Data-Driven decision-making class - Spring 1401

Exercise #2

Managing staff

Mahsa Choopannezhad – student number: 98207477

Overview

تمرین دوم درس مدلسازی و تصمیم گیری داده محور، بیشتر تمرکز تمرین روی ارائه یک برنامه کاربردی براساس مدل های متفاوت پیش بینی یک دیتاست است.

Goals

باتوجه به مقاله داده شده به شرکت AdviseInvest در راستای بهبود بهره‌وری کارکنانش مشاوره دهید.

در سه قسمت اصلی این تمرین را پی می‌گیریم:

- Data visualization and descriptive analytics
- Predictive analytics
- Prescriptive analytics

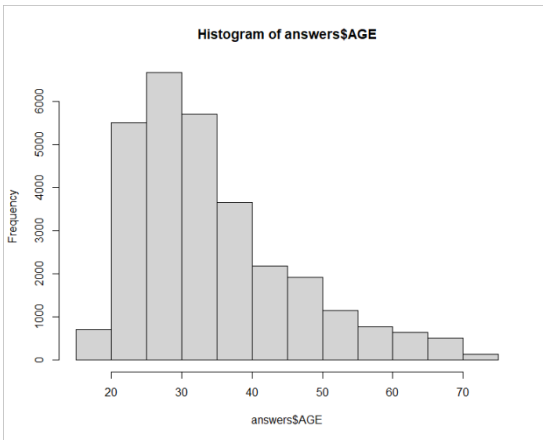
Introduction to the dataset:

قبل از اینکه سراغ کار با دیتاست برویم، مواردی که لازم است را فاکتور میکنیم و با چک کردن خلاصه متغیرهای دیتاست و چک و حذفNA های احتمالی و چک کردن ستون سن که نمونه غیرمنطقی نداشته باشیم.

Data visualization and descriptive analytics

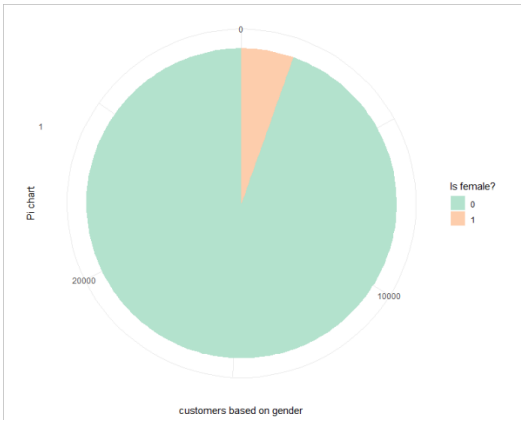
بعد از تمیزکردن دیتاست، در این بخش برای بدست آوردن دید اولیه از دیتا و متغیرهای موجود و ارتباط های احتمالی آنها نمودارهای متفاوت زیر را رسم میکنیم.

شکل۱ توزیع مشتری‌ها براساس سن آنها را نشان میدهد. مشاهده میشود که باتوجه به چولگی نمودار به سمت چپ، اکثر مشتریان این مرکز را مشتریان زیر ٤٠ سال و قشر جوان تشکیل میدهد.



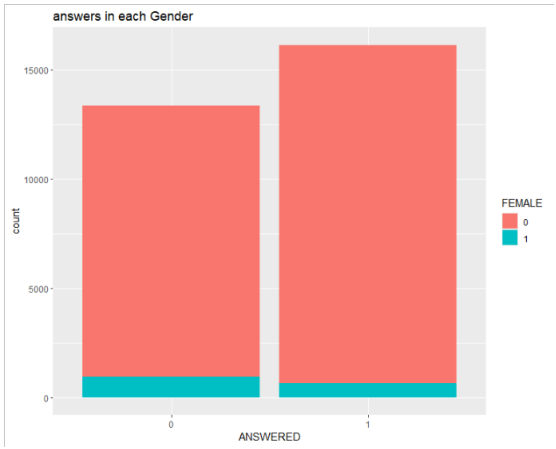
شکل ۱

مورد بعدی که از دیتای دموگرافیک پروفایل ها مورد بررسی قرار میدهیم بحث جنسیت مشتری ها است. شکل ۲ توزیع مشتریان براساس جنسیت را نشان میدهد. تقریبا چیزی حدود ۵ درصد مشتریان را خانم ها تشکیل میدهند.



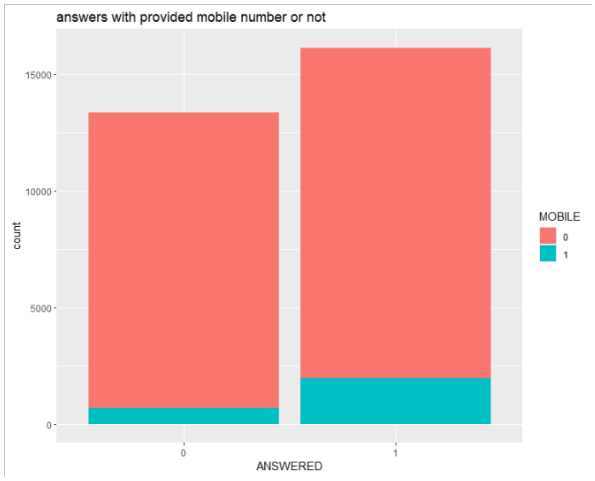
شکل ۲

شکل ۳ که رفتار مشتریان در پاسخ را براساس جنسیت نشان میدهد، بیان میکند که خانم ها رفتار متفاوتی در پاسخ به تلفن نداشته‌اند و این درحالی است که درصد بیشتری از اقایان به تلفن پاسخ داده اند.



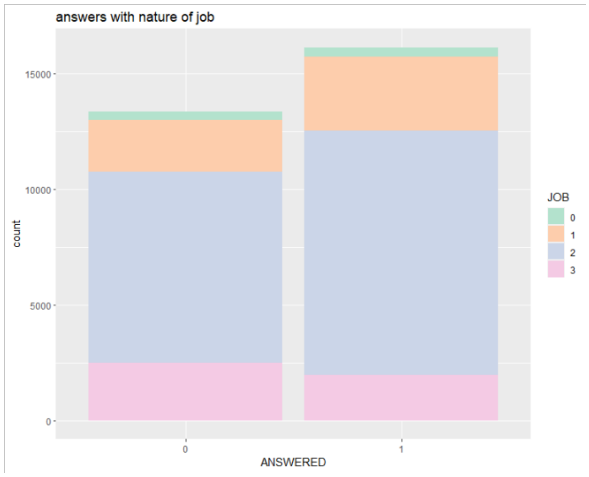
شکل ۳

شکل ٤ توزیع افراد را براساس پر کردن فیلد تلفن همراه و نرخ پاسخگویی آنها نشان میدهد. که میتوان برداشت کرد که افرادی که این فیلد را پر کرده اند تعداد پاسخ بیشتری هم به تلفن داشته اند. (لذا شاید یکی از اقدام های موثر میتواند اجباری کردن پر کردن این فیلد برای پروفایل باشد.)



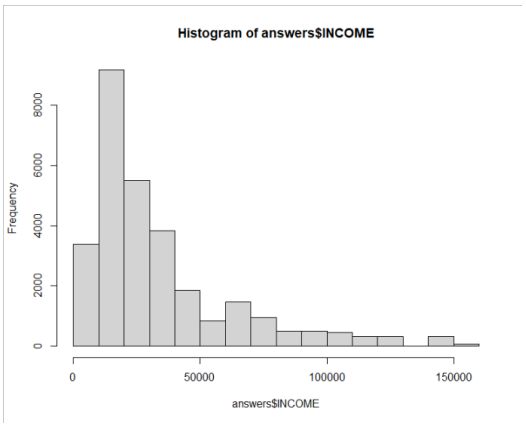
شکل ٤

شکل ۵ که رفتار پاسخ دهی را براساس نوع شغل افراد نشان میدهد که افراد با نوع شغل ۲ و ۳ (افرادی که میدلول به بالا هستند و شغلی بالاتر از متوسط و مدیریتی دارند) نرخ پاسخگویی بالاتری داشته اند.



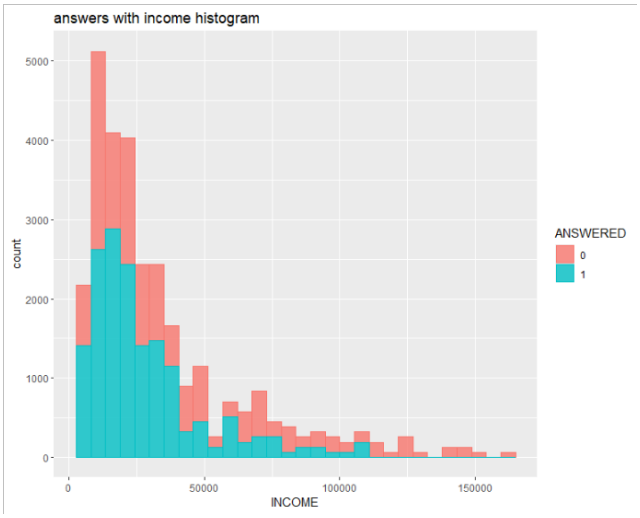
شکل ٥

شکل ٦ هم که توزیع درآمدی افراد را نشان میدهد، بیان میکند که باتوجه به چولگی به سمت چپ نمودار، بیشتر افراد و مشتریان را افرادی با درآمد زیر 50,000 تشکیل میدهد.



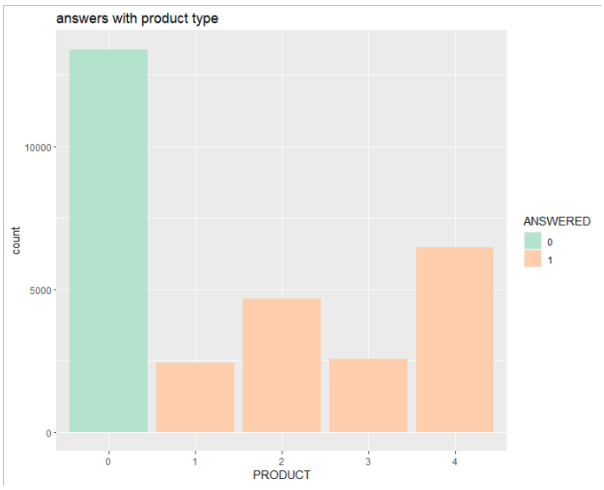
شکل ٦

شکل شماره ٧ دو هیستوگرام و توزیع افراد را براساس رفتار پاسخگویی آنها نشان میدهد و همانطور که مشخص است باتوجه به رفتار پاسخگویی افراد، توزیع درآمدی آنها عوض نشده است و همچنان قله نمودار و چولگی درهر توزیع یکسان است.



شکل ۷

شکل شماره ۸ که متغیر Product را براساس پاسخدهی آنها رسم کرده است، نشان میدهد که اگر افراد تلفن را جواب دهند تعداد افرادی که پلن پیشرفته و مبتدی را میخرند بیشتر است.



شکل ۸

Predictive analytics

باتوجه به نتایج بررسی در قسمت قبل و همبستگی کامل معنادار دو متغیر product و answered دو راهکار کلی داریم:

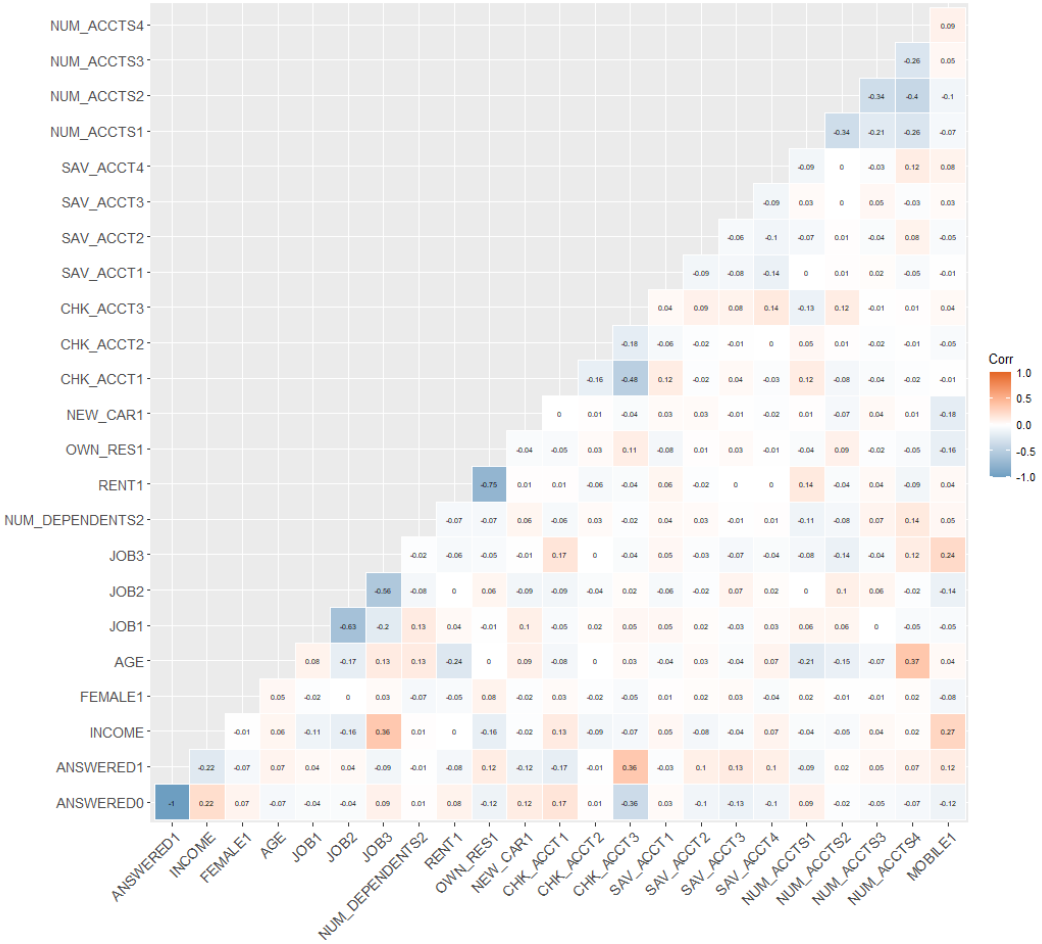
۱- متغیر answered را به عنوان متغیر هدف در نظر بگیریم و به عنوان یک متغیر باینری آن را پیش بینی کنیم در این صورت چون متغیر product همبستگی کاملی با متغیر پاسخ دارد آن را از دیتاست حذف میکنیم.

۲- متغیرproduct را به عنوان متغیر هدف در نظر بگیریم و با کلاس بندی مسئله را حل کنیم. در این حالت متغیر answered که همبستگی کامل با متغیر پاسخ دارد را از دیتاست حذف میکنیم.

در ادامه با مبنا قرار دادن راهکار اول با answers_excludeProduct کار میکنیم:

Correlation analysis

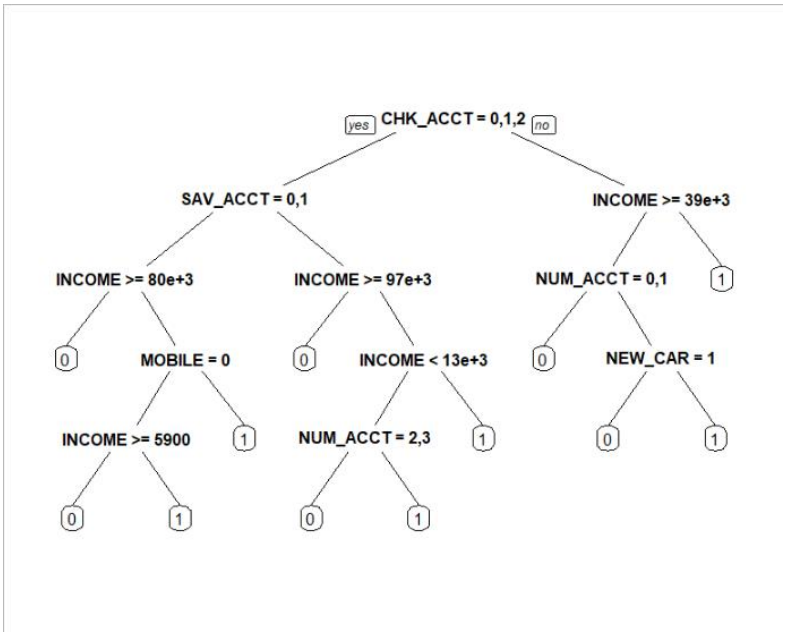
ماتریس کورلیشن را برای متغیرهای مختلف نسبت به یکدیگر میکشیم تا اگر چند متغیر خیلی بایکدیگر همبستگی شدید دارند آنها را از مدل حذف کنیم که با توجه به ماتریس رسم شده و هیت مپ قرار داده شده، متیغرهای خیلی همبسته شدید نداریم و مشکلی نداریم.



شکل ۹

CART model

با فراهم کردن دیتاست train و دیتاست test و باتوجه به حجم دیتاست جداکردن ۱۰ درصد دیتا به عنوان test مناسب است، متغیر پاسخ را متغیرansweredدر نظر میگریم و مدل درختی مناسب را رسم میکنیم، شکل ۱۰ درخت حاصل را نشان میدهد.



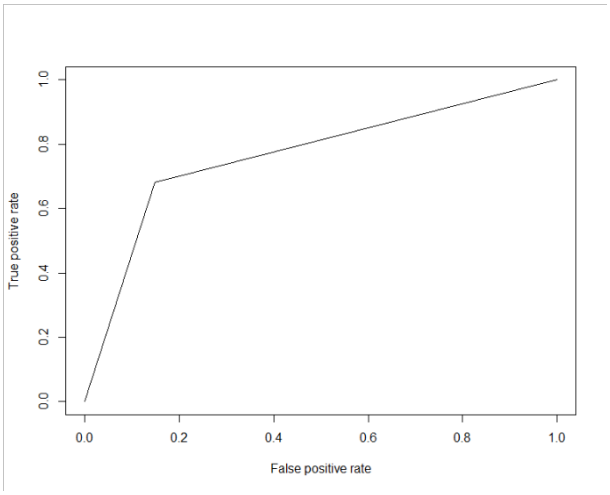
شکل ۱۰

Confusion matrix of tree

	۰	۱
۰	۱۱۴۰	۱۹۸
۱	۵۱۳	۱۱۰۰

باتوجه به نتیجه ماتریس پرفورمنس، هم sensitivityو هم specify قابل قبول داریم و accuracy برابر 75%است.

شکل ۱۱ منحنی ROC را برای مدل نشان میدهد که مساخت زیر نمودار آن (AUC) برابر0.7669885 است.



شکل ۱۱

Logistic regression model

با در نظر گرفتن متغیرanswered به عنوان متغیر پاسخ مدل رگرسیون را حساب میکینم و برای مقادیری از پیش بینی که احتمالی بزرگتر از ۰/۵ میگیرند کلاس ۱ را اختصاص میدهیم. شکل ۱۲ میزان تاثیرگذاری و significant بودن هر متغیر (دامی شده متغیرهای کتگوریکال) را نشان میدهد.

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.020e+00  1.270e-01  -8.037 9.23e-16 ***
INCOME         -2.147e-05  6.214e-07 -34.542  < 2e-16 ***
FEMALE1        -5.824e-01  6.445e-02  -9.036  < 2e-16 ***
AGE             1.871e-02  1.565e-03  11.956  < 2e-16 ***
JOB1           -4.482e-01  1.083e-01  -4.139 3.49e-05 ***
JOB2           -5.122e-01  1.045e-01  -4.900 9.60e-07 ***
JOB3           -6.527e-01  1.063e-01  -6.141 8.19e-10 ***
NUM_DEPENDENTS2 -1.483e-01  4.376e-02  -3.389 0.000702 ***
RENT1           5.675e-02  6.103e-02   0.930 0.352418
OWN_RES1        4.153e-01  5.354e-02   7.758 8.66e-15 ***
NEW_CAR1       -5.531e-01  3.478e-02 -15.903  < 2e-16 ***
CHK_ACCT1       2.163e-01  3.710e-02   5.829 5.56e-09 ***
CHK_ACCT2       3.422e-01  6.033e-02   5.672 1.42e-08 ***
CHK_ACCT3       1.698e+00  3.886e-02  43.707  < 2e-16 ***
SAV_ACCT1      -3.940e-02  4.849e-02  -0.812 0.416548
SAV_ACCT2       7.181e-01  6.604e-02  10.873  < 2e-16 ***
SAV_ACCT3       1.252e+00  8.137e-02  15.389  < 2e-16 ***
SAV_ACCT4       4.858e-01  4.402e-02  11.037  < 2e-16 ***
NUM_ACCTS1      5.695e-01  8.035e-02   7.087 1.37e-12 ***
NUM_ACCTS2      7.012e-01  7.790e-02   9.002  < 2e-16 ***
NUM_ACCTS3      1.191e+00  8.172e-02  14.579  < 2e-16 ***
NUM_ACCTS4      8.738e-01  7.876e-02  11.095  < 2e-16 ***
MOBILE1         1.606e+00  6.348e-02  25.306  < 2e-16 ***
---
```

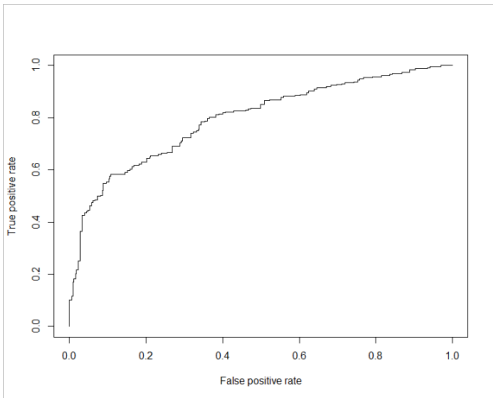
شکل ۱۲

Confusion matrix of logistic regression

	۰	۱
۰	947	391
۱	476	1137

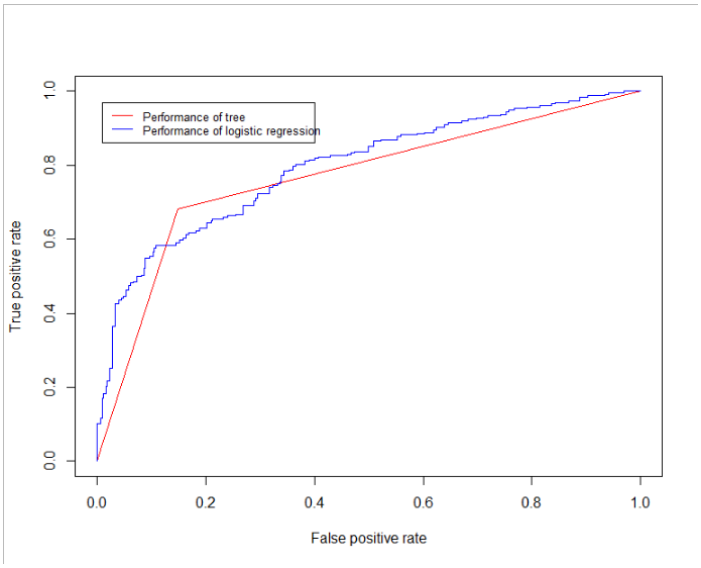
باتوجه به نتیجه مدل accuracy برابر ۷۰٪ است.

شکل ۱۳ منحنی ROC را برای مدل لاجستیک رگرسیون نشان میدهد و مساحت زیر نمودار آن برابر0.7951857 است.



شکل ۱۳

درنهایت شکل ۱٤ مقایسه پرفورمنس این دو مدل را درکنار یکدیگر نشان میدهد.



شکل ۱٤

باتوجه به نتیجه و مقایسه شکل فلان و بزرگتر بودن مساحت زیر نمودار ROC مدل لاجستیک رگرسیون را انتخاب میکنیم و سراغ اختصاص منابع میرویم.

دیتا فریمی از دیتای تست می‌سازیم و احتمال محاسبه شده توسط مدل رگرسیون را به آن اضافه میکنیم، این دیتافریم جدید با نام compare_result ساخته شده است حالا با مرتب‌سازی دیتا براساس ستون احتمال محاسبه شده، منابع (فروشندگان) را تخصیص می‌دهیم. همچنین باتوجه به درآمدی که هر تماس پاسخ داده شده ۱۱۰ دلار درامد دارد، ستون درآمد احتمالی را هم ایجاد میکنیم.

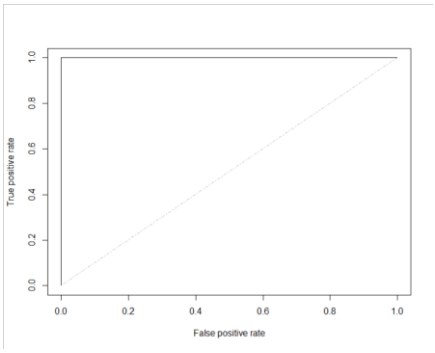
	ANSWERED	INCOME	FEMALE	AGE	JOB	NUM_DEPENDENTS	RENT	OWN_RES	NEW_CAR	CHK_ACCT	SAV_ACCT	NUM_ACCTS	MOBILE	c.log_pred.	expected_value
3741	1	-0.4006348122	0	0.72366403	2	2	0	1	0	3	4	3	1	0.9828609	110
6694	1	-0.4006348122	0	0.72366403	2	2	0	1	0	3	4	3	1	0.9828609	110
7904	1	-0.4006348122	0	0.72366403	2	2	0	1	0	3	4	3	1	0.9828609	110
9201	1	-0.4006348122	0	0.72366403	2	2	0	1	0	3	4	3	1	0.9828609	110
15325	1	-0.4006348122	0	0.72366403	2	2	0	1	0	3	4	3	1	0.9828609	110
16810	1	-0.4006348122	0	0.72366403	2	2	0	1	0	3	4	3	1	0.9828609	110
2736	1	-0.0001242613	0	-1.03599265	2	1	1	0	0	3	3	4	1	0.9749677	110
3396	1	-0.0001242613	0	-1.03599265	2	1	1	0	0	3	3	4	1	0.9749677	110
3424	1	-0.0001242613	0	-1.03599265	2	1	1	0	0	3	3	4	1	0.9749677	110
3629	1	-0.0001242613	0	-1.03599265	2	1	1	0	0	3	3	4	1	0.9749677	110
8888	1	-0.0001242613	0	-1.03599265	2	1	1	0	0	3	3	4	1	0.9749677	110
8973	1	-0.0001242613	0	-1.03599265	2	1	1	0	0	3	3	4	1	0.9749677	110
11569	1	-0.0001242613	0	-1.03599265	2	1	1	0	0	3	3	4	1	0.9749677	110
19059	1	-0.0001242613	0	-1.03599265	2	1	1	0	0	3	3	4	1	0.9749677	110
22385	1	-0.0001242613	0	-1.03599265	2	1	1	0	0	3	3	4	1	0.9749677	110
23440	1	-0.0001242613	0	-1.03599265	2	1	1	0	0	3	3	4	1	0.9749677	110
26959	1	-0.0001242613	0	-1.03599265	2	1	1	0	0	3	3	4	1	0.9749677	110
29496	1	-0.0001242613	0	-1.03599265	2	1	1	0	0	3	3	4	1	0.9749677	110
5878	1	-0.8237189723	0	-0.42011281	2	1	0	1	0	3	3	3	0	0.9670635	110
16157	1	-0.8237189723	0	-0.42011281	2	1	0	1	0	3	3	3	0	0.9670635	110
26672	1	-0.8237189723	0	-0.42011281	2	1	0	1	0	3	3	3	0	0.9670635	110
27951	1	-0.8237189723	0	-0.42011281	2	1	0	1	0	3	3	3	0	0.9670635	110

شکل ۱۵آ

با استفاده از مدل احتمال آموزش داده شده و محاسبه احتمال پاسخ دهی هر مشتری باتوجه به اطلاعات پروفایل آن در هر بازه زمانی این مرتب سازی براساس احتمال را انجام میدهیم و با توجه به این احتمال نیروهای فروش را تخصیص میدهیم تا به حداکثر درآمد برسیم و از اتلاف وقت منابع جلوگیری کنیم.

Random forest model

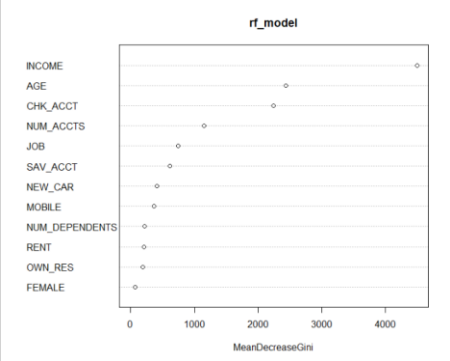
مدل جنگل را هم استفاده میکنیم تا ببینیم به نتیجه بهتری میرسیم یا نه. با در نظر گرفتن پارامترهای تنظیم نشده به دقت ۱ رسیدیم که خب عجیب به نظر میرسه و به نظر میرسه که یه جایی overfit کرده جنگلمون. شکل ۱۱ نمودار ROC جنگل را نشان میدهد.



شکل ۱۶

Importance variable

متغیرهایی که در جنگل اهمیت بیشتری داشته اند را در شکل ۱۶و۱۷ مشاهده میکنید.



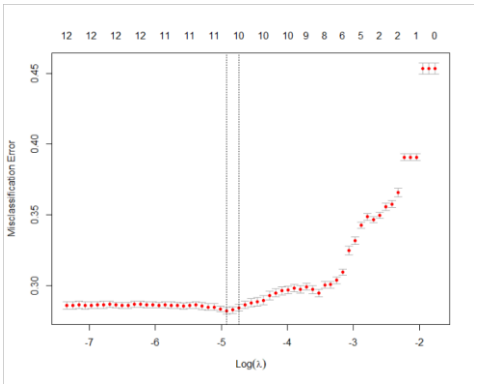
شکل ۱۷

```
> importance(rf_model)
                MeanDecreaseGini
INCOME               4495.53930
FEMALE                 70.67262
AGE                  2441.25002
JOB                   741.25203
NUM_DEPENDENTS        219.65891
RENT                  211.97446
OWN_RES               190.54994
NEW_CAR               409.60646
CHK_ACCT             2241.97956
SAV_ACCT              616.99933
NUM_ACCTS            1152.76795
MOBILE                366.99501
```

شکل ۱۸

Lasso regression model:

درنهایت با استفاده از مدل رگرسیون لاسو هم سعی میکنیم دیتاست را بررسی کنیم، تا با استفاده از این مدل هم به تفسیر پذیری بهتری دست پیدا کنیم. مقادیر لاندا متفاوت را بررسی میکنیم تا به مقادیر بهینه لاندا برسیم.



شکل ۱۹

cv.lasso$lambda.min

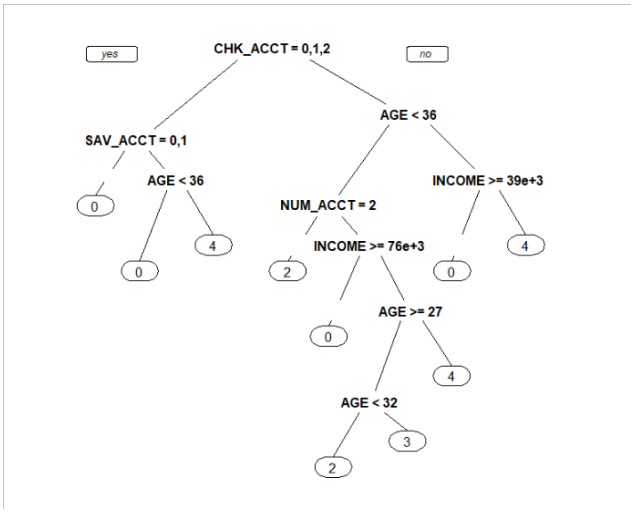
۰/۰۰۷۲۷۲۱۴۹

cv.lasso$lambda.1se

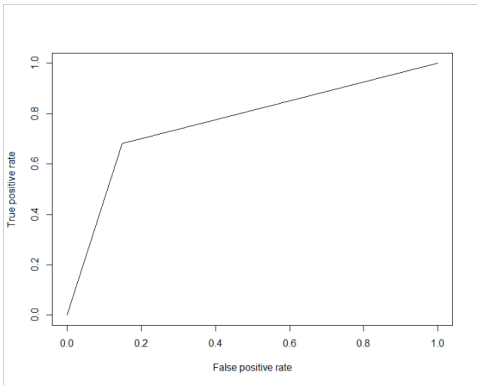
۰/۰۰۸۷۵۹۳۳

استفاده از راهکار دوم:

در راهکار دوم درواقع با حذف متغیر answered به پیش بینی متغیر product میپردازیم که در درون تفسیرش، معنای answered را هم شامل میشود. شکل ۲۰ درخت پیش بینی برای کلاس های این متغیر را نشان میدهد و شکل ۲۱ منحنی ROC درخت را نشان میدهد که مقدار AUC ان برابر ۰/۷۶۶۹۸۸۵ است و شکل ۲۲ ماتریس پرفورمنس کلاس ها را نشان میدهد.



شکل ۲۰



شکل ۲۱

t_pred	0	1	2	3	4
0	1157	0	81	19	81
1	150	0	66	15	12
2	196	0	185	0	86
3	124	0	25	34	73
4	256	0	24	24	342

شکل ۲۲