2022

پروژه درس مدلسازی و تصمیم‌گیری داده محور

استاد درس: دکتر نفیسه صدقی

معصومه محمودی – محمد زارعی

دانشگاه صنعتی شریف | نیم‌سال دوم ۱۴۰۰ – ۱۴۰۱

فهرست

فهرست اشکال

چکیده

در ابتدا بر روی داده‌های این دیتاست عملیات مورد پیش‌پردازش (Preprocessing) قرار گرفت تا داده هایی که در فرایند بررسی و مصور سازی و پیش بینی و نتیجه گیری کاربرد زیادی ندارند و بیشتر باعث اختلال در عملکرد می شوند تا حد امکان کنار گذاشته شوند. سپس با استفاده از نمودارهای مختلف ارتباط بین متغیرها و تاثیری که بر روی یکدیگر دارند را بررسی شد. با استفاده از نمودارها و Visualization داده ها اطلاعات کلی از مدل ارتباط متغیرها با یکدیگر بدست می آوریم. در ادامه با توجه به این که متغیر پاسخ از نوع دسته‌ای (Categorical) می‌باشد، چند مدل طبقه‌بندی (Classification) ایجاد شده و از میان آن‌ها یک مدل که عملکرد بهتری داشته برای پیش‌بینی متغیر پاسخ در دیتاست test استفاده شده است.

واژگان کلیدی: سرمایه گذاری، خودرو، مبتدی، بانکداری، رگرسیون لجستیک

۱ توضیح مساله و داده‌ها

۱-۱ وام خودرو

خودروها رایج ترین دارایی های غیرمالی در میان افراد هستند. تقریباً سه چهارم خریدهای خودرو از طریق اعتبار تأمین می‌شود و وام‌های خرید خودرو یکی از رایج‌ترین اشکال استقراض خانوارها است. وام دهندگان در بازار خودرو با خطراتی مواجه هستند. اولین و آشکارترین ریسک، عدم بازپرداخت وام است. یعنی شخصی که برای خرید خودرو وام گرفته است و آن را پس نمی‌دهد. دومین ریسک مهم برای وام دهندگان در این بازار، ریسک پیش پرداخت است. یعنی خریدار خودرو وام را زودتر پرداخت می کند و جریان پرداخت بهره وام دهنده را کاهش می دهد.

مؤسسه مالی غیر بانکی (NBFI)[1] یا شرکت مالی غیر بانکی (NBFC)[2] مؤسسه مالی است که مجوز کامل بانکی ندارد یا توسط یک آژانس نظارتی بانکی ملی یا بین‌المللی نظارت نمی‌شود. NBFC خدمات مالی مرتبط با بانک را تسهیل می کند، مانند سرمایه گذاری، تجمیع ریسک[3]، پس انداز قراردادی[4] و کارگزاری بازار[5]. یک NBFI به دلیل افزایش عدم پرداخت در رده وام خودرو در تلاش برای نشان دادن سود است.

هدف شرکت تعیین توانایی های بازپرداخت وام مشتری و درک اهمیت نسبی هر پارامتری است که به توانایی وام گیرنده برای بازپرداخت وام کمک می کند.

در این پروژه تلاش بر این است که با توجه به متغیرها و مشخصاتی که از مشتریان (وام گیرندگان) در اختیار داریم و تعدادی از مشتریان که در بخش train وضعیت بازپرداخت آنها مشخص شده، پیش‌بینی کنیم آیا مشتریان بازپرداخت وام خودرو را انجام می دهند یا خیر؟

[1] Non-Banking Financial Institution
[2] Non_Bank Financial Comapny
[3] Risk pooling
[4] Contractual Savings
[5] Market Brokering

۱-۲ معرفی دیتاست (Dataset)

این مسئله شامل دو دیتاست با نام‌های Train_Dataset و Test_Dataset است. مدل‌های تحلیلی با کمک بخشی از دیتاست Train_Dataset ساخته شده و بر روی بخش دیگری از آن، مورد ارزیابی قرار گرفت. سپس، داده‌های دیتاست Test_Dataset به عنوان ورودی پیش‌بینی به مدل با عملکرد بهتر داده شد. با توجه به این که متغیر پاسخ در دیتاست Test_Dataset وجود ندارد، لازم است که مدل‌های ارائه‌شده پیش از پیش‌بینی نهایی با استفاده از بخشی از دیتاست Train_Dataset مورد آزمایش قرار گیرد.

۱-۳ Data Dictionary

دیتاست مورد بررسی دارای ۱۲۱۸۵٦ مشاهده(observation) است که تعداد مشتریان بررسی شده می باشد و از ٤۰ متغیر مختلف درباره‌ی اطلاعات مربوط به هر مشتری استفاده شده که در جدول زیر انواع آنها و توضیحات هر کدام قابل مشاهده است. در این جدول، منظور از «Cat.» ، دسته‌ای (Categorical) و منظور از «Num.» ، عددی (Numerical) است.

جدول ۱-۱

شماره	متغیر	type	توضیح	سطوح داده (Cat.)
۱	ID	Num.	شناسه درخواست وام مشتری	
۲	Client_Income	Num.	درآمد مشتری به واحد دلار	
۳	Car_Owned	Cat.	خودرویی که قبل از درخواست وام برای خودروی دیگر متعلق به مشتری باشد	۰: متعلق به مشتری نباشد ۱: متعلق به مشتری باشد
٤	Bike_Owned	Cat.	دوچرخه متعلق به مشتری (۰ به معنای خیر و ۱ به معنای غیر آن است)	۰: متعلق به مشتری نباشد ۱: متعلق به مشتری باشد
٥	Active_Loan	Cat.	وام فعال دیگری در زمان درخواست وام	۰: وام فعال ندارد ۱: وام فعال دارد
٦	House_Own	Cat.	مشتری دارای خانه می باشد یا خیر	۰: مشتری خانه ندارد

				۱: مشتری خانه دارد
۷	Child_Count	Cat.	تعداد فرزندانی که مشتری دارد	
۸	Credit_Amount	Num.	مبلغ اعتبار وام به واحد دلار	
۹	Loan_Annuity	Num.	سالیانه وام به واحد دلار	
۱۰	Accompany_Client	Cat.	چه کسی مشتری را هنگام درخواست وام همراهی کرد	
۱۱	Client_Income_Type	Cat.	نوع درآمد مشتریان	
۱۲	Client_Education	Cat.	سطح تحصیلات مشتری	
۱۳	Client_Marital_Status	Cat.	وضعیت تأهل مشتری	D: طلاق گرفته S: مجرد M: متاهل W: بیوه
۱٤	Client_Gender	Cat.	جنسیت مشتری	
۱٥	Loan_Contract_Type	Cat.	نوع وام	CL: وام نقدی RL: وام گردان
۱٦	Client_Housing_Type	Cat.	وضعیت مسکن مشتری	
۱۷	Population_Region_Relative	Num.	جمعیت نسبی منطقه ای که مشتری در آن زندگی می کند	ارزش بالاتر یعنی مشتری در جای خوبی زندگی می کند
۱۸	Age_Days	Num.	سن مشتری در زمان ارسال درخواست وام	
۱۹	Employed_Days	Num.	چند روز قبل از درخواست وام، مشتری شروع به کسب درآمد کرد	
۲۰	Registration_Days	Num.	چند روز قبل از درخواست وام، مشتری ثبت نام خود را تغییر داد	
۲۱	ID_Days	Num.	چند روز قبل از درخواست وام، مشتری مدرک هویتی خود را تغییر داد که با آن وام درخواست دهد.	
۲۲	Own_House_Age	Num.	سن خانه مشتری به سال	
۲۳	Mobile_Tag	Cat.	شماره موبایل توسط مشتری ارائه شده یا خیر	۰: شماره ثبت نشده

				۱: شماره ثبت شده
۲٤	Homephone_Tag	Cat.	شماره تلفن خانگی توسط مشتری ارائه شده یا خیر	۰: شماره ثبت نشده ۱: شماره ثبت شده
۲۵	Workphone_Working	Cat.	آیا شماره تلفن کار قابل دسترسی بود یا خیر	۰: شماره در دسترس نبود ۱: شماره در دسترس بود
۲٦	Client_Occupation	Cat.	نوع شغل مشتری	
۲۷	Client_Family_Members	Cat.	تعداد اعضای خانواده مشتری	
۲۸	Cleint_City_Rating	Cat.	رتبه شهر مشتری	۱: متوسط ۲: خوب ۳: عالی
۲۹	Application_Process_Day	Cat.	روزی از هفته که مشتری درخواست وام کرده است	۰: یکشنبه ۱: دوشنبه ۲: سه شنبه ۳: چهارشنبه ٤: پنجشنبه ۵: جمعه ٦: شنبه
۳۰	Application_Process_Hour	Num.	ساعتی از روز که مشتری درخواست وام کرده است	
۳۱	Client_Permanent_Match_Tag	Cat.	آدرس تماس مشتری با آدرس دائمی مطابقت دارد یا خیر	Yes: مطابقت دارد No: مطابقت ندارد
۳۲	Client_Contact_Work_Tag	Cat.	آدرس کار مشتری با آدرس تماس مطابقت دارد یا خیر	Yes: مطابقت دارد No: مطابقت ندارد
۳۳	Type_Organization	Cat.	نوع سازمانی که مشتری در آن کار می کند	

۳٤	Score_Source_1	Num.	امتیاز از منبع دیگری گرفته شده است. (نرمال شده)	
۳۵	Score_Source_2	Num.	امتیاز از منبع دیگری گرفته شده است. (نرمال شده)	
۳٦	Score_Source_3	Num.	امتیاز از منبع دیگری گرفته شده است. (نرمال شده)	
۳۷	Social_Circle_Default	Num.	چند نفر از دوستان/عضو خانواده مشتری در ٦۰ روز گذشته پرداخت وام را نپذیرفته اند	
۳۸	Phone_Change	Num.	مشتری چند روز قبل از درخواست وام تلفن خود را عوض کرده است	
۳۹	Credit_Bureau	Cat.	تعداد کل درخواست های وام در سال گذشته	
٤۰	Default	Cat.	پرداخت یا عدم پرداخت وام	۰: پرداخت وام ۱: عدم پرداخت وام

۲ تحلیل کاوش‌گرایانه داده (EDA)

به فرآیند کاوش داده‌ها با بکارگیری خلاصه‌های عددی و نمودارها[6]، برای شناسایی روابط احتمالی میان متغیرها، Exploratory Data Analysis (EDA) گفته می‌شود. به کمک EDA ، می‌توان ناهنجاری‌هایی همچون داده‌های پرت (Outliers) و یا مشاهدات غیرعادی را یافت، الگوها را کشف نمود و پرسش‌هایی ایجاد کرد که بعدها توسط روش‌های آماری رسمی‌تر، مورد بررسی قرار گیرد. در اجرای EDA ، تحلیلگر یا دانشمند داده همچون کارآگاهی عمل می‌کند که به دنبال یافتن سرنخ‌ها و بینش‌هایی برای شناسایی دلایل ریشه‌ای چالشی است که در تلاش برای حل آن است.

اگرچه جداول خلاصه آماری شامل اطلاعاتی همچون میانگین و انحراف استاندارد نیز بخشی از EDA هستند، بیشتر افراد در اجرای EDA بر گراف‌ها تمرکز دارند. تحلیلگر گراف‌ها و دیگر ابزارهای کاوش‌گرایانه را بکار می‌گیرد و به جایی می‌رود که داده‌ها او را هدایت می‌کنند. هرگاه گراف یا تحلیلی نتواند اطلاعات کافی در اختیار تحلیلگر قرار دهد، او داده‌ها را از جنبه دیگری مورد بررسی قرار می‌دهد.

دو بخش مهم EDA، تحلیل تک متغیره (Univariate Analysis) است که در سطوح اولیه به کمک جداول خلاصه آماری و مصورسازی داده (Data Visualization) انجام می‌شود.

۲-۱ بررسی آماری دیتاست

نخستین نکته‌ای که لازم است پس از بارگزاری دیتاست در نرم‌افزار RStudio مورد بررسی قرار گیرد، فرمت ذخیره‌سازی متغیرهای دیتاست است. در این دیتاست نیز برخی متغیرهای عددی به صورت کاراکتر (Character) ذخیره شده بودند که لازم بود پیش از بررسی آماری آن‌ها را به فرمت عددی تبدیل نمود. از میان ٤٠ متغیری که در این دیتاست وجود دارد، یک متغیر ID[7]، ١٦ متغیر .Num و ٢٣ متغیر .Cat دیده می‌شود. البته برخی متغیرهای دسته‌ای ترتیبی (.Ordinal Cat) به گونه‌ای هستند که با توجه به نوع تحلیل و نیاز می‌توان آن‌ها را .Cat و یا .Num در نظر گرفت.

برای بررسی اولیه آماری دیتاست در R، معمولا از دستور ()summary استفاده می‌شود اما در این پروژه از دستور ()describe از پکیج «Hmisc» استفاده شد که اطلاعات بیشتری همچون پنج عدد بزرگترین و

[6] Visualizations
[7] Identifier

کوچکترین برای متغیرهای .Num و درصد هر یک از سطوح متغیرهای .Cat را ارائه می‌دهد. در ادامه به بیان مهم‌ترین نکات بررسی آماری متغیرهای دیتاست اشاره شده است.

۲-۲ تحلیل تک متغیره (Univariate Analysis)

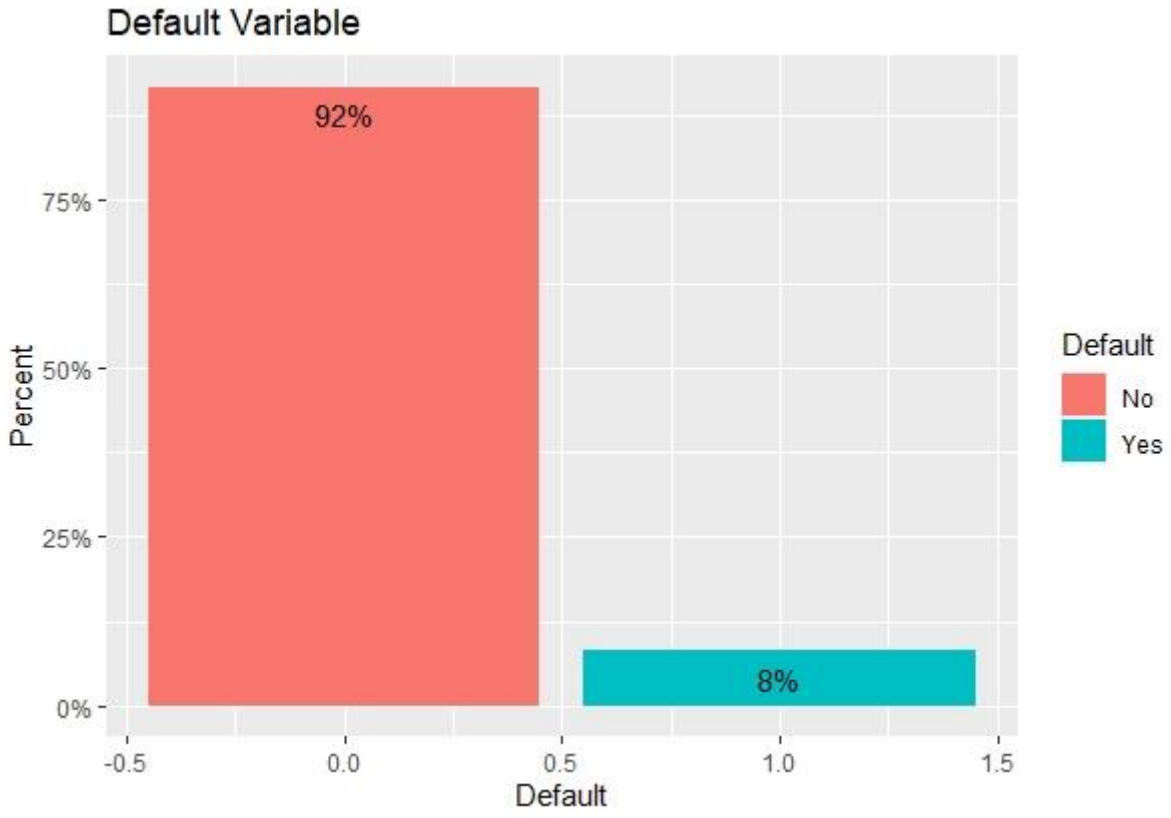
تحلیل تک متغیره ساده‌ترین شکل تحلیل داده است که در آن داده‌های مورد تجزیه و تحلیل، تنها از یک متغیر تشکیل شده است. از آنجایی که تنها حضور یک متغیر مطرح است، با دلایل و روابط سر و کار ندارد. هدف اصلی تحلیل تک متغیره توصیف داده‌ها و یافتن الگوهای موجود در آن‌هاست. با توجه به تعداد بالای متغیرها در این دیتاست، در ادامه، به بیان مهم‌ترین نتایج تحلیل تک متغیره در این دیتاست که به کمک خلاصه‌های آماری و نمودارها بدست آمده، پرداخته شده است:

- **Client_Income** : میانگین و میانه این متغیر به ترتیب ۱۶۸۶۵ و ۶۷۵۰ بوده است که نشان از چولگی زیاد و نرمال نبودن توزیع داده‌ها دارد. نکته دیگر این است که بزرگترین ۵ داده این متغیر از ۳۸۲۵۰۰ تا ۱۸۰۰۰۰۹ هستند. حضور داده ای به بزرگی یک میلیون و ۸۰۰ هزار در کنار داده‌هایی با میانه ۶۷۵۰ ، صرف نظر از درست بودن یا نبودن داده، می‌تواند بر روی برازش مدل، تاثیر منفی داشته باشد.
- **Child_Count** : این متغیر از ۰ تا ۱۹ مقدار گرفته است ولی توزیع داده‌ها نامتوازن بوده است، به گونه ای که اگر هر مقدار را یک دسته در نظر بگیریم، هر چه از صفر به سمت ۱۹ حرکت می‌کنیم، تعداد افراد کاهش می‌یابد، این کاهش به قدری شدید است که افراد بدون فرزند، ۸۲۸۳۴ نفر بوده اند، در حالی که افراد با ۵ فرزند، ۳۴ نفر و افراد دارای بیش از ۷ فرزند، انگشت شمار بوده اند.
- **Accompany Client** : این متغیر .Cat، هفت سطح دارد که یکی از آن‌ها تنها دارای ۱۲ داده است و نام آن هم «##» است که نامشخص است.
- **Client_Income_Type** : این متغیر .Cat، هشت سطح دارد ولی چهار مورد از آن‌ها به ترتیب یک، دو، شش و هشت داده دارند.
- **Client_Gender** : برای این متغیر سه سطح تعریف شده است که یکی از آن‌ها «XNA» است که دقیقا مشخص نیست منظور از این مقدار چه بوده است، با این حال این سطح تنها سه داده دارد.

- **Employed_Days** : در توضیحات این متغیر آمده که چند روز پیش از درخواست وام، شخص آغاز به کسب درآمد کرده است، با این حال بیش از ۲۱۰۰۰ داده از این متغیر مقداری بیش از ۳۰۰۰۰۰ دارند که تقریبا معادل ۹۰۰ سال است.
- **Mobile_Tag** : این یک متغیر Boolean است ولی سطح صفر آن تنها یک داده دارد و ۱۲۱۸۵۵ داده مقدار ۱ دارند.
- **Client_Family_Members** : این متغیر نیز شرایطی همچون متغیر Child_Count دارد و انتظار می‌رود که رابطه‌ای میان این دو متغیر برقرار باشد.
- **Type_Organization** : یک متغیر .Cat است که ۵۸ سطح دارد. به نظر می‌رسد برخی از سطوح این متغیر، زیرمجموعه ای از یک سطح کلی بزرگتر باشند، برای نمونه ۱۳ سطح این متغیر دارای نام‌های «Industry: type 1» تا «Industy: Type 13» هستند. همچنین یک سطح نامشخص به نام «XNA» وجود دارد.
- **Source_Score_1** ، **Source_Score_2** و **Source Score 3** : دقیقا مشخص نیست که این سه متغیر بیانگر چه ویژگی هستند و در راهنمای دیتاست نیز توضیحی در این باره ارائه نشده است، تنها می‌دانیم که این سه متغیر به صورت پیوسته مقادیر صفر تا یک را اختیار می‌کنند.

Default : این متغیر، متغیر پاسخ مساله است و همان‌گونه که پیش از این گفته شد، چون یک متغیر Boolean است، برای پیش‌بینی آن لازم است که از مدل‌های Classification استفاده شود. نکته ای که درباره این متغیر وجود دارد عدم توازن آن است که می‌تواند بر دقت مدل برازش شده تاثیر منفی داشته باشد. (

- (



شکل ۲-۱ نمودار Bar Chart برای متغیر Default

۳ آماده‌سازی داده (Data Preparation)

برای دستیابی به یک مدل پیش‌بینی با دقت مناسب، لازم است که داده‌ها پس از بررسی مورد ویرایش قرار گیرند تا مدل نهایی خطای کمتری داشته باشد. این آماده‌سازی می‌تواند شامل مراحل مختلفی باشد که در این دیتاست، شامل مدیریت داده‌های پرت یا نادرست و بررسی و جایگزینی (Imputation) داده‌های گم شده (Missing Values) است. در طی فرآیند آماده‌سازی داده‌ها ممکن است تصمیم گرفته شود که بخشی از داده‌ها حذف شوند. در این پروژه به منظور مقایسه و دستیابی به مدل مناسب دو دیتاست با دو رویکرد متفاوت ایجاد می‌شود به گونه‌ای که:

۱. **df.norm** : در دیتاست اول هیچ داده‌ای حذف نمی‌گردد و تنها به گونه‌ای که الگوی داده‌ها حفظ شود، جایگزینی صورت می‌گیرد.
۲. **df_omit** : ترکیبی از حذف و جایگزینی صورت می‌گیرد.

۳-۱ داده‌های پرت (Outliers)

یک تابع برای مشخص نمودن داده‌های پرت نوشته شد و تعداد داده‌های پرت متغیرهای .Num محاسبه شد، در دیتاست دوم تمامی داده‌های پرت حذف شدند ولی در دیتاست اول، موارد زیر اعمال شد:

- **Client_Income** : تعداد داده‌های پرت، ۵۳۹۲ بدست آمد که تعداد نسبتا زیادی است، با این حال در مورد این متغیر، گسترده بودن دامنه داده‌ها تا حدودی طبیعی است، بنابراین سطح ۱۰۰۰۰۰ به عنوان مبنا در نظر گرفته شد و داده‌های بیشتر از این سطح با عدد ۱۰۰۰۰۰ جایگزین شد، بدین ترتیب مقدار ۱۰۱ داده جایگزین شد.
- **Child_Count** : تعداد داده‌های پرت این متغیر، ۱۶۵۹ بدست آمد ولی با بررسی سطوح داده عدد ٤ به عنوان پایه در نظر گرفته شد، مقادیر بیشتر از ٤ با عدد ٤ جایگزین شد.
- **Client_Family_Members** : همانند متغیر پیشین با این تفاوت که عدد ٦ به عنوان مبنا در نظر گرفته شد.
- **Credit_Bereau** : همانند متغیر پیشین با این تفاوت که عدد ۱۰ به عنوان مبنا در نظر گرفته شد.

- **Employment_Days** : تعداد داده‌های پرت ۲۱۴۸۸ بدست آمد، با توجه به غیر منطقی بودن مقادیر این داده‌ها، تمامی آن‌ها به داده گم شده یا ناموجود (NA)[8] تبدیل شد تا با روش‌هایی که در ادامه بیان خواهد شد، با مقدار مناسب جایگزین شود.
- **Score_Source_2** : تعداد داده‌های پرت، ٦ بدست آمد که به NA تبدیل شد.

در مورد متغیرهای .Cat ، نیز موارد زیر اعمال شد:

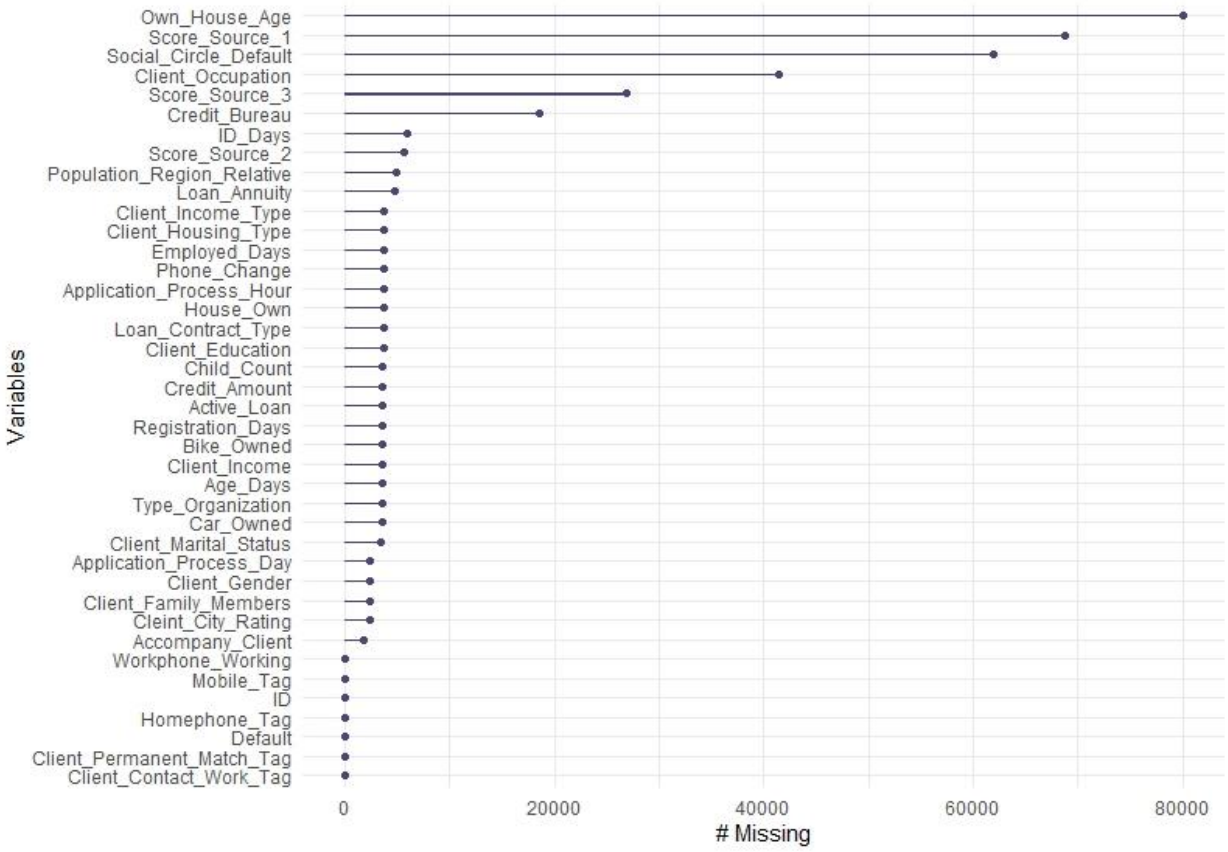
- **Client_Income_Type** : چهار سطح این متغیر که تعداد بسیار کمی داده داشتند، در یک سطح جدید به نام «Other»، ادغام شدند.
- **Accompany_Client** : سطح نامشخص «##» در دیتاست اول به NA تبدیل شد و در دیتاست دوم حذف شد.
- **Client_Gender** : سطح نامشخص «XNA» در دیتاست اول به NA تبدیل شد و در دیتاست دوم حذف شد.
- **Type_Organization** : سطوح مشابه این متغیر در یکدیگر ادغام شدند و به این ترتیب تعداد سطوح متمایز از ۵۸ به ۳۵ کاهش یافت. با این حال این تعداد نیز زیاد است و در اجرای الگوریتم های جایگزینی که در ادامه بیان خواهد شد مشکل ایجاد خواهد کرد، از این رو، داده‌های NA این متغیر به سطح نامشخص «XNA» تبدیل شد که پیش از این حدود ۲۲۰۰۰ داده داشت.
-

۲-۳ داده‌های گم‌شده یا ناموجود (NA - Missing)

برخی از داده‌ها در دیتاست به صورت NA ثبت شده اند ولی برخی دیگر به ویژه در فرمت کاراکتر خالی (NULL) هستند و این مساله سبب می‌شود که در نگاه سطحی به عنوان NA در نظر گرفته نشوند. بدین منظور در هنگام وارد نمودن دیتاست، لازم است در دستور read.csv() ، ویژگی na.strings را برابر با مقدار TRUE قرار دهیم. در ادامه برای بررسی داده‌های NA ، تعداد کل سطرهای کامل دیتاست شمارش شد که عدد ۲۵٦۸ بدست آمد که معادل ۲ درصد از کل داده ها است، به عبارت دیگر، حدود ۹۸ درصد از سطرهای دیتاست، دست کم یک مقدار NA دارند. سپس تعداد NA ها در متغیرهای مختلف مورد بررسی

[8] Not Available

قرار گرفت و مشخص شد که از میان ٤٠ متغیر موجود در داده، شش مورد بیش از پنج درصد مقدار NA دارند که این مقادیر به ترتیب از ٦٥ تا ١٥ درصد است (شکل ٣-١).



شکل ٣-١ تعداد مقادیر NA متغیرها

تغییرات مرحله آماده‌سازی داده که تا کنون بیان شده در قالب یک تابع، پیاده‌سازی شد. بدین ترتیب، در نتیجه اعمال این تغییرات، دیتاست اول بدون حذف هیچ گونه داده‌ای دارای ١٢١٨٥٦ مشاهده و دیتاست دوم دارای ١١٠٨٨١ مشاهده که نشان از حذف ١٠٩٧٥ سطر دارد. همچنین تعداد سطرهای بدون NA از ٢٥٦٨ مورد به ٢٦٣٥ مورد در دیتاست اول و ٢١٤٢ مورد در دیتاست دوم تغییر کرده است.

٣-٣ مقداردهی (Imputaion)

مدیریت داده‌های از دست رفته از اهمیت بسیار بالایی برخوردار است، زیرا که بسیاری از الگوریتم‌های یادگیری آماری (Statistical learning) و یادگیری ماشین (Machine learning) از NA ها پشتیبانی نمی‌کنند، با این حال الگوریتم‌های K-nearest و Naïve Bayes از داده‌های از دست رفته پشتیبانی می‌کنند. دلیل دیگر این است که ممکن است این داده‌ها در نهایت منجر به ایجاد یک مدل جانبدارانه (Biased

Model) و یا عدم دقت در تجزیه و تحلیل ها گردند. به طور کلی دو راه برای مدیریت داده‌های از دست رفته وجود دارد که یکی حذف و دیگری مقداردهی (Imputation) داده‌های از دست رفته است. حذف داده‌های از دست رفته می‌تواند از راه حذف سطرها و یا ستون‌های دیتاست صورت گیرد.

Imputation به معنای جایگزینی یک مقدار از دست رفته (NA) با یک مقدار دیگر بر اساس یک برآورد معقول است. به عبارت دیگر از داده‌های موجود در دیتاست برای مقداردهی داده‌های از دست رفته به منظور دستیابی به یک دیتاست کامل‌تر استفاده می‌شود. روش‌های متفاوتی برای Imputation وجود دارد که از میان آن‌ها می‌توان به انتخاب داده تصادفی، بکارگیری میانه یا میانگین برای داده‌های .Num و بکارگیری مد برای داده‌های .Cat نام برد. با این حال این روش‌ها هرگز توصیه نمی‌شوند زیرا می‌توانند کیفیت داده‌ها را کاهش داده و الگوی داده‌ها را به طور کلی تغییر دهند. برای نمونه در مورد این دیتاست، متغیر Own_House_Age حدود ۸۰ هزار داده از دست رفته از ۱۲۱ هزار مشاهده دارد، حال اگر این داده‌ها را با میانه جایگزین کنیم، به یکباره ٦٥ درصد به فراوانی میانه افزوده می‌شود و این روش قطعا الگوی داده‌ها را به طور کلی تغییر خواهد داد.

یکی دیگر از روش‌های Imputation بکارگیری الگوریتم‌های پیش‌بینی برای داده‌های از دست رفته است که در این پروژه از این روش استفاده شده است. بدین منظور از پکیج «mice»[9] استفاده شده است که یک پکیج مقداردهی چندگانه (Multiple Imputation) است. روش کار mice به این صورت است که در ابتدا زنجیره‌ای از معادلات بر اساس دیگر متغیرها برای متغیری که قرار است مقادیر از دست رفته آن impute شوند تشکیل می‌دهد.سپس، الگوریتم تعیین شده برای متغیر مورد نظر را به صورت تکرارشونده به تعداد تکرارهای از پیش تعیین شده و بر اساس زنجیره معادله تشکیل شده اجرا می‌نماید. هربار که تعداد تکرارهای تعیین شده به پایان می‌رسد، یک imputation به صورت یک dataframe بر اساس نتایج تکرارها برای دیتاست مورد نظر ایجاد می‌شود. تعداد imputation ها نیز قابل تنظیم است و پس از پایان اجرای mice می توان تعیین کرد که یکی از imputation ها انتخاب شود و یا برآیندی از آن‌ها به کمک دستور pool() محاسبه و سپس جای‌گذاری شود. در پکیج mice ، ۲۷ الگوریتم متفاوت برای گونه‌های مختلف داده در نظر گرفته شده است که می‌توان آن‌ها را به صورت دستی یا پیش فرض برای impute کردن داده‌های از دست رفته بکار گرفت. به طور کلی بکارگیری پکیج‌های imputation برای متغیرهای .Cat با تعداد سطوح بالا

[9] Multivariate Imputation via Chained Equations

آسان نیست و در این پژوهش نیز موفق نشدیم متغیر Type_Organization را به وسیله imputation مقداردهی کنیم و به همین دلیل داده‌های NA این متغیر با سطح نامشخص «XNA» جایگزین شد.

در این پروژه از الگوریتم Predictive Mean Matching برای متغیرهای .Num و .Ordinal Cat ، الگوریتم Polytomous Logistic Regression برای متغیرهای .Cat با تعداد سطوح بیشتر از دو و الگوریتم Logistic Regression برای متغیرهای Boolean استفاده شده است. همچنین به دلیل زمان‌بر بودن فرآیند imputation ، تنها یک imputation حاصل از پنج بار تکرار برای دیتاست های اول و دوم که در گام پیشین آماده شده اند، در نظر گرفته شده است. در ادامه به توضیح دو الگوریتم بکارگرفته شده در این پروژه پرداخته شده است.

۱-۳-۳ Predictive Mean Matching (PMM)

PMM مقدار پیش‌بینی شده برای متغیر هدف Y را به کمک مدل imputation خاص خود محاسبه می‌کند. برای هر مقدار از دست رفته، متد PMM مجموعه کوچکی از کاندیداها (معمولا با سه، پنج یا ده عضو) از تمامی سطرهایی که مقدار پیش‌بینی شده برای آن‌ها نزدیک‌ترین مقدار به مقدار پیش‌بینی شده برای سطر با مقدار از دست رفته است، تشکیل می‌دهد. یک کاندیدا به طور تصادفی از میان کاندیداها انتخاب می‌شود و مقدار مشاهده شده آن برای جایگزینی مقدار از دست رفته در نظر گرفته می‌شود. فرض بر این است که توزیع سلول از دست رفته با داده‌های مشاهده شده کاندیداها یکسان است.

PMM یک روش ساده و همه‌کاره است، این روش برای تمامی گونه‌های متغیرها کاربرد دارد، با این حال عملکرد آن برای متغیرهای عددی بهتر است. این متغیر همچنان در برابر تبدیل متغیر هدف نیز انعطاف پذیر است برای نمونه وارد نمودن log(Y) می‌تواند نتایجی نزدیک به ورود exp(Y) داشته باشد. این روش امکان بکارگیری برای متغیرهای گسسته را نیز فراهم می‌آورد. مقداردهی ها بر اساس مقادیر مشاهده شده در سطرهای دیگر دیتاست است، بنابراین این مقداردهی ها واقع‌بینانه هستند، محاسبات خارج از محدوده داده‌های موجود رخ نخواهد داد و خبری از مشکلات مربوط به مقداردهی های نادرست و بی‌معنی همچون منفی شدن وزن و غیره نخواهد بود. همچنین مدل این روش یک مدل ضمنی (implicit) است به این معنی که نیاز به یک مدل صریح (explicit) برای مقادیر از دست رفته نیست، بنابراین این روش نسبت به بسیاری از روش‌های دیگر آسیب‌پذیری کمتری دارد.

۲-۳-۳ Polytomous (Multinomial) Logistic Regression

رگرسیون لجستیگ چندگانه یا چندجمله‌ای یک روش طبقه‌بندی (Classification) است که رگرسیون لجستیک را به مسائل چند سطحی (Multiclass) تعمیم می‌دهد، بدین معنی که بیش از دو نتیجه گسسته احتمالی وجود داشته باشد. به عبارت دیگر، مدلی برای پیش‌بینی احتمال نتایج ممکن یک متغیر وابسته Cat. که بیش از دو سطح دارد، بر اساس مجموعه‌ای از متغیرهای مستقل (که می‌توانند حقیقی، Boolean ، Cat. و ... باشند)، ارائه می‌دهد. رگرسیون لجستیک چندگانه با نام‌های مختلفی همچون Polytomous LR ، Multiclass LR ، Softmax Regression، Multinomial logit (mlogit) ، The Maximum entropy (MaxEnt) Classifier و Conditional Maximum entropy شناخته می‌شود.

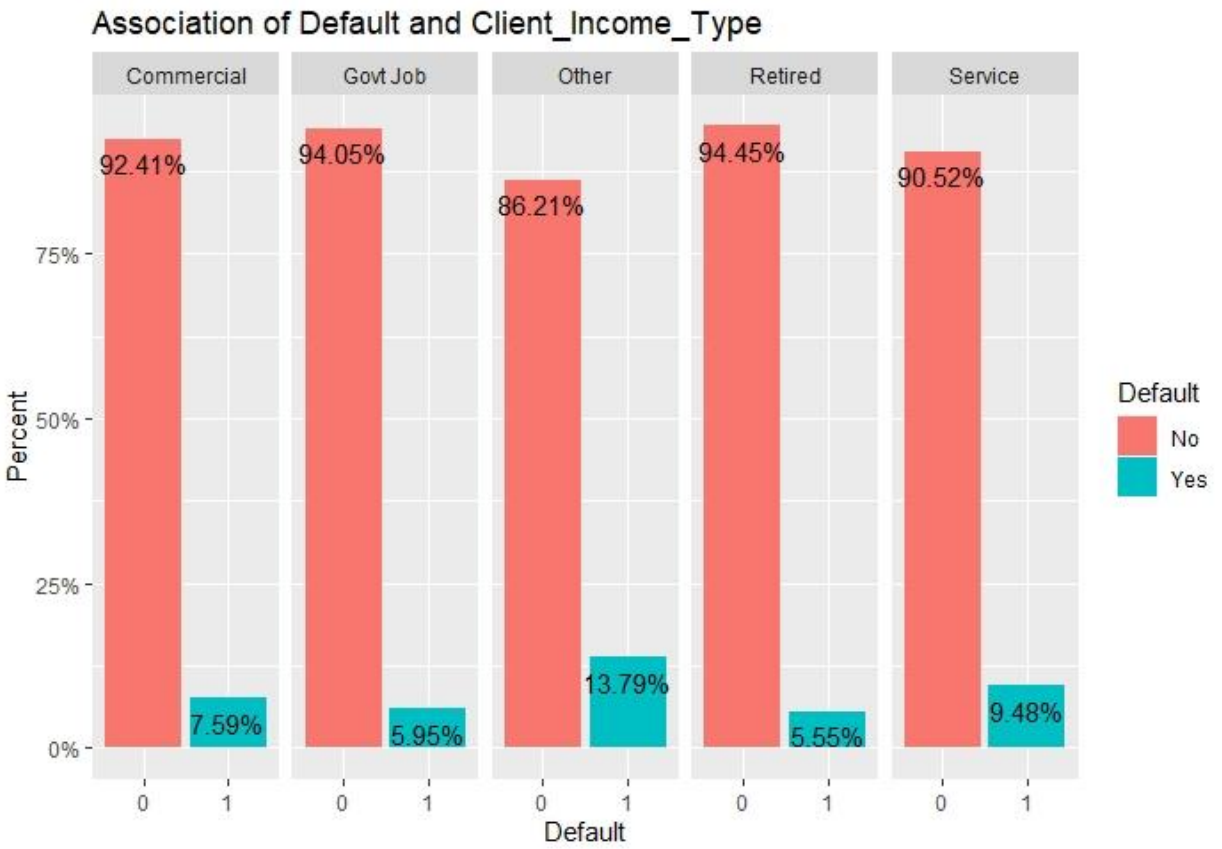
نقطه قوت این روش نسبت به روش رگرسیون لجستیک باینری، بکارگیری اندازه نمونه تمام دسته‌های نتایج در تخمین احتمال پارامترها و واریانس است، زیرا روش باینری تنها از اندازه نمونه دو دسته نتیجه در تخمین احتمال پارامترها و واریانس استفاده می‌کند. یکی از مهم‌ترین فرض‌های این مدل این است که پاسخ‌های نهایی از یکدیگر مستقل اند. در تجزیه و تحلیل رگرسیون چندگانه، بیش از یک مدل logit بر داده‌ها برازش می‌گردد، سپس برآیند تمامی مدل‌های logit ، مدل رگرسیون چندگانه را تشکیل می‌دهد و همه با هم برای پیش‌بینی احتمال هر نتیجه بکار گرفته می‌شوند.

۴ مصورسازی داده (Data Visualization)

در این بخش برای رسم تمامی نمودارها از دیتاست اول استفاده شده است. در نمودارهای Box plot برای نمایش بهتر نمودار، پیش از رسم، داده‌های پرت حذف شده اند.

۴-۱ رابطه میان Default و Client_Income_Type

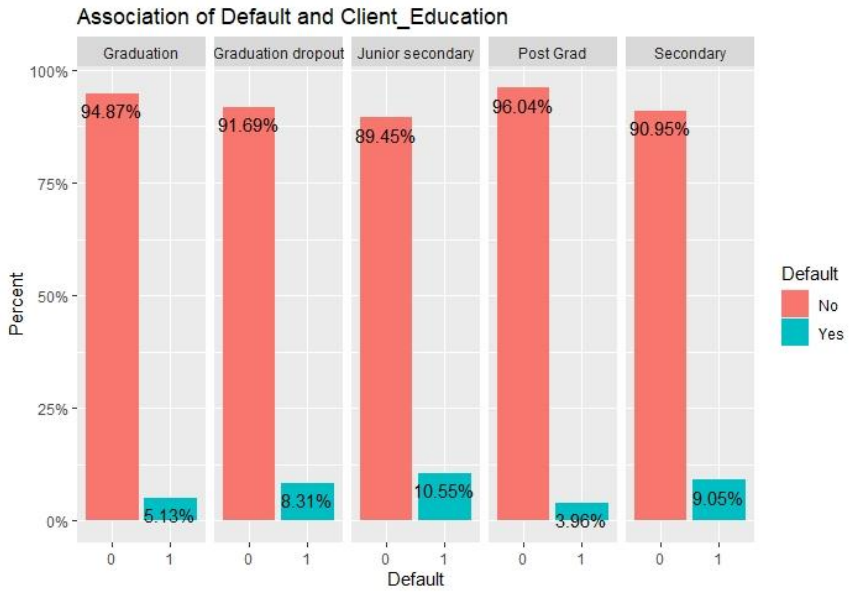
به نظر می‌رسد افراد دارای منبع درآمد خدماتی و یا تجاری درصد default بیشتری داشته اند.



شکل ٤-١ رابطه میان Default و Client_Income_Type

۴-۲ رابطه میان Default و Client_Education

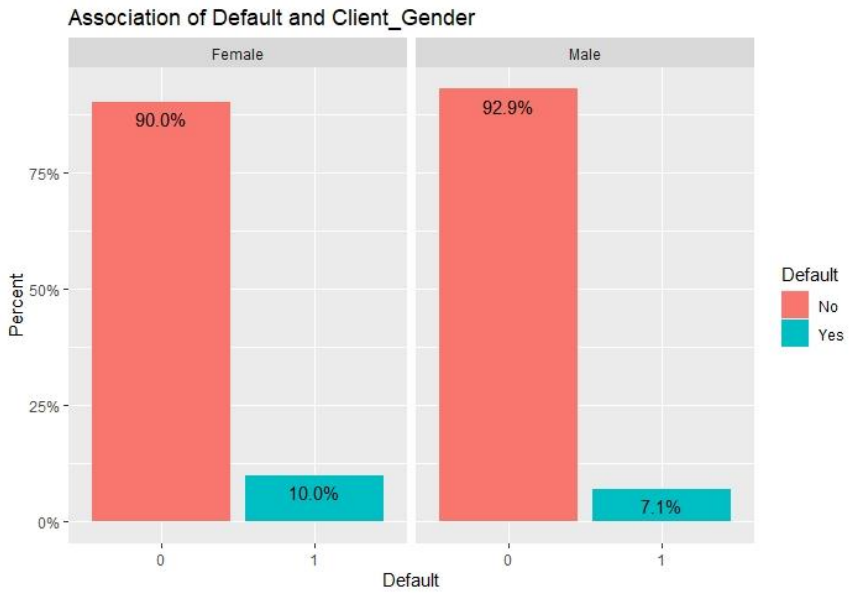
به نظر می‌رسد افراد دارای سطح تحصیلات Junior Secondary بیشترین درصد Default و افراد دارای سطح تحصیلات Post Grad دارای کمتری درصد default بوده اند.



شکل ٤-۲ رابطه میان Default و Client_Education

۴-۳ رابطه میان Default و Gender

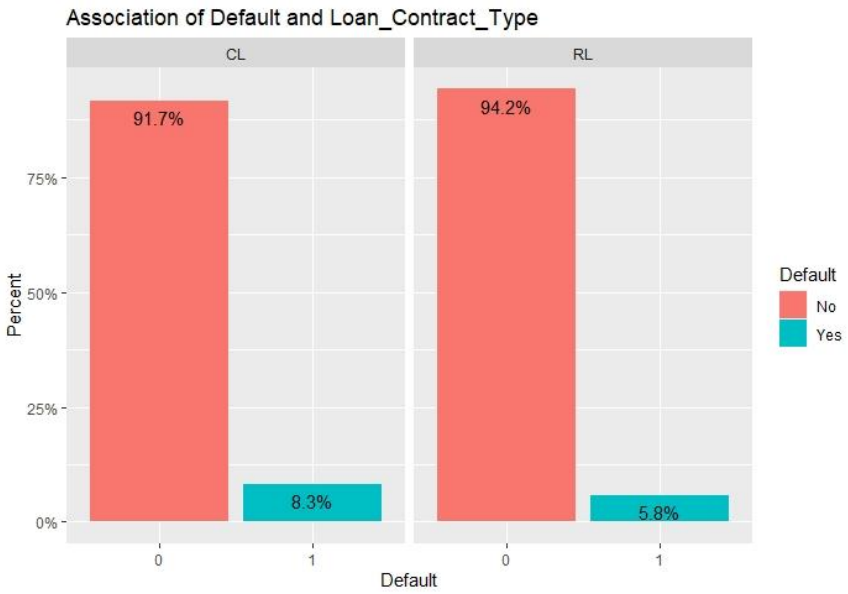
به نظر می‌رسد درصد default مردان اندکی بیشتر از زنان بوده است.



شکل ٤-۳ رابطه میان Default و Gender

۴-۴ رابطه میان Default و Loan_Contract_Type

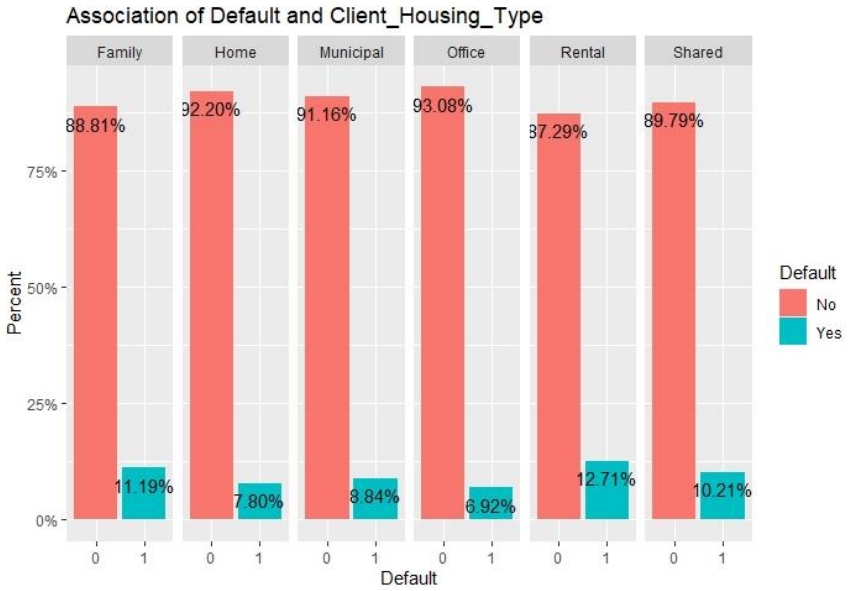
به نظر می‌رسد افراد دارای وام از نوع چرخشی درصد default کمتری داشته اند.



شکل ٤-٤ رابطه میان Default و Loan_Contract_Type

۵-۴ رابطه میان Default و Client_Housing_Type

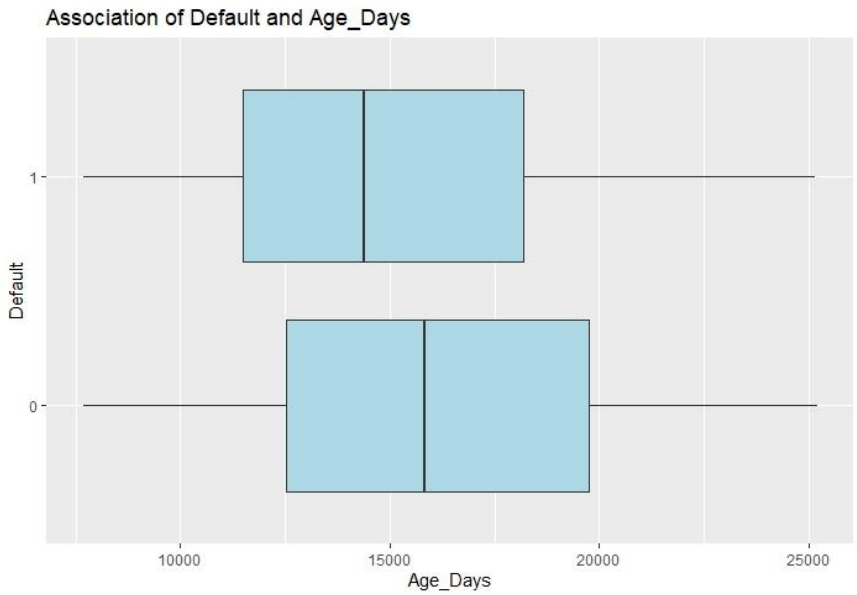
به نظر می‌رسد افراد اجاره نشین درصد default بیشتری داشته اند.



شکل ٤-٥ رابطه میان Default و Client_Housing_Type

۴-۶ رابطه میان Default و Age_Days

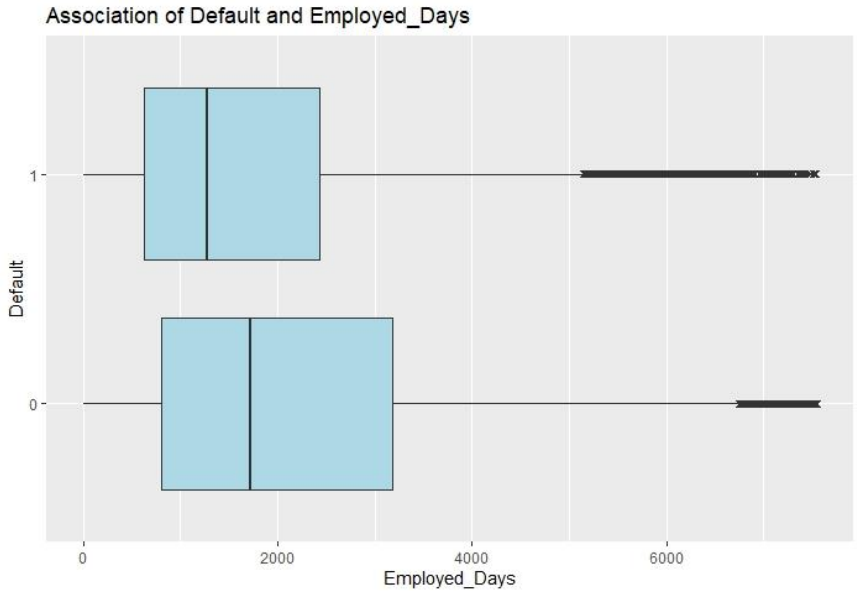
به نظر می‌رسد افرادی که default کرده اند، به طور متوسط جوان تر بوده اند.



شکل ٤-٦ رابطه میان Default و Age_Days

۴-۷ رابطه میان Default و Employed_Days

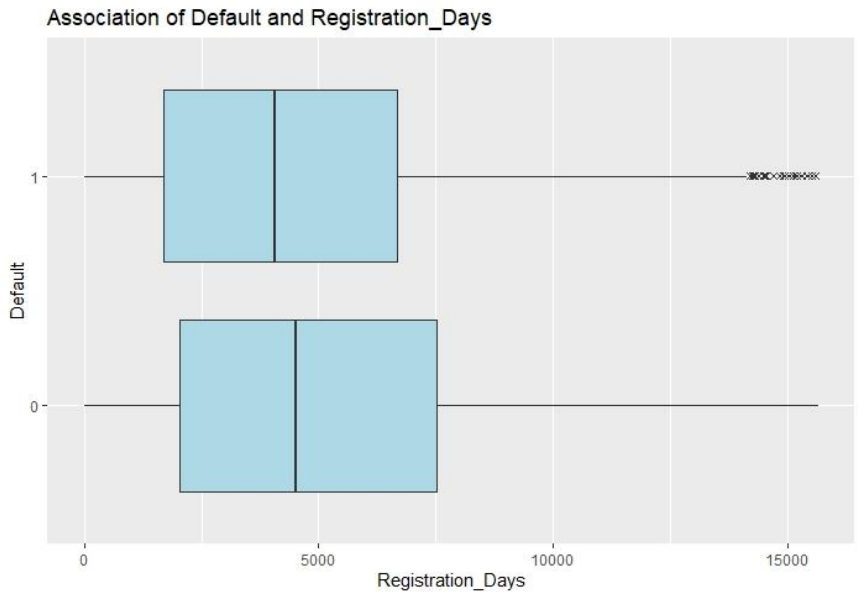
به نظر می‌رسد افرادی که default کرده اند، به طور متوسط سابقه کار کمتری داشته اند.



شکل ٤-٧ رابطه میان Default و Employed_Days

۸-۴ رابطه میان Default و Registration_Days

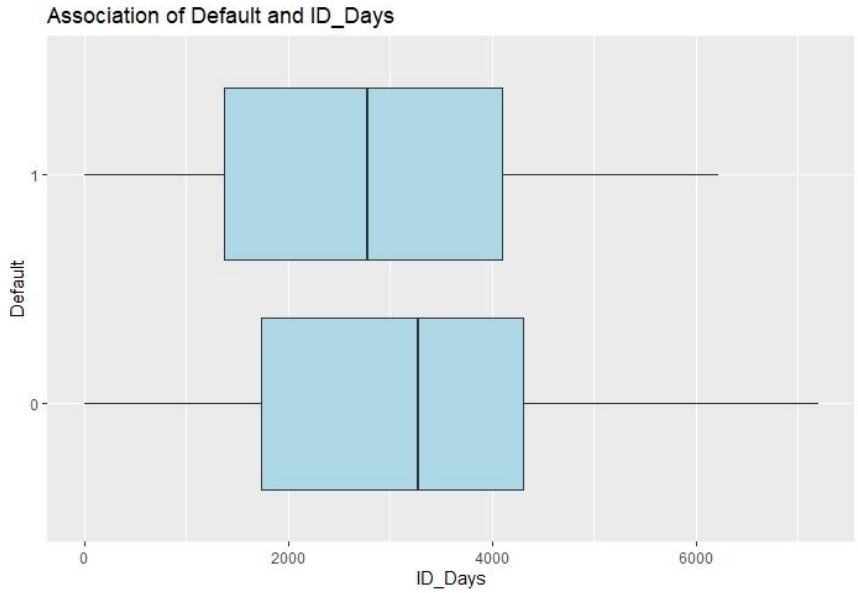
به نظر می‌رسد افرادی که default کرده اند، به طور متوسط عمر حساب بانکی آن‌ها کمتر بوده است.



شکل ٤-٨ رابطه میان Default و Registration_Days

۹-۴ رابطه میان Default و ID_Days

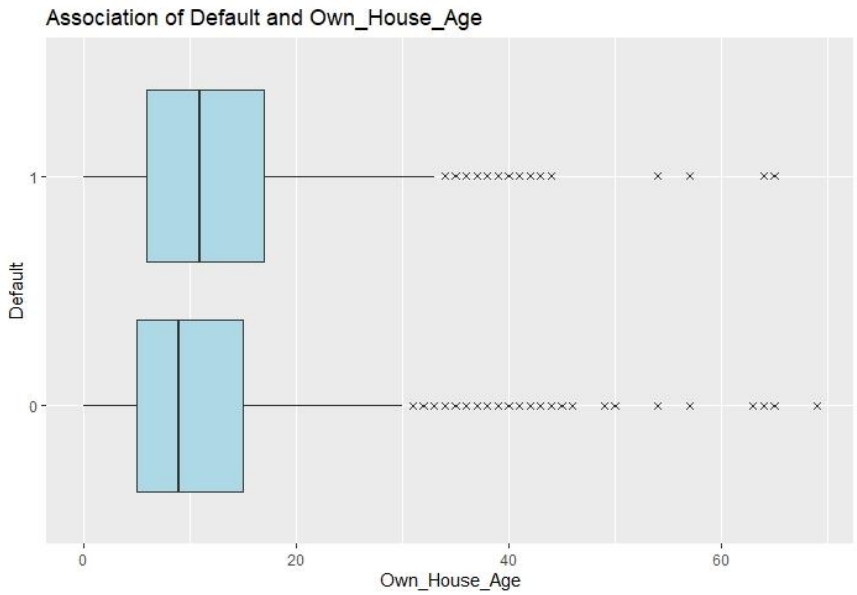
به نظر می‌رسد افرادی که default کرده اند، به طور متوسط مدارک هویتی خود را زودتر تغییر داده اند.



شکل ٤-٩ رابطه میان Default و ID_Days

۴-۱۰ رابطه میان Default و Own_House_Age

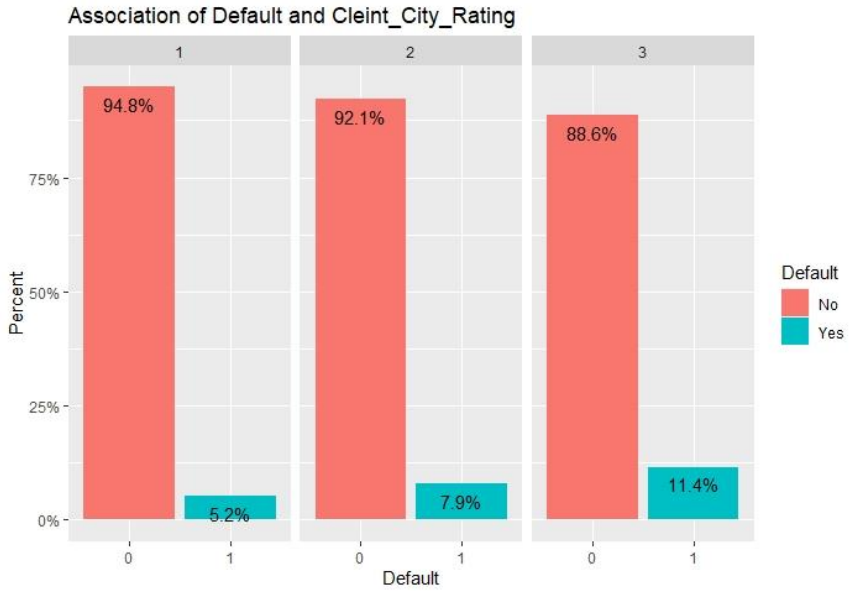
به نظر می‌رسد افرادی که default کرده اند، به طور متوسط سابقه مالکیت خانه بیشتری داشته اند.



شکل ٤-١٠ رابطه میان Default و Own_House_Age

۴-۱۱ رابطه میان Default و Client_City_Rating

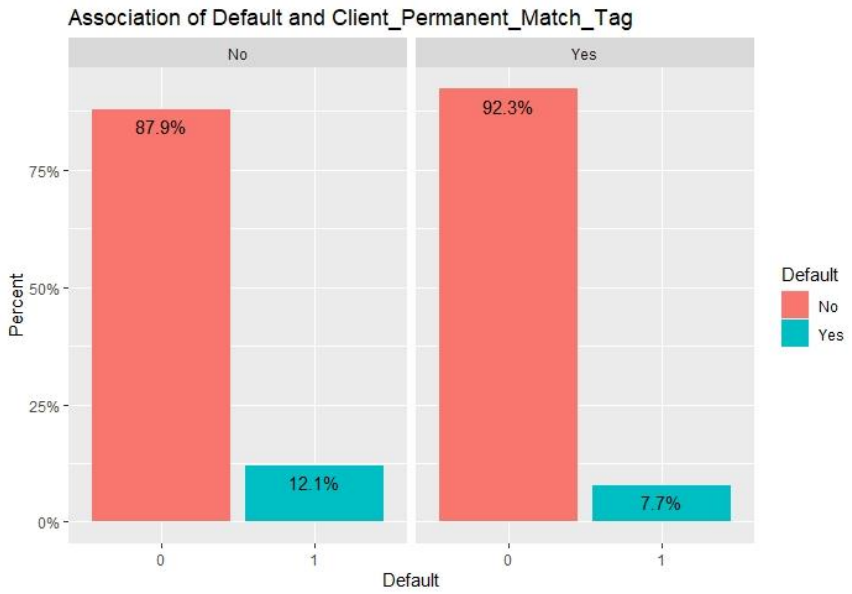
به نظر می‌رسد بالاتر بودن رتبه شهری رابطه عکس با default کردن داشته باشد.



شکل ٤-١١ رابطه میان Default و Client_City_Rating

۱۲-۴ رابطه میان Default و Client_Permanent_Match_Tag

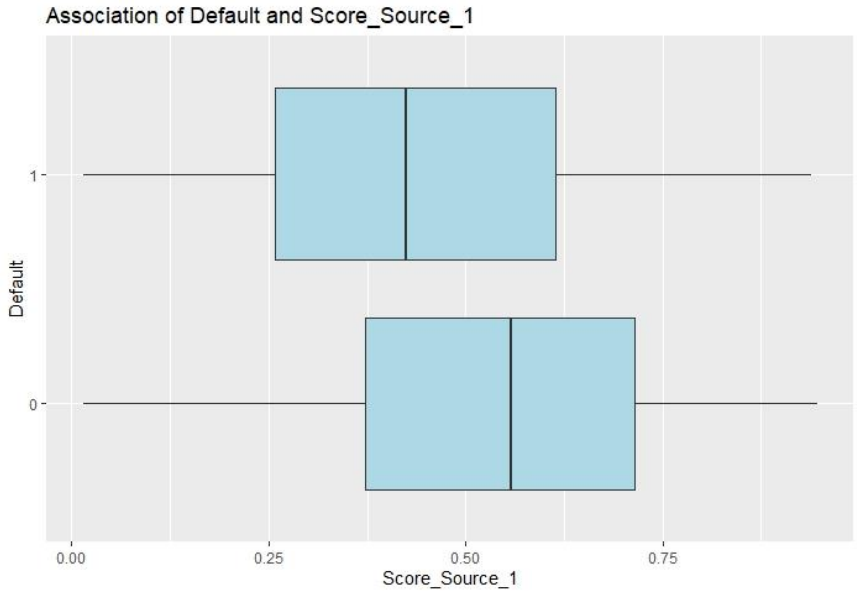
به نظر می‌رسد افرادی که مطابقت آدرس نداشته اند، بیشتر default کرده اند.



شکل ٤-۱۲ رابطه میان Default و Client_Permanent_Match_Tag

۱۳-۴ رابطه میان Default و Score_Source_1

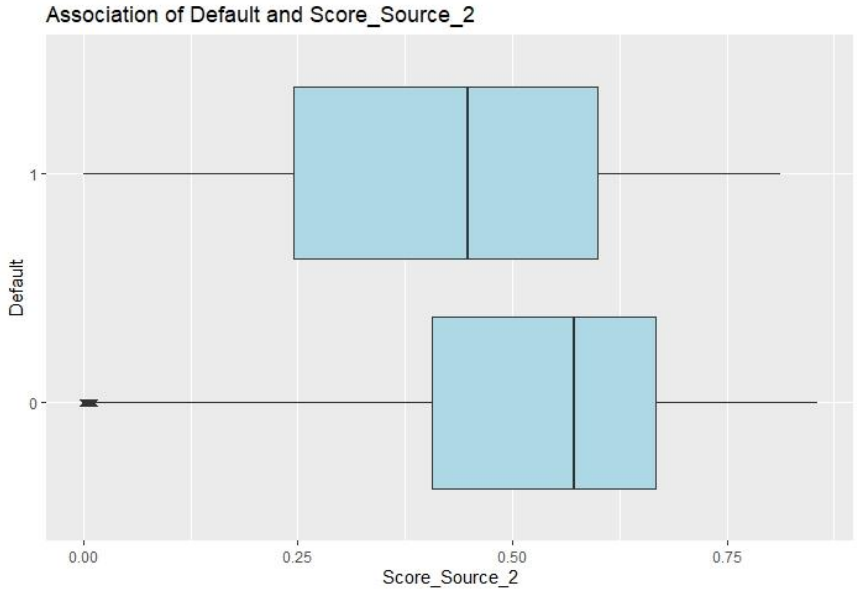
به نظر می‌رسد افرادی که default کرده اند، به طور متوسط مقدار Score_Source_1 کمتری داشته اند.



شکل ٤-۱۳ رابطه میان Default و Score_Source_1

۱۴-۴ رابطه میان Default و Score_Source_2

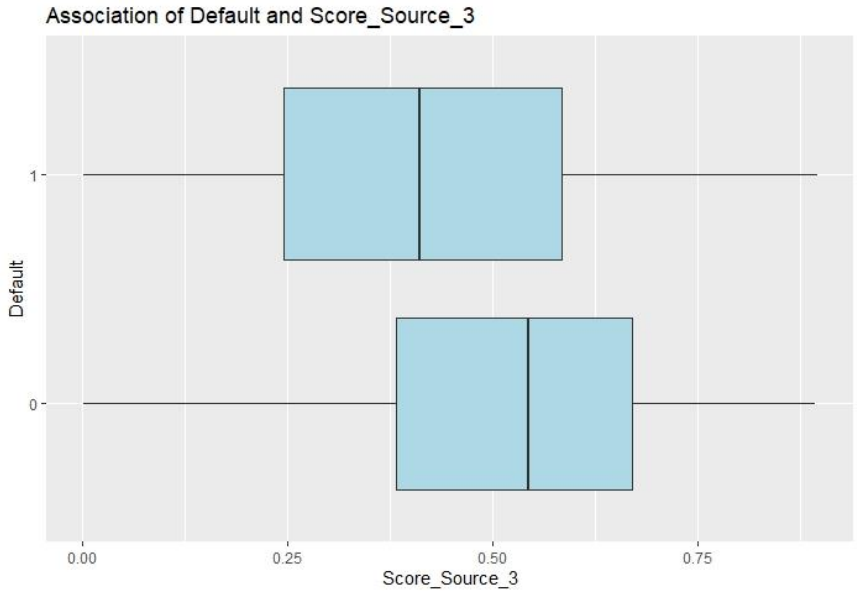
به نظر می‌رسد افرادی که default کرده اند، به طور متوسط مقدار Score_Source_2 کمتری داشته اند.



شکل ٤-١٤ رابطه میان Default و Score_Source_2

۱۵-۴ رابطه میان Default و Score_Source_3

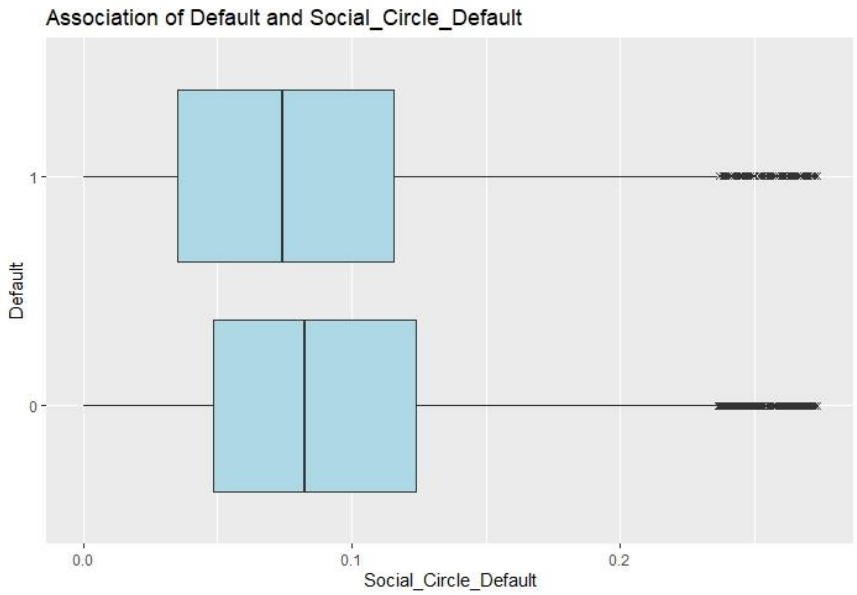
به نظر می‌رسد افرادی که default کرده اند، به طور متوسط مقدار Score_Source_3 کمتری داشته اند.



شکل ٤-١٥ رابطه میان Default و Score_Source_3

۱۶-۴ رابطه میان Default و Social_Circle_Default

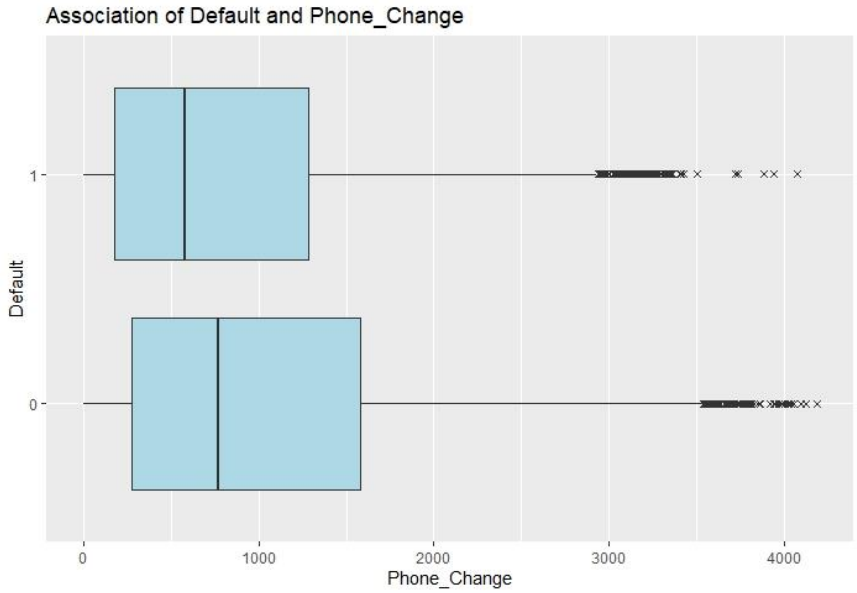
به نظر می‌رسد افرادی که default کرده اند، سابقه خانوادگی default کمتری داشته اند.



شکل ٤-١٦ رابطه میان Default و Social_Circle_Default

۱۷-۴ رابطه میان Default و Phone_Change

به نظر می‌رسد افرادی که default کرده اند، به طور متوسط شماره تلفن خود را زودتر تغییر داده اند.



شکل ٤-١٧ رابطه میان Default و Phone_Change

۵ آمار توصیفی (Descriptive Statistics)

در این بخش یا آزمون تاثیرگذار بودن یک عامل مورد بررسی قرار گرفته و یا این که یک آزمون ایجاد شده است.

۵-۱ رابطه میان Default و Client_Education

افرادی که سطح تحصیلات post grad دارند با احتمال بیشتری وام های خود را پرداخت می کنند نسبت به افرادی که سطح تحصیلات graduation دارند.

از تست proportion test استفاده می کنیم:

H0: P1 > P2

H1: P1 < P2

N1: تعداد افراد با سطح تحصیلات post grad

N2: تعداد افراد با سطح تحصیلات graduation

Win1: مجموع عدم بازپرداخت وام توسط افراد با سطح تحصیلات graduation

Win2: مجموع عدم بازپرداخت وام توسط افراد با سطح تحصیلات post grad

با انجام تست به نتیجه زیر میرسیم:

P1: 0.03960396

P2: 0.05127774

در نتیجه با توجه به تست انجام شده فرض صفر رد می شود و افرادی که سطح تحصیلات post grad دارند با احتمال کمتری وام های خود را پرداخت نمی کنند.

۵-۲ رابطه میان Default و Client_Gender

ادعا شده که جنسیت مشتریان بر روی پرداخت و عدم پرداخت قسط های وام موثر است. برای تست این فرضیه از آزمون مربع کای دو (استقلال) استفاده می کنیم:

ابتدا یک جدول با متغیرهای Gender و Defult ایجاد میکنیم و سپس از تست chisq.test استفاده می کنیم.

چون مقدار p-value برابر ۲.۲e-16 است و از ۰.۰۵بیشتر شده آزمون معنادار می شود.

۵-۳ آزمون میانگین درآمد

معاون بانک ادعا کرده که میانگین درآمد مشتریان درخواست کننده وام بیشتر از ۱۶۰۰۰ دلار می باشد.

H0: M <= 20000

H1: M > 20000

برای تست این فرض از آزمون نرمال z-test با فرض انحراف معیار برابر ۲ استفاده کنیم.

مقدار بحرانی برای تست درآمد را با ۰.۰۵ در نظر میگیریم و تست را با این فرضیات انجام می دهیم.

پس از آن با توجه به اینکه خود مسیله مقدار انحراف معیار را نداده و مجهول است از t.test استفاده کردیم و نتایج را مجددا بررسی کردیم.

همچنین فرض کردیم اگر به اندازه ۰.۲ اختلاف داشتیم میزان قدرت در تشخیص را اندازه میگیریم.

میزان قدرت به اندازه ۰.۹۷ در تشخیص می باشد که میزان قدرت مناسبی می باشد.

۵-۴ بازه اطمینان درآمد

در این بخش برای درآمد مشتریان می خواهیم بازه اطمینان محاسبه کنیم.

برای اینکار از نمونه های ۳۰ تایی استفاده میکنیم که ۱۰۰۰ بار نمونه گیری شود و میانگین مربوط به هر نمونه ذخیره شود. سپس مقدار Z استاندارد را بدست آورده و در فرمول بازه اطمینان جایگذاری میکنیم که فاصله اطمینان بصورت زیر می باشد:

(۲۰۱۴۲.۱۹, ۱۳۴۶۰.۸۹)

۵-۵ رابطه میان Default و Client_Housing_Type

ادعا شده افرادی که دارای خانه‌ی از نوع rental می باشند احتمال عدم بازپرداخت وام بیشتری در مقایسه با سایر مشتریان دارند.

H0: P1 < P2

H1: P1 > P2

P1: احتمال عدم بازپرداخت وام توسط افراد با خانه ی rental

P2: احتمال عدم بازپرداخت وام توسط افراد با خانه ی غیر از rental

N1: تعداد افراد با خانه ی rental

N2: تعداد افراد با خانه ی غیر از rental

Win1: مجموع عدم بازپرداخت وام توسط افراد با با خانه ی rental

Win2: مجموع عدم بازپرداخت وام توسط افراد با با خانه ی غیر از rental

برای انجام این آزمون فرض از proportion test استفاده میکنیم.

پس از انجام آزمون مقدارهای زیر بدست می آید:

P1: 0.12707775

P2: 0.08007267

با توجه به نتایج بدست آمده فرض صفر رد می شود و افرادی که دارای خانه ی اجاره ای هستند توان بازپرداخت وام کمتری نسبت به سایر افراد دارند.

۵-۶ آزمون سن

ادعا شده افرادی که دارای سن بیشتر از ٤١ سال (١٥٠٠٠ روز) می باشند با احتمال بیشتری بازپرداخت وام های خود را در مقایسه با افراد دارای سن کمتر از ٤١ سال انجام می دهند.

H0: P1 < P2

H1: P1 > P2

P1: احتمال بازپرداخت وام توسط افراد با سن بیشتر از ٤١ سال

P2: احتمال بازپرداخت وام توسط افراد با سن کمتر از ٤١ سال

N1: تعداد افراد با سن بیشتر از ٤١ سال

N2: تعداد افراد با سن کمتر از ٤١ سال

Win1: مجموع بازپرداخت وام توسط افراد با سن بیشتر از ٤١ سال

Win2: مجموع بازپرداخت وام توسط افراد با سن کمتر از ٤١ سال

برای انجام این آزمون فرض از proportion test استفاده میکنیم.

مقدار احتمال بازپرداخت برا یافراد با بیشتر از ٤١ سال سن برابر ۰.۹۲۸۱۱۵۲ و برای افراد با کمتر از ٤١ سال سن برابر ۰.۹۰۳۸۲۰۸ می باشد. در نتیجه فرض صفر رد می شود. اما همانطور که از عدد احتمال ها مشخص است تفاوت چشمگیری با یکدیگر ندارند.

۷-۵ رابطه میان Default و Client_City_Rating

طبق بررسی هایی که تیم این موسسات انجام دادند به این نتیجه رسیدند که اگر سطح شهر زندگی مشتریان متوسط باشد توان بازپرداخت وام آنها کاهش می یابد.

برای آزمون این فرض یکبار سطح شهر متوسط با خوب و یکبار سطح شهر متوسط با عالی را بررسی میکنیم.

سطح های متوسط و خوب:

H0: P1 > P2

H1: P1 < P2

P1: احتمال عدم بازپرداخت وام توسط افراد با سطح شهر متوسط

P2: احتمال بازپرداخت وام توسط افراد با سطح شهر خوب

N1: تعداد افراد با سطح شهر متوسط

N2: تعداد افراد با سطح شهر خوب

Win1: مجموع بازپرداخت وام توسط افراد با سطح شهر متوسط

Win2: مجموع بازپرداخت وام توسط افراد با سطح شهر خوب

برای انجام این آزمون فرض از proportion test استفاده میکنیم.

پس از انجام تست احتمال عدم بازپرداخت وام برای افراد در سطح شهر متوسط برابر ۰.۰۵٦۵۳۳۸۵ و برای افراد در سطح شهر خوب برابر ۰.۰٦۸۵٤۱۳٤می باشد و فرض صفر رد می شود.

سطح های متوسط و عالی:

P1: احتمال عدم بازپرداخت وام توسط افراد با سطح شهر متوسط

P2: احتمال بازپرداخت وام توسط افراد با سطح شهر عالی

N1: تعداد افراد با سطح شهر متوسط

N2: تعداد افراد با سطح شهر عالی

Win1: مجموع بازپرداخت وام توسط افراد با سطح شهر متوسط

Win2: مجموع بازپرداخت وام توسط افراد با سطح شهر عالی

پس از انجام تست احتمال عدم بازپرداخت وام برای افراد در سطح شهر متوسط برابر ۰.۰۵۱٦٤٦۲۷ و برای افراد در سطح شهر خوب برابر ۰.۱۱٤۱۸۳۸۷ می باشد و فرض صفر رد می شود.

۶ مدل‌های پیش‌بینی (Predictive Models)

سه مدل Logistic Regression ، Decision Tree و Random Forest برای پیش‌بینی در نظر گرفته شد. با توجه به این که دیتاست Test_Dataset متغیر Default را ندارد، لازم است برای بررسی عملکرد مدل‌های پیش‌بینی از دیتاست Train_Dataset استفاده شود. بنابراین، دیتاست آماده‌سازی شده به دو بخش test و train تقسیم شد. برای این منظور، نسبت تقسیم (Split Ratio) برابر با ۰.۸ در نظر گرفته شد که در نتیجه آن، ۹۷٤۸۵ داده به دیتاست train و ۲٤۳۷۱ داده به مجموعه تست اختصاص یافت. همچنین در تمامی مدل‌ها متغیرهای ID و Mobile_Tag از مدلسازی حذف شدند. تلاش شده برای آستانه مقداری انتخاب شود که تا حد امکان، مقدار Sensitivity و Specificity با هم برابر شوند، بدین ترتیب شرایط مقایسه مدل‌ها تقریبا برابر خواهد بود.

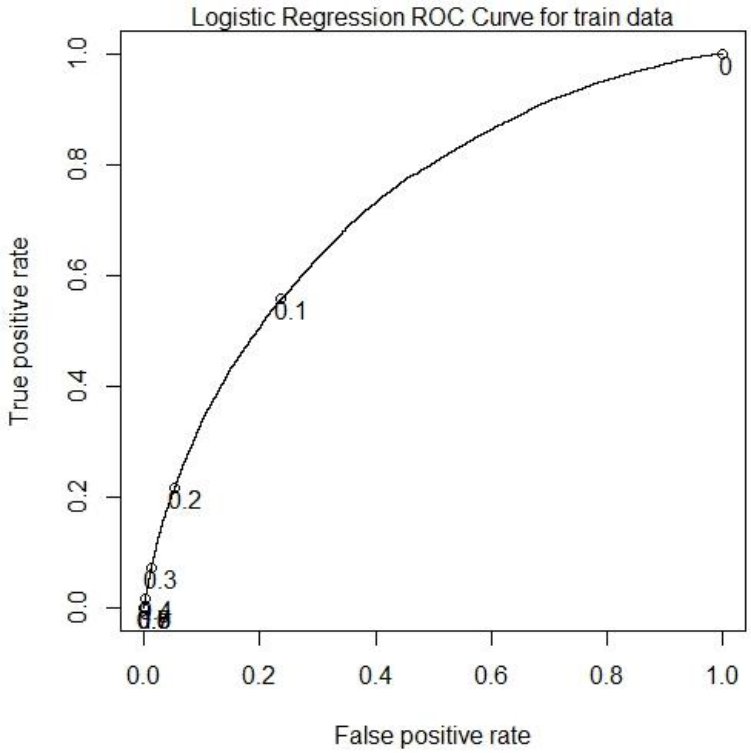
۶-۱ Logistic Regression

این مدل در ابتدا با بکارگیری همه متغیرها به جز ID و Mobile_Tag برازش شد، سپس متغیرهایی که موثر تشخیص داده نشده بودند، به ترتیب بزرگترین مقدار p-value از مدل حذف شدند. در خروجی مدل نهایی، مقدار AIC[10] که می‌توان گفت بیانگر خطای مدل است برابر با ٤۹۷۸۹ مشاهده شد . همچنین، با در نظر گرفتن آستانه برابر با ۰.۰۸، مقادیر accuracy برابر با ٦٦.۷۹٪ ، sensitivity برابر با ٦٦.۸۱٪ و specificity برابر با ٦٦.۵۳٪ برای داده‌های train بدست آمد.

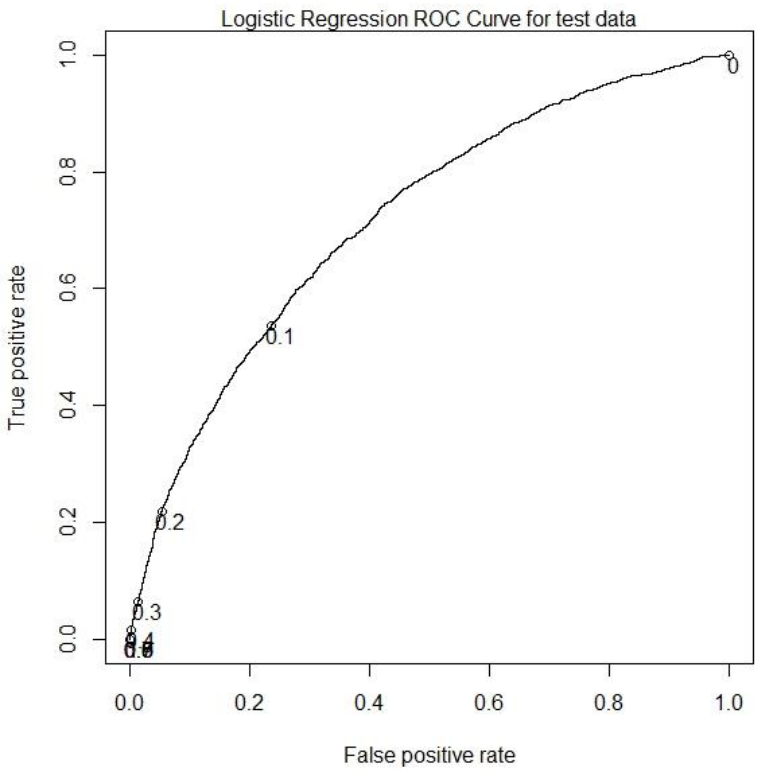
در ادامه، نمودار ROC [11] مدل نهایی برای داده‌های train، رسم شد که در شکل ٦-۱ نشان داده شده است.

با بکارگیری مدل برای داده‌های test و در نظر گرفتن آستانه ۰.۰.۸ ، مقادیر accuracy برابر با ٦٦.۸٦٪ ، sensitivity برابر با ٦۷.۰۳٪ و specificity برابر با ٦۵.۰۱٪ بدست آمد. نمودار ROC مدل برای داده‌های test درشکل ٦-۲ نشان داده شده است و سطح زیر نمودار ROC (AUC)[12] نیز برابر با ۷۱.۹۳٪ بدست آمده است.

[10] Akaike Information Criterion
[11] Receiver Operating Characteristics
[12] Area Under the Curve



شکل ٦-١ نمودار ROC مدل Logistic Regression برای داده‌های train

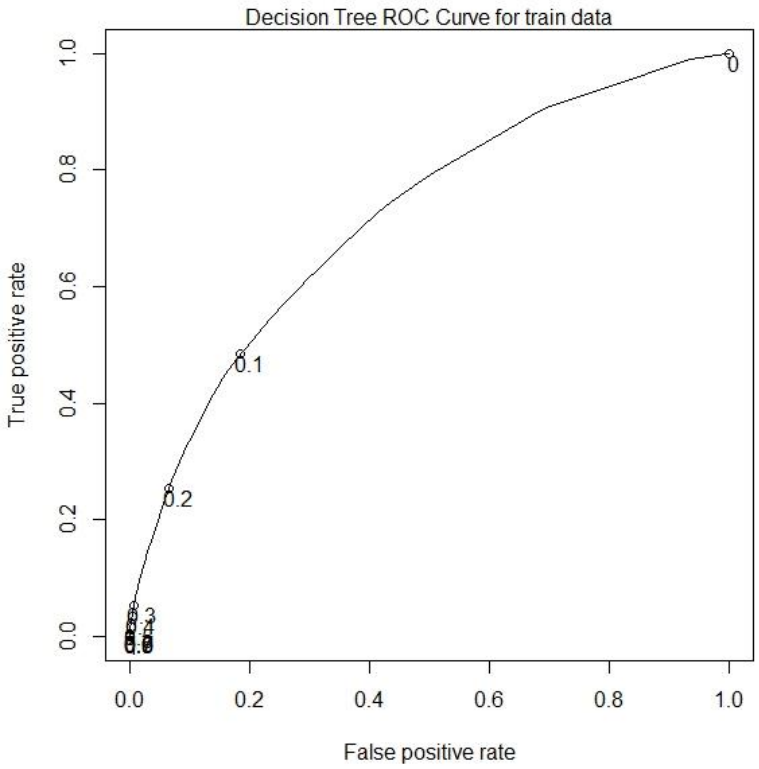


شکل ٦-٢ نمودار ROC مدل Logistic Regression برای داده‌های test

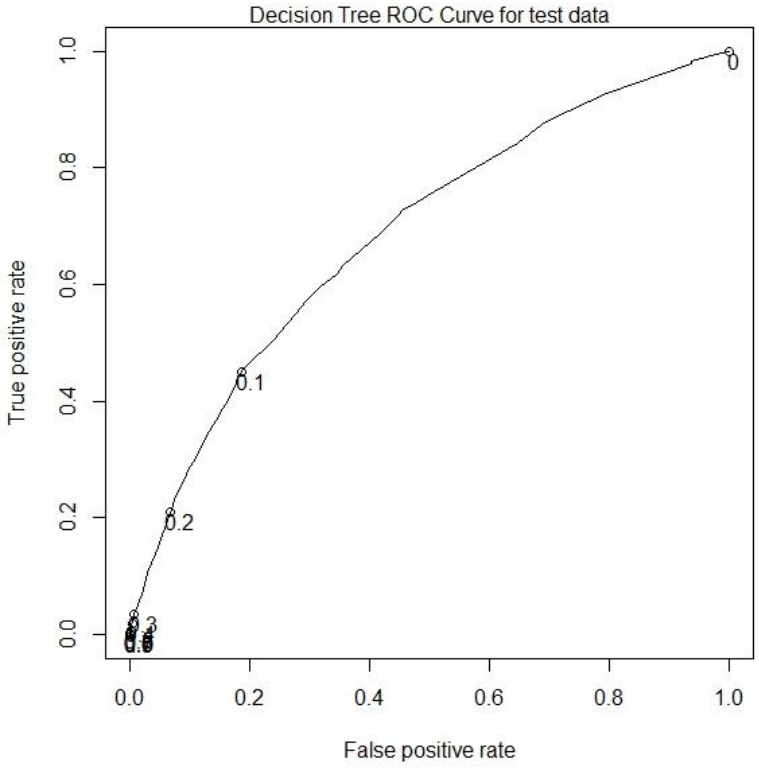
۲-۶ Decision Tree

پیش از بکارگیری این مدل داده های عددی استاندارد سازی شدند. در ادامه به دلیل تعداد بالای متغیرها مقدار پارامتر cp برابر با ۱- قرار داده شد تا مدل بر خلاف پیچیدگی بالا ایجاد شود، در ادامه تعداد شاخه زدن های مدل محدود شد تا یک نتیجه معقول حاصل شود. نتایج در جدول آمده است. همچنین نمودار این مدل با توجه به تعداد زیاد شاخه ها نمایش قابل فهمی ندارد.. با در نظر گرفتن مقدار آستانه برابر ۰.۰۸، معیارهای accuracy برابر با ٦٥.٧٤٪ ، sensitivity برابر با ٦٥.٧١٪ و specificity برابر با ٦٦.٠٦٪ برای داده‌های train بدست آمد. نمودار ROC مدل برای داده‌های train در شکل ٦-٣ رسم شده است.

با بکارگیری مدل برای داده‌های test ، معیارهای accuracy برابر با ٦٥.٠٦٪ ، sensitivity برابر با ٦٥.٣٤٪ و specificity برابر با ٦١.٨٦٪ بدست آمد. به نظر می‌رسد، این مدل کارایی معادل با مدل Logistic Regression داشته است. در ادامه، نمودار ROC مدل برای داده‌های test رسم شد که در شکل ٦-٤ نشان داده شده است، همچنین مقدار AUC، برابر با ٦٨.٧٠٪ بدست آمد که نسبت به مدل Logistic Regression بدتر است.



شکل ٦-٣ نمودار ROC مدل Decision Tree برای داده‌های train



شکل ٦-٤ نمودار ROC مدل Decision Tree برای داده‌های test

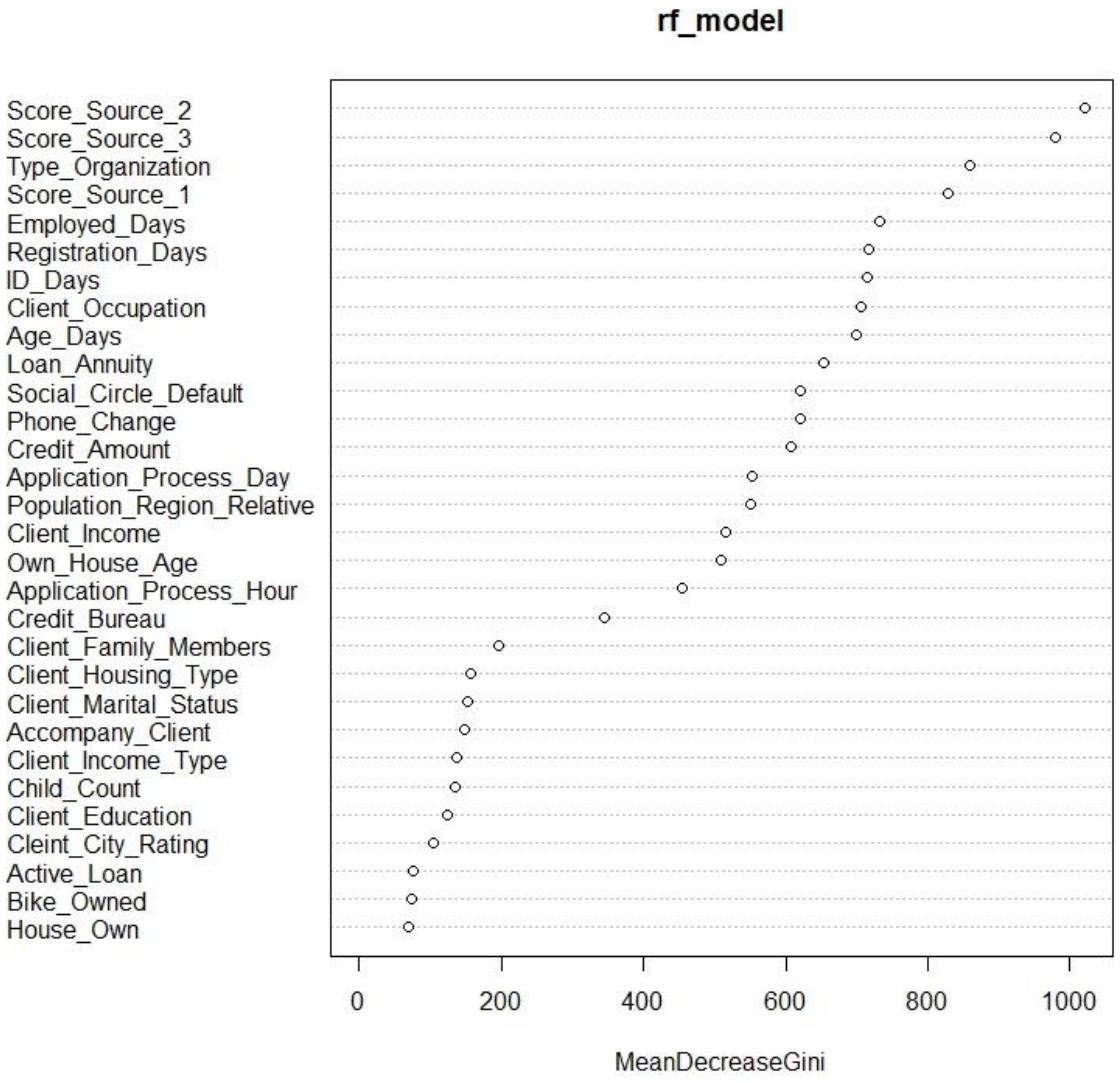
۳-۶ Random Forest

این مدل در ابتدا با بخش بزرگی از متغیرها تشکیل شد. با تشکیل مدل، نمودار اهمیت متغیرها مورد بررسی قرار گرفت. برای بهبود عملکرد مدل، کم اهمیت ترین متغیرها در نمودار اهمیت که در شکل ٦-٥ نشان داده شده است، به ترتیب حذف شدند تا جایی که دیگر پارامتر AUC بهبود نیابد. پس از حذف پنج متغیر روند بهبود مقدار AUC به پایان رسید و مدل نهایی با مقدار خطای OOB [۱۳] برابر با ۷.۷۱٪ ایجاد شد. با در نظر گرفتن مقدار آستانه برابر با ۰.۱۱ برای دیتاست train ، معیارهای accuracy برابر با ۷۰.۵۷٪ ، sensitivity برابر با ۷۰.۸٤٪ و specificity برابر با ٦۷.٤۲٪ بدست آمد. همچنین نمودار ROC مدل برای دیتاست train در شکل ٦-٦ رسم شده است.

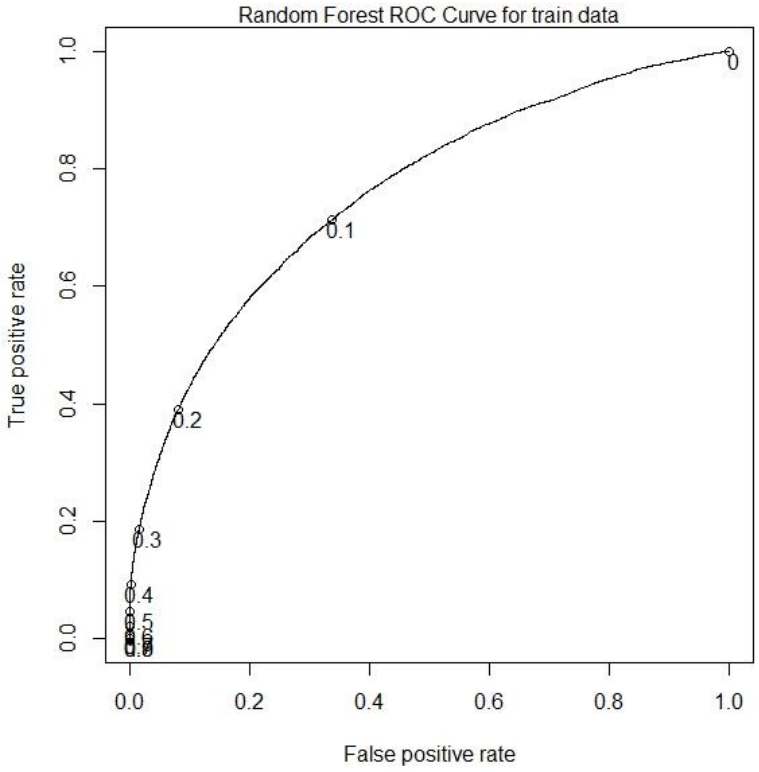
با بکارگیری مدل برای داده‌های test ، معیارهای accuracy برابر با ۷۰.٤٦٪ ، sensitivity برابر با ۷۰.۷۱٪ و specificity برابر با ٦۷.۷۰٪ بدست آمد. به نظر می‌رسد، کارایی این مدل بر خلاف دو مدل پیشین، بر

[۱۳] Out Of Bag Error

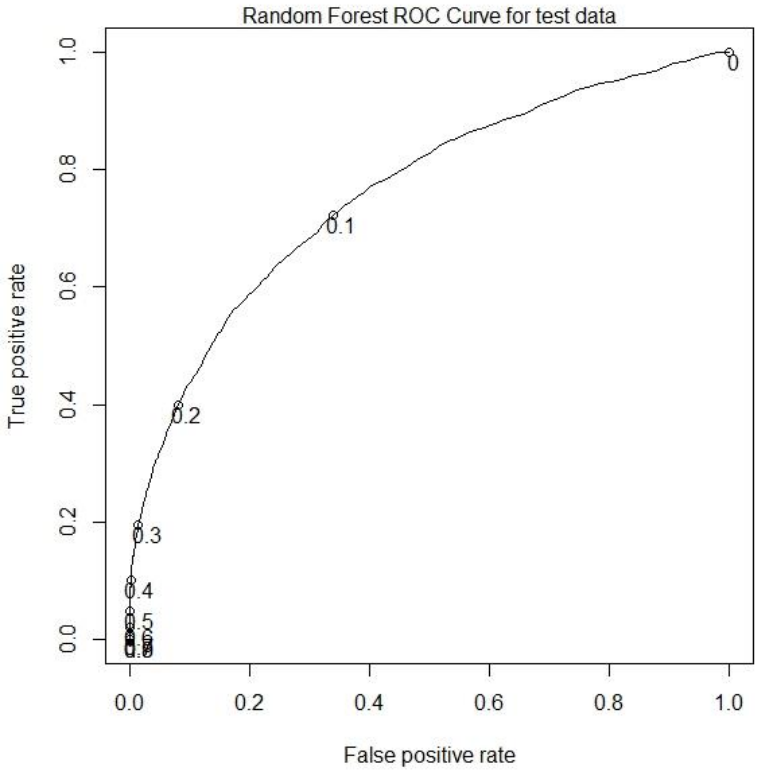
روی داده‌های test بهتر بوده است. در ادامه، نمودار ROC مدل برای داده‌های test رسم شد که در شکل ٦-٧ نشان داده شده است، همچنین مقدار AUC، برابر با ٧٦.١٨٪ بدست آمده که به وضوح عملکرد بهتری را نسبت به دو مدل پیشین نشان می‌دهد.



شکل ٦-٥ نمودار اهمیت متغیرها در مدل Random Forest



شکل ٦-٦ نمودار ROC مدل Random Forest برای داده‌های train



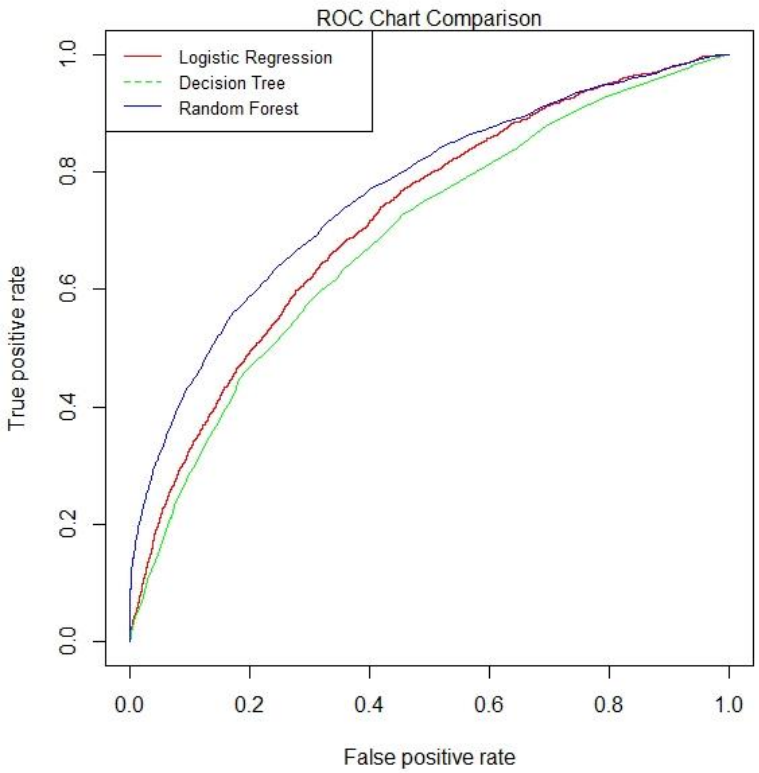
شکل ٦-٧ نمودار ROC مدل Random Forest برای داده‌های test

۷ نتیجه گیری

مهم‌ترین معیارهای مربوط به سه مدل پیش‌بینی در جدول ۲ آورده شد هاست. همچنین منحنی مربوط به مدل‌های ارائه شده در مقایسه شده است. در نتیجه می‌توان گفت مدل Random Forest عملکرد بهتری داشته است.

جدول ۲ – مقایسه معیارهای مدل‌های پیش‌بینی ارائه شده

AUC	Specificity	Sensitivity	Accuracy	مدل
۷۱.۱۹٪	٦٥.٠١٪	٦٧.٠٣٪	٦٦.٨٦٪	Logistic Regression
٦٨.٧٠٪	٦١.٨٦٪	٦٥.٣٤٪	٦٥.٠٦٪	Decision Tree
٧٦.١٨٪	**٦٧.٧٠٪**	**٧٠.٧١٪**	**٧٠.٤٦٪**	**Random Forest**



شکل ۷-۱ منحنی ROC مدل‌های ارائه شده

۸ مراجع

[1] O. MONK, "Automobile Loan Default Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/saurabhbagchi/dish-network-hackathon. [Accessed 28 06 2022].

[2] Anonymous, "Exploratory Data Analysis," JMP Statistical Discovery LLC, [Online]. Available: https://www.jmp.com/en_hk/statistics-knowledge-portal/exploratory-data-analysis.html. [Accessed 31 07 2022].

[3] M. Restori, "What is Exploratory Data Analysis," Charito, [Online]. Available: https://chartio.com/learn/data-analytics/what-is-exploratory-data-analysis/. [Accessed 31 07 2022].

[4] N. Tamboli, "All You Need To Know About Different Types Of Missing Data Values And How To Handle It," Analytics Vidhya, 25 07 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/. [Accessed 31 07 2022].

[5] P. Bhandari, "Missing Data | Types, Explanation, & Imputation," Scribbr, 08 12 2021. [Online]. Available: https://www.scribbr.com/statistics/missing-data/. [Accessed 31 07 2022].

[6] Anonymous, "Tutorial on 5 Powerful R Packages used for imputing missing values," Analytics Vidhya, 05 07 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/#:~:text=MICE%20Package,of%20uncertainty%20in%20missing%20values.. [Accessed 31 07 2022].

[7] Anonymous, "mice: mice: Multivariate Imputation by Chained Equations," RDocumentation, [Online]. Available: https://www.rdocumentation.org/packages/mice/versions/3.14.0/topics/mice. [Accessed 31 07 2022].

[8] J. Josse, "Handling missing values with R," Julie Josse, [Online]. Available: http://juliejosse.com/wp-content/uploads/2018/06/DataAnalysisMissingR.html. [Accessed 31 07 2022].

[9] S. Buuren, " Predictive mean matching," Stevan Buuren, [Online]. Available: https://stefvanbuuren.name/fimd/sec-pmm.html. [Accessed 31 07 2022].

[10] Anonymous, "Extensions to Multinomial Regression," Columbia Public Health, 07 07 2022. [Online]. Available: https://www.publichealth.columbia.edu/research/population-health-methods/extensions-multinomial-regression#:~:text=Multinomial%20(Polytomous)%20Logistic%20Regression&text=In%20polytomous%20logistic%20regression%20analysis,outcome%20is%20compared%20to%20it. [Accessed 31 07 2022].