***توجه : نمونه کار این فایل تماما با نرم افزار SPSS قابل انجام است. در اینجا هدف از آورده شدن این فایل نشان دادن توانایی آماری و تسلط متقاضی به مباحث آماری می باشد***

در این پروژه فایلی در اختیار قرار گرفت که اطلاعات آن مربوط به هزار نفر از مشتریان یک بانک بود. در جدول ارائه شده 14 متغیر وجود داشت. از آنجا که فایل ارائه شده در فرمت اکسل بود، برای آنالیز در نرم­افزار R به فرمت .CSV تبدیل شد و در فایل پروژه R که پیش تر ایجاد کرده بودیم، آپلود شد.



* فایل اکسل در اختیار



* دستور بارگزاری فایل داده­ها در نرم­افزار R

همانطور که مشخص است، سه ستون اول مربوط به شماره ردیف، آیدی مشتریان و نام خانوادگی آن­هاست که ارزش آماری نداشتند و از آنالیز حذف شدند. سپس با استفاده از دستور na.omit ، داده­های missing حذف شدند تا مراحل مدلسازی آغاز شود. صورت مسئله از ما خواسته بود خروج مشتریان از بر اساس پارامترهای ارائه شده مدلسازی شود. بنابراین بایستی متغیر Exited را به عنوان متغیر وابسته و سایر متغیرها مستقل در نظر گرفته شوند و با ضریب تخمینی مدل متغیر Exited را ایجاد کنند. بنابراین، ابتدا با دستور set.seed(123)، توالی داده­های شبه تصادفی خاصی انتخاب شدند و سپس داده­ها به دو گروه تقسیم شدند: 80% از داده­ها به صورت تصادفی انتخاب شده و در گروه training قرار گرفتند تا از طریق آنها، یادگیری ماشین و training انجام شود. و مابقی داده­ها برای آزمودن یادگیری جدا شدند. در اینجا ما از معیار تخمین آکایک (AIC) برای تخمین مدل استفاده کردیم و انتخاب متغیر­ها برای مدل به صورت backward بود به این صورت که ابتدا تمام پارامتر­ها اضافه شدند و مرحله به مرحله تا رسیدن به AIC بهینه حذف شدند. دستور زیر این عمل را انجام داد:

model1 <- glm(Exited ~., data = training.data1, family = binomial) %>% stepAIC(trace = 1)

خروجی­های دستور بالا به شرح زیر بودند:

1. در مرحله آغاز، هر 10 پارامتر وارد مدل شدند و AIC اولیه ارائه شد:



1. در مرحله­ی بعدی متغیر HasCrCard از مدل حذف شد و AIC کاهش یافت



1. نهایتا در مدل نهایی متغیر EstimatedSalary حذف و AIC نهایی بدست آمد.



با استفاده از دستور summary خروجی نهایی مدل به شرح زیر بود :



متغیرهای کشور(جغرافیا)، سن، جنسیت و عضو فعال بودن بیشترین تاثیر را برروی خروج یا عدم خروج افراد از بانک دارند.

حال نوبت به بررسی مدل به روی 20% داده­ی باقی مانده برای آزمایش مدل می­رسد. در واقع در این مرحله مشخص می­شود آیا مدل بدست آمده پیش بینی کننده­ی داده­های باقی مانده هستند یا نه. با مقایسه احتمالاتی مدل پیش بینی شده با داده­های آزمون، مدل ارائه شده با شانس 80.95% به داده­های باقی­مانده منطبق شدند (یعنی مدل با شانس 8.95% پیش­بینی صحیحی از خروج یا عدم خروج می­کند). کدهای زیر جهت بررسی این موضوع به کار گرفته شدند:

probabilities1 <- model1 %>% predict(testing.data1, type = "response")

predicted.classes1 <- ifelse(probabilities1 > 0.5, 1, 0)

mean(predicted.classes1 == testing.data1$Exited)

در نهایت جهت ارائه، برای چهار پارامتر با کمترین p-value (کشور(جغرافیا)، سن، جنسیت و عضو فعال بودن)، نمودار میله­ای رسم شد. لازم به ذکر است کلیه نمودار­ها با استفاده از پکیج ggplot2 و در محیط R رسم شدند :

1. نمودار توصیف ارتباط خروج از بانک با جنسیت : مشخص است که مردان نسبت به زمان تمایل کمتری به خروج از بانک دارند.



1. نمودار توصیف ارتباط خروج از بانک با میزان فعالیت : همراهی معنی داری بین عضوفعال بودن و عدم خروج وجود دارد



1. نمودار توصیف ارتباط خروج از بانک با سن: میانگین سنی افرادی که بانک را ترک کرده­اند از افرادی که ترک نکرده اند بیشتر است. (مشتریان با سن بالاتر، با احتمال بیشتری بانک را ترک می­کنند)



1. نمودار توصیف ارتباط خروج از بانک با کشور: این نمودار مشخص می­کند فرانسوی و اسپانیایی بودن مشتری شانس عدم ترک بانک را افزایش می­دهد درحالیکه آلمانی بودن مشتری نمی­تواند بطور قطعی خروج/عدم خروج مشتری را نشان دهد.

