RMSE و همچنین اختلاف کم خطای مدل و خطای تست این مدل به خوبی عمل کرده و به داده‌ها فیت شده است.(البته به نظر اینجانب!!!)

ه) با توجه به ضریب متغیرهای Time on App و Time of Website مشاهده می‌شود که ضریب Time on App بسیار بیشتر از ضریب Time on Website است یعنی تاثیر زمان کار با اپلیکیشن بیشتر از تاثیر کار با وبسایت است. همچنین ازآنجایی که p-value متغیر Time on Website بزرگتر از α است، فرض صفر بودن ضریب آن رد نمی‌شود به عبارت دیگر با متغیر پاسخ

(Yearly Amount Spent) رابطه‌ای ندارد. بنابراین شرکت با تاکید بر اپلیکیشن نتیجه بهتری خواهد گرفت.

سوال سوم.

ب) Accuracy Score نسبت تعداد کلاس بندی‌های درست به کل کلاس بندی‌ها (کل داده) می‌باشد بنابراین هر چه بیشتر به یک نزدیکتر باشد مدل بهتر عمل کرده است.

	Train	Test
$K = 1$	0.539	0.472
$K = 30$	0.541	0.551

Table 1: Accuracy Score

با توجه به مقادیر جدول هیچکدام دقت خوبی ندارند و طبق این معیار این مدل با پارامترهای داده شده خوب نیست.
Confusion Matrix برای هر کلاس تعداد آن داده‌هایی را که درست کلاس بندی کرده و آنهایی را که اشتباه کلاس بندی کرده نشان می‌دهد. با توجه به نمودار رنگی هر چه قطر نمودار کمرنگ تر باشد نشان می‌دهد مدل دقت خوبی در دسته بندی داشته است. در اینجا مدل در هر ۴ حالت فقط برای کلاس ۰۱ دقت خوبی داشته است و در سایر کلاس‌ها رنگ نمودار نزدیک مشکی است که به معنای عدم دسته بندی درست می‌باشد.
معیارهای Recall و Precision بسته به هزینه ناشی از دسته بندی اشتباه می‌تواند معیار خوب بودن مدل باشد. اگر برای هر جفت دسته P و N خطای ناشی از اشتباه تشخیص N به شرط P برای ما ارزش بیشتری داشته باشد از معیار Recall و برعکس آن از Precision استفاده می‌کنیم. معیار F1-Score از میانگین این دو معیار به دست می‌آید که هر چه مقدار آن بیشتر باشد مدل بهتر است.
در اینجا میانگین هر سه معیار ذکر شده در $K = 1$ هم در داده‌های تست و هم آموزش بیشتر از $K = 30$ می‌باشد لذا مدل با پارامتر $K = 1$ بهتر عمل می‌کند.
در نمودار ROC هر چه فاصله خم رسم شده از نقطه چین بیشتر باشد یا به عبارت دیگر مساحت زیر نمودار ROC بیشتر باشد آن مدل بهتر است. در اینجا میانگین مساحت‌های زیر نمودار به ازای تمام دسته‌ها برای $K = 1$ بیشتر از $K = 30$ است لذا مدل با $K = 1$ طبق معیار ROC بهتر عمل کرده است.
برای پیشبینی داده‌های آموزش و محاسبه خطا از k-fold Cross Validation استفاده کردم زیرا در این صورت مدل با داده‌های جدید و نه با داده‌های آموزش قبلی تمرین داده می‌شود.

ج) بهترین پارامترها برای مدل KNN $K = 2$ و فاصله منهتن به دست آمده است.
معیار Accuracy و میانگین F1-Score برای داده‌های آموزش بیشتر از داده‌های تست شده است که به نظر می‌آید Overfitting رخ داده است و اصطلاحا مدل داده‌های آموزش را حفظ کرده است. قطر اصلی نمودار رنگی Confusion Matrix برای داده‌های آموزش وضعیت بهتری نسبت به داده‌های تست دارند و در دسته‌های اول، دوم و دهم دسته بندی به با دقت بیشتری نسبت به سایر دسته‌ها انجام شده که دلیل آن بالانس نبودن دسته‌ها می‌باشد.
در نمودار های ROC مشاهده می‌شود که مساحت زیر نمودار برای دسته ۱۵ بیشتر از بقیه است. آن نمودارهایی که مقدار NaN دارند به دلیل وجود نداشتن آن دسته در داده‌های تست می‌باشد.