**فهرست مطالب:**

1. شرح دقیق کاربرد سامانه هوشمند و مساله‌ای که حل می‌کند.
2. ارزش افزوده و جذابیت‌های ایجاد شده در محصول بواسطه استفاده از هوش مصنوعی
3. میزان تاثیرگذاری هوش مصنوعی در چرخه توسعه محصول
4. سابقه و میزان مرسوم بودن استفاده از مؤلفه‌های هوشمند در کاربرد مشابه (در سطح ملی و بین المللی)
5. روش اجرایی:
   1. موجود بودن کد منبع آموزش و مجموعه داده آموزشی در شرکت
   2. متناسب بودن حجم دادگانی که در فرآیند مدلسازی، ارزیابی و اعتبارسنجی مورد استفاده قرار گرفته است.
   3. برخورداری از ابزار و فرایندهای مدیریت کیفیت داده.
   4. ذکر نوع (یا نام) الگوریتم‌های هوشمند مورد استفاده.
   5. میزان تناسب الگوریتم انتخاب شده با کاربرد هدف گذاری شده
   6. توصیف پذیری (Explainability) عملکرد هوشمند
   7. گستره دامنه مورد پشتیبانی /میزان قابلیت تعمیم الگوریتم.
   8. ارائه کردن دقیق دامنه ورودی‌ای که عملکرد الگوریتم هوشمند در آن تضمین شده است.
   9. برخورداری از توانایی آموزش، توسعه و سفارشی سازی الگوریتم‌های هوشمند و معماری آنها
   10. شاخص‌های کمی و کیفی مورد استفاده برای اندازه‌گیری کارایی الگوریتم هوش مصنوعی و کیفیت خروجی محصول (مثلا دقت ترجمه)
   11. آشنایی، شناخت و استفاده از معیارهای ارزیابی کارآیی، دقت، مقاومت و … متناسب با کاربرد
   12. برخورداری از فرآیند ارزیابی تخصصی الگوریتم‌های هوشمند (ترجیحا به شکل خودکار)
   13. مقاومت Robustness سامانه (در برابر نویز و تخریب ورودی)
   14. برخورداری از فرآیند تضمین کیفیت الگوریتم‌های هوشمند برای حالتهای پیش‌بینی شده و پیش‌بینی نشده (ترجیحا به شکل خودکار)
   15. لحاظ کردن جنبه های قابلیت اتکا(Dependable AI Aspects)
6. **شرح دقیق کاربرد سامانه هوشمند و مساله ای که حل می‌کند:**

ریسک نقدینگی ناشی از ناتوانی بانک در بازپرداخت به موقع بدهی ها، انجام تعهدات یا ناتوانی در گسترش سبد دارایی های پربازده با هزینه های متعارف است. به عبارت دیگر، زمانی که بانکی نقدینگی کافی نداشته باشد، نمی تواند با افزایش بدهی یا تبدیل دارایی ها وجوه کافی را به سرعت و با قیمت مناسب تامین کند که این امر بر سودآوری بانک تاثیر می گذارد. می توان گفت علت اصلی ریسک نقدینگی در بانک ها عدم تطابق ارزش و سررسید بدهی ها و دارایی ها و در نتیجه ایجاد شکاف نقدینگی منفی است.

محاسبه نرخ بهره شامل ایجاد نمودار سررسید ترازنامه و محاسبه مازاد یا کسری (نرخ بهره) است. این تفاوت با تحلیل وضعیت مالی آتی بانک بر اساس وضعیت جاری دارایی ها، بدهی ها و اقلام غیرجاری آتی (در صورت وجود) محاسبه می شود. چنین مطالعاتی زمانی را برای مدیریت کسری نرخ بهره قبل از وقوع هر اتفاقی در آینده فراهم می کند. در اين ارتباط بانكها مي بايست شكاف نقدينگي را براي يك يا 3 ماه بعد، شش ماه بعد و يا براي يك سال آتي محاسبه كنند. همانطور كه اشاره گرديد، ساخت جدول شكاف نقدينگي بانك مستلزم تدوين سناريوهاي مختلفي است. در هر سناريويي، بانك بايد هر نوع جريان مثبت يا منفي نقدينگي را كه ميتواند رخ دهد، مد نظر قرار دهد. در واقع با پیش‌بینی میزان منابع و پیش‌بینی میزان تسهیلات مشتریان در دوره‌های آتی می‌توان به پیش‌بینی ریسک نقدینگی با اطمینان بالایی اقدام کرد.

در حال حاضر و در سناريوي شرايط نرمال و در خصوص جريانات وجوه، فرض می‌شود كه بسياري از وام‌هاي در حال سررسيد در زمان مقرر بازپرداخت شده و يا سپرده‌هاي بانكي به سهولت جايگزين شده و سپرده‌هاي قبلي تجديد مي‌شوند که این خود یک فرض بسیار ساده است و در واقعیت این امر به این سادگی محقق نمی‌شود لذا نیاز است تا ضمن شناسایی عوامل موثر بر میزان منابع و مانده تسهیلات مشتریان در آینده، با روش‌ها و الگوریتم‌های پیش‌بینی کننده نسبت به تخمین این شاخص‌ها به ازای هر مشتری در آینده اقدام کرد.

یکی از مهم ترین ریسک هایی که بانک ها و موسسات اعتباری سپرده گذاری با آن مواجه هستند، ریسک نقدینگی است که ناشی از عدم تطابق موقت دارایی ها و بدهی های آنهاست. بانک‌ها و مؤسسات اعتباری باید یک سیستم مدیریت ریسک نقدینگی مناسب داشته باشند تا علاوه بر نظارت و کنترل ریسک نقدینگی، از قدرت و ظرفیت نقدینگی کافی برای ایفای مؤثر نقش «واسطه‌گری پولی عادی و متعارف» برخوردار باشند.

ریسک نقدینگی یک تهدید مالی ویرانگر برای بانک ها است که در صورت ارزیابی نادرست یا نادیده گرفته شود می تواند عواقب جبران ناپذیری داشته باشد. کنترل بهینه پدیده ای مانند ریسک نقدینگی نیازمند یک روش اندازه گیری دقیق است که در صورت ارزیابی نادرست یا سهل انگاری ممکن است منجر به عوارض جبران ناپذیری شود. کنترل بهینه پدیده ای مانند ریسک نقدینگی ، به یک روش اندازه گیری دقیق نیاز دارد.

از سوی دیگر، در بانك‌ها ممكن است از نسبت‌هاي متنوعي براي مديريت نقدينگي بهره گرفته شود اما نكته قابل توجه آن است كه درباره اكثر نسبت‌ها (به جز نسبت‌هاي تعريف شده از سوي كميته بال) رقم استانداردي وجود نداشته و هر بانكي به تناسب ساختار، ويژگي‌ها و شرايط اقتصادي پيرامون خود، رقم خاصي را به عنوان نسبت مطلوب در نظر مي‌گيرد که ممکن است در طول زمان هم متغیر باشند.

همانظور که ذکر شد، گرچه بانک‌ها نسبت‌ها و روش‌های مختلفی جهت محاسبه میزان ریسک نقدینگی خود دارند اما برخي از مهمترين و متداول‌ترین نسبت‌ها به شرح زيرند:

1. نسبت تسهيلات به سپرده ها
2. نسبت تمركز سپرده (100 سپرده گذار عمده)
3. نسبت تسهيلات بلندمدت به سپرده هاي كوتاه مدت
4. نسبت سپرده‌هاي ديداري (حساب جاری) به كل سپرده‌ها
5. نسبت مجموع سپرده‌ها به كل بدهي‌ها
6. نسبت مصارف به منابع

با این حال، ارائه تعریف مناسب برای آن یک مانع جدی است. علاوه بر این، مسئله تعیین عوامل موثر و مرتبط و فرموله بندی فرم مناسب تابعی برای تخمین‌زنی و پیش‌بینی آن، کاری دشوار و پیچیده است.

در این پروژه سعی شده است تا پیش‌بینی میانگین منابع مشتریان به تفکیک نوع حساب (جاری، قرض الحسنه، کوتاه مدت و بلند مدت) در 3 ماه آینده و نیز میانگین میزان تسهیلات مشتریان در 3 ماه آینده با استفاده از تکنیک‌های یادگیری ماشین از مدل‌های سنتی تا روش‌های جدیدتر مانند یادگیری عمیق و شبکه‌های عصبی بازگشتی صورت گیرد. پس از پیش‌بینی این مقادیر، نسبت‌های مذکور در ردیف‌های 1 تا 4 در دوره 3 ماه آینده تخمین زده شود.

همچنین در صورتیکه مقادیر مصارف و بدهی‌های بانک برای کاربر سامانه مشخص باشد، با توجه به پیش‌بینی میزان منابع بانک در 3 ماه آینده، ردیف‌های 5 و 6 نیز قابل تخمین اند.

1. **ارزش افزوده و جذابیت‌های ایجاد شده در محصول بواسطه استفاده از هوش مصنوعی**

مزایای حاصل از اجرای این طرح، در حوزه مدیریت و پیش‌بینی ریسک نقدینگی و به تبع کاهش این ریسک به عنوان سیتمی کمک تصمیم برای مدیران است. مزایای این پروژه بصورت خلاطه بدین ترتیب است:

* پیش‌بینی ریسک نقدینگی بانک با توجه به 4 شاخص اصلی (نسبت‌های 1 تا 4 ذکر شده)
* اجرای دستورالعمل بانک مرکزی در حوزه مدیریت ریسک نقدینگی
* شناخت وضعیت موجود و شکاف تحلیل

از طرفی چون موتور تحلیلی این پروژه با ریزدانگی مشتری عمل می‌کند، لذا از سایر دستاوردها و مزایای آن بدین صورت می‌توان نام برد:

* پیش‌بینی منابع مشتریان به تفکیک نوع حساب و بالطبع پیش‌بینی رویگردانی مشتریان از منظر منابع بانکی
* پیش‌بینی میزان بدهکاری (خوش حسابی) مشتری در آینده و ایجاد سیستم هشدار سریع بمنظور کاهش احتمال نکول تسهیلات
* اعتبارسنجی مشتریان و محاسبه ریسک بازپرداخت با توجه به مبلغ تسهیلات
* کمک به تدوین مطلوب‌تر برنامه‌ریزی‌های بلند مدت و کوتاه مدت با توجه به نقدینگی و درآمدهای بانک

1. **میزان تاثیر هوش مصنوعی در چرخه توسعه محصول**
2. **کارهای پیشین**

تاکنون در ایران تحقیق جامعی درباره امکان استفاده از شاخص‌های حسابداری و شبکه عصبی برای مدل‌سازی برای پیش‌بینی ریسک نقدینگی در بانک‌های دولتی انجام نشده است که یکی از دلایل آن ممکن است دسترسی محدود به اطلاعات و داده‌های لازم برای تحقیق باشد. زیرا متأسفانه این اطلاعات به صورت شفاف منتشر نمی‌شود (Ansor, 1400, 191) و جمع‌آوری این اطلاعات و داده‌ها کاری دشوار و سخت است، اما مطالعات مختلفی در این زمینه انجام شده است، به عنوان مثال یانو سانگ (2022) در تحقیقی با استفاده از یکی از مدل‌های شبکه عصبی پیشرفته و با استفاده از ۲۵ شاخص ریسک نقدینگی به‌عنوان متغیرهای ورودی مدل، عوامل مؤثر بر ریسک نقدینگی بانک‌های تجاری شناسایی و یک مدل موفقیت برای هشدار ریسک نقدینگی تهیه شد. در سال 2020 و منطقه جغرافیایی مورد مطالعه، نمونه ای منتخب از بانک های تجاری چین Mishraz et al (2021) در مطالعه ای به مدت 5 سال از سال 2015 تا 2019 با داده های حسابداری 75 بانک هندی، عملکرد مصنوعی بود. مدل‌های شبکه عصبی، تحلیل تشخیصی، لجستیک و خطی (LDA) را با یکدیگر مقایسه کردند، نتایج تحقیق نشان داد که مدل‌های شبکه عصبی مصنوعی دقیق‌تر از تحلیل خطی هستند.

در سال 2020، کریم الدحیدهی و همکاران در مطالعه ای با استفاده از داده های حسابداری 16 بانک عراقی جمع آوری شده بین سال های 2004 تا 2018، توانستند ریسک های مالی (ریسک نقدینگی، ریسک اعتباری و ریسک سهام) را با استفاده از مدل های شبکه عصبی مصنوعی پیش بینی کنند.

دوانا و همکاران (2018) یک مدل ارزیابی ریسک نقدینگی با استفاده از شبکه عصبی مصنوعی و شبکه بیزی طراحی کردند. در این تحقیق از روش شبکه عصبی مصنوعی با استفاده از اطلاعات ماهانه از دفتر کل بانک های دولتی طی هشت سال و با استفاده از 10 نسبت نقدینگی مانند نسبت جاری، نسبت جاری و سایر نسبت های انتخابی محققین استفاده شده است. پتانسیل شبکه برای پیش بینی ریسک نقدینگی در بانکداری مورد آزمایش قرار گرفت، نتایج این تحقیق نشان دهنده عملکرد بالا، دقت و انعطاف پذیری شبکه های عصبی مصنوعی در مدل سازی ریسک نقدینگی بود.

فرامرزی و همکاران (1395) در مطالعه خود به مدل‌سازی ریسک نقدینگی بانک خصوصی با استفاده از شاخص‌های حسابداری پرداختند و نتایج پژوهش توانایی مدل را در پیش‌بینی ریسک نقدینگی در یک نمونه انتخابی نشان داد.

در ادامه به چندین مطالعه در مورد مقایسه عملکرد شبکه عصبی مصنوعی با سایر روش‌های پیش‌بینی ریسک اشاره می‌شود، به عنوان مثال، عزیزاده و منصوری در سال 1392 عملکرد مدل‌های کلاسیک و شبکه‌های عصبی را در ارزیابی ریسک و اعتبار مقایسه کردند. کارایی مشتریان بانک تجارت این مطالعه مدل های کلاسیک (رگرسیون خطی و لجستیک) را با شبکه های عصبی مقایسه کرد که نتیجه مطالعه نشان داد که شبکه های عصبی نسبت به مدل های کلاسیک کارایی بیشتری دارند.

سه مدل تحلیل پوششی داده‌ها، شبکه‌های عصبی و رگرسیون لجستیک برای پیشبینی رتبه اعتباری مشتریان حقوقی متقاضی وام بانکی در تحقیق ابراهیمی و دریابر در سال 1391 با یکدیگر مقایسه شدند، نتایج به دست آمده بیانگر قدرت بالای شبکههای عصبیدر پیش‌بینی ریسک است. کارایی شبکههای عصبی در مقایسه با مدل خطی آریما در تحقیقی با عنوان پیش بینی شاخص بورس اوراق بهادار با استفاده از شبکه‌های عصبی در سال 1384 توسط سینایی و همکاران مقایسه شد، نتیجه تحقیق نشان داد که شبکه‌های عصبی با خطای کمتری موفق به پیش‌بینی شاخص قیمت نسبت به مدل آریما شده است.

از جمله این تحقیقات می توان به تحقیق کفایی و راهزنی در سال 1395 با عنوان «بررسی تأثیر متغیرهای کلان اقتصادی بر ریسک نقدینگی» اشاره کرد که به صورت مدل رگرسیونی و با استفاده از روش داده های تابلویی فصلی انجام شده است. از اطلاعات 14 بانک کشور برای محاسبه ریسک نقدینگی و سنجش نقدینگی با استفاده از «نسبت دارایی‌های نقدی و مطالبات کوتاه‌مدت تقسیم بر بدهی‌ها و سپرده‌های کوتاه‌مدت» انصاری و همکاران، از «نسبت دارایی‌های نقدی» استفاده می‌شود.

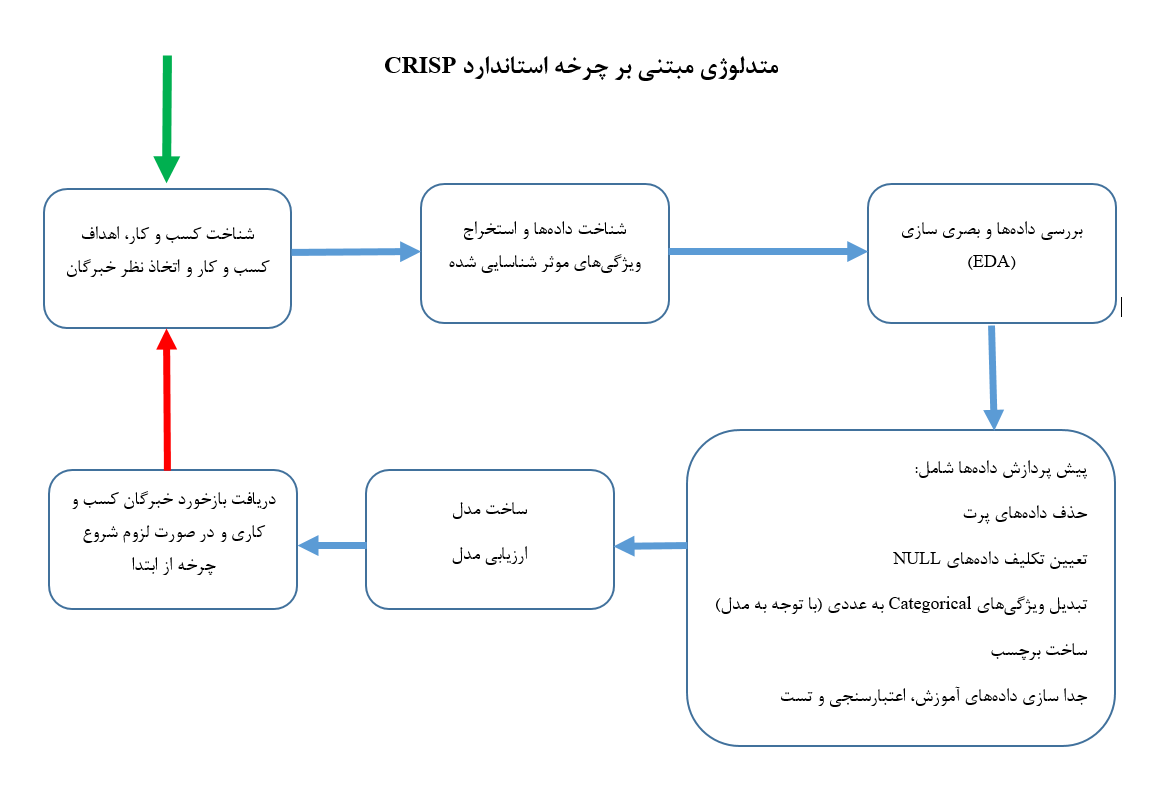
در نهایت مطالعه ای توسط اصغری اسکی (1392) با عنوان استفاده از شبکه های عصبی در پیش بینی سری های زمانی به بررسی عوامل مختلف ساختاری، روش های آموزشی مختلف و انتخاب داده ها و مکان یابی در فرآیند پیش بینی با استفاده از شبکه های عصبی پرداخته است. . در این تحقیق از داده های اقتصادی کشور استفاده شد و عواملی مانند روش پیش بینی، الگوریتم های آموزش شبکه های عصبی، تعداد بهینه ورودی و خروجی شبکه عصبی مورد بررسی قرار گرفت و پس از بررسی ساختارهای مختلف شبکه های عصبی، الگوریتم های موفق با ساختار مناسب به دست آمد. و ماشین های پیش بینی با عملکرد بهینه انتخاب و مورد ارزیابی قرار گرفتند. نتیجه این تحقیق نشان داد که انتخاب یک شبکه عصبی با ساختار و الگوریتم آموزشی مناسب ابزار بسیار قدرتمندی برای پیش‌بینی سری‌های زمانی است، به‌قدری قدرتمند که سری‌های زمانی را گام به گام پیش‌بینی می‌کند.

**4.2.نوآوری**

* استفاده از دیتاست جامع شامل اطلاعات مالی، تراکنشی، دموگرافیک مشتریان و شاخص‌های کلان اقتصادی مانند تورم، CPI، نرخ دلار، شاخص GDP در 5 سال اخیر
* محاسبه نقدینگی به ازای تمامی مشتریان: بر خلاف بقیه که یک عدد کلی رو پیش بینی میکنند اینجا با ریزدانگی مشتری سر و کار داریم که این امر موجب میشود تا مقدار ریسک نقدینگی را بتوان به روش ها و نسبتهای مختلفی محاسبه کرد
* امکان استفاده از دستاوردهای این پروژه در سایر پروژه ها و تصمیمات کلان بهره مانند:
  + پیش‌بینی منابع مشتریان به تفکیک نوع حساب و بالطبع پیش‌بینی رویگردانی مشتریان از منظر منابع بانکی
  + پیش‌بینی میزان بدهکاری (خوش حسابی) مشتری در آینده و ایجاد سیستم هشدار سریع بمنظور کاهش احتمال نکول تسهیلات
  + اعتبارسنجی مشتریان و محاسبه ریسک بازپرداخت با توجه به مبلغ تسهیلات
  + کمک به تدوین مطلوب‌تر برنامه‌ریزی‌های بلند مدت و کوتاه مدت با توجه به نقدینگی و درآمدهای بانک
* محاسبه ریسک به ازای نوع مشتری (حقیقی، حقوقی، صنف و یا تمام مشتریان)
* گرچه در این پروژه از روش‌ها و الگوریتم‌های کلاسیک و سنتی یادگیری ماشین نیز استفاده شد اما برای اولین بار ریسک نقدینگی با استفاده از شبکه های عصبی بازگشتی بعنوان یکی از الگوریتمهای حوزه یادگیری عمیق انجام شد.

1. **روش اجرایی**

منابع یا داده‌های ورودی شامل پارامترهای مؤثر داخلی بانک شامل اطلاعات مالی و غیر مالی مشتریلان و پارامترهای تأثیرگذار اقتصادی و کلان است. در فرآیند حل این مسئله به طور کلی از چرخه استاندارد CRISP استفاده شده است. خلاصه ای از این چرخه را در نمودار زیر مشاهده می‌کنید:



انتظار می‌رود مدل بتواند، میزان منابع و تسهیلات را به ازای تمامی مشتریان در 3 ماه آتی پیش‌بینی کند و سپس با محاسبه 4 نسبت:

1. **"نسبت تسهيلات به سپرده ها"**
2. **"نسبت تمركز سپرده ( 100 سپرده گذار عمده)"**
3. **"نسبت تسهيلات بلندمدت به سپرده هاي كوتاه مدت"**
4. **"نسبت سپرده‌هاي ديداري (حساب جاری) به كل سپرده‌ها"**

اصلی‌ترینی شاخص‌های ریسک نقدینگی را پیش‌بینی کند.

در بخش بعدی، نحوه ساخت دیتاست و نحوه برچسب گذاری داده‌ها تشریح خواهد شد.

1. **موجود بودن کد منبع آموزش و مجموعه داده آموزشی در شرکت**

داده‌های برای هر مشتری در 5 سال گذشته و به صورت ماهانه استخراج شده است. اگر یک مشتری در بازه کمتر از 6 ماه در بانک حضور داشته باشد، اطلاعات وی از دیتاست حذف می‌گردد و خروجی مدل برای وی بصورت پیش‌فرض اطلاعات آخرین ماه وی خواهد بود.

ویژگی‌هایی که با توجه به تجربیات مشابه بنظر می‌رسید که در تعیین خروجی دقیق مدل موثرند، با در نظر گرفتن اولویت و میزان تاثیر بشرح ذیل است.

در ادامه به ازای هر نوع از مشتری مشخصات ویژگی‌ها ذکر می‌شود.

همچنین کد آموزش الگوریتم های یادگیری ماشین و یادگیری عمیق استفاده شده در این پروزه نیز به پیوست موجود است که در ادامه نیز در هر بخش به کد مربوطه مورد نظر در مجموعه کدهای آموزشی نوشته شده اشاره می‌گردد. در مرحله بعدی دادگان را به تفکیک نوع مشتری و نوع حساب جدا کرده تا برای هر یک مدل جداگانه ای با توجه به مختصات و مشخصات مربوطه ایجاد کنیم.

همان طور که پیش تر نیز ذکر شد انجام پیش‌بینی در این پروژه با دو روش پیوسته و گسسته انجام شد. در روش گسسته با کلاسه بندی دادگان سعی در دسته بندی آنها جهت افزایش دقت و تخمین مناسب تری از آینده مشتریان در بازه های مناسبی خواهیم داشت.

نحوه برچسب‌گذاری داده‌ها:

هدف نهایی از اجرای مدل اینست که به ازای هر مشتری مقادیر زیر را بصورت میانگین در 3 ماه آینده پیش‌بینی کند:

- معدل حساب‌های کوتاه مدت

- معدل حساب‌های بلند مدت

- معدل حساب‌های جاری

- معدل حساب‌های قرض الحسنه

- مانده تسهیلات نزد مشتری

در صورت تخمین این 5 مقدار، همانطور که پیش‌تر نیز اشاره شد 4 مورد از اصلی‌ترین نسبت‌های لازم جهت محاسبه ریسک نقدینگی به شرح زیر قابل محاسبه است:

1. **"نسبت تسهيلات به سپرده ها"**
2. **"نسبت تمركز سپرده ( 100 سپرده گذار عمده)"**
3. **"نسبت تسهيلات بلندمدت به سپرده هاي كوتاه مدت"**
4. **"نسبت سپرده‌هاي ديداري (حساب جاری) به كل سپرده‌ها"**

لذا برای یک مشتری خاص، بخشی از دیتاست و 3 مورد از برچسب‌های آن بشرح ذیل خواهد بود:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| میانگین مانده تسهیلات در 3 ماه بعد (ماه‌‌های 140202 تا 140204) | میانگین معدل حساب کوتاه در 3 ماه بعد (ماه‌‌های 140202 تا 140204) | میانگین معدل حساب کوتاه در 3 ماه بعد (ماه‌‌های 140202 تا 140204) | میانگین معدل حساب جاری در 3 ماه بعد (ماه‌‌های 140202 تا 140204) | ویژگی 3 | ویژگی 2 | معدل حساب جاری | سال و ماه | شماره مشتری |
|  |  |  | y |  |  | x | 140201 | 1234 |
|  |  |  |  |  |  |  | 140202 | 1234 |

اما از آنجا که این نوع برچسب ممکن است مقادیر از صفر تا بی‌نهایت (اعداد بسیار بزرگ) داشته باشد، لذا مثلا بجای "میانگین معدل حساب جاری در 3 ماه بعد" میزان تغییرات این مقدار نسبت به "معدل حساب جاری مشتری در ماه ثبت رکورد" بصورت زیر در نظر گرفته شده است:

برچسب حساب جاری =

برای سایر برچسب‌ها نیز به همین منوال عمل شده است.

اما به منظور انجام پیش‌بینی و برحسب اجرای مدل‌های دسته‌بندی (Classification) و یا رگرسیون لازمست تا مقادیر برچسب‌ها به شکل گسسته یا به شکل پیوسته ذخیره شوند. از آنجا که هر 2 رویکرد فوق در این پروژه انجام شده است، لذا در حالت استفاده از روش‌های رگرسیونی از برچسب محاسبه شده در مرحله قبل استفاده می‌شود اما در حالت استفاده از روش‌های دسته‌بندی، برچسب محاسبه شده در مرحله قبل براساس مقدارش بصورت زیر گسسته سازی می‌گردد:

|  |  |
| --- | --- |
| مقدار برچسب رگرسیونی (محاسبه شده در مرحله قبل) | برچسب جدید (برای حالت دسته‌بندی) |
| کمتر از -50 درصد | 1 |
| بین -30 تا -50 درصد | 2 |
| بین -5 تا -30 درصد | 3 |
| بین -5 تا +5 درصد | 4 |
| بین +5 تا +30 درصد | 5 |
| بین +30 تا +50 درصد | 6 |
| بیش از +50 درصد | 7 |

در ادامه این گزارش، هر کجا از عبارت "برچسب رگرسیونی" استفاده شد منظور برچسب محاسبه شده بر اساس فرمول است و هر کجا از عبارت "برچسب دسته‌بندی" استفاده شد منظور برچسب گسسته سازی شده بر اساس جدول فوق است.

در بخش‌های بعدی ضمن تشریح مدل‌های اجرایی و پارامترهای درنظر گرفته شده، به بررسی نتایج مدل‌ها بر اساس هر 2 روش برچسب زنی پراخته می‌شود.

1. **متناسب بودن حجم دادگانی که در فرآیند مدلسازی، ارزیابی و اعتبارسنجی مورد استفاده قرار گرفته است.**

داده های ما حجم بسیار زیادی دارد که با توجه به مدل هایی که در ادامه آموزش خواهیم داد می تواند مقدار مناسبی از دادگان با Information کافی از فضای واقعی جامعه باشد و simulation دقیقی را از جامعه پشتیبانی کند. این موضوع با استفاده از الگوریتم MonteCarlo Simulation نیز در اسکریپت پایتون نوشته شده بررسی شده است. همچنین بازه اطمینان 95 درصد به ازای هر ویژگی استخراج شده که می تواند پشتیبانی خوب این sample داده ای ایجاد شده از جامعه واقعی مشتریان می باشد. (dataExplanation.py خط شماره 83)

همچنین این دادگان با نسبت 0.2 و 0.8 به دادگان test و train تقسیم شده اند. (این عملیات در فایل PreProcess.py در خط 54 انجام شده است.) انجام این عملیات جداسازی برای روشی پیش بینی با داده های کلاسه بندی شده به صورت Stratified و برای روشی که به دنبال پیش بینی داده ها به صورت پیوسته هستیم به صورت random انجام شده است.

1. **برخورداری از ابزار و فرایندهای مدیریت کیفیت داده.**

در این بخش به بررسی کیفیت داده و روش ها و ابزار های مورد استفاده برای پایش آن می‌پردازیم. برای انجام این عملیات در طول اسکریپت نوشته شده ما انواع مصورسازی یا visualization را در کلاس نوشته شده در فایل dataExplanatory جهت انجام فاز EDA (Explanatory Data Analysis) پروژه منطبق بر چرخه CRISP به صورت اتوماتیک انجام می‌دهیم. در این بخش انواع نمودارها و بررسی ها بر روی دادگان انجام می‌شود. همچنین همان طور که پیش تر ذکر شد عملیات بررسی های آماری بر روی دادگان مانند آزمون فرض، MonteCarlo Simulation و Central Limit Theorem را مورد بررسی قرار می‌گیرد.

بررسی های انجام شده در فاز EDA : (تمامی این بررسی‌ها در فایل DataExplanatory.py از ابتدا تا انتها قرار دارد)

Correlation: در این بخش میزان ارتباط میان ویژگی‌های مختلف بدست می‌آید.

Scatter Plot: در این بخش میان تمامی ویژگی های موجود که پیش از این ذکر شد(متغیر های مستقل) و ویژگی هدف (متغیر وابسته) یک نمودار Scatter رسم می‌شود.

UniqueValues : در این بخش با در نظر گرفتن مقادیر منحصر به فرد هر ویژگی در صورت وجود مقدار منحصر به فردی که از حیطه و گستره مقادیر مجاز یک ویژگی خارج است، آن را تشخصی داده و برای حل آن می‌کوشیم.

Distribution: در این بخش با ملاحظه توزیع هر یک از ویژگی های دادگان می‌توانیم از وجود مقادیر نامتعارف آگاه شده و آن ها را مدیریت کنیم همچنین با دانستن توزیع دادگان برای fit کردن مدل های مختلف نیز شرایط بسیار ساده تری را در پیش خواهیم داشت.

CategoryCount: در این بخش با شمارش هر مقدار منحصر به فرد در هر ویزگی میتوانیم مقادیر منحصر به فرد نویز و همچنین مقادیر منحصر به فرد بسیار کمیاب و نادر را تشخیص و از مقادیر دادگان خارج کنیم.

ViolonPlot:

BoxPlot:

پیش پردازش عمومی داده‌ها:

به مانند هر پروژه تحلیلی دیگر نیاز است تا بمنظور استفاده بهینه از روش‌های یادگیری ماشین، پردازش‌های اولیه بر روی دیتاست انجام شود. در ادامه و با توجه به موارد ذکر شده در بخش 2-6 (تفسیر داده‌ها و آماده‌سازی اولیه داده‌ها) نسبت به حذف ویژگی‌ها با همبستگی (Correlations) بالا، حذف داده‌های پرت (Outliers) و نرمال سازی داده‌ها به روش‌خای مختلف پرداخته می‌شود.

خاطر نشان می‌گردد در این بخش تنها به پردازش‌هایی پرداخته می‌شود که فارغ از نوع مدل و نحوه برچسب گذاری باشد. به‌همین دلیل از پردازش‌هایی چون PCA و نحوه تبدیل داده‌های Categorical به مقادیر عددی که وابسته نوع مدل اند، در این بخش اجتناب شده و در مرحله بعدی و بعنوان یکی از گام‌های مدل‌های اجرایی بیان خواهد شد. (PreProcess.py)

1. **ذکر نوع (یا نام) الگوریتم‌های هوشمند مورد استفاده.**

در این پروژه از روش های مختلف و الگوریتم های کلاسیک و مدرن یادگیری ماشینی و یادگیری عمیقی استفاده شده است که دقت مدل را در شرایط مختلف محاسبه کرده و بیشترین میزان قابلیت اتکا را برای مدل های ما ایجاد کند. برخی از مدل های مورد استفاده در این پروژه به شرح زیر است: (این مدل ها در اسکریپت Models.py پیاده سازی کرده ایم)

1. مدل جنگل تصادفی (Random Forest) RF:
2. مدل XGBoost
3. مدل MLP
4. مدل GaussianProcess
5. Logistic Regression
6. ExtraTree:
7. مدلSVM (با استفاده از کرنل های مختلف آن)
8. مدل LinearRegression:
9. Polynomial Regression:
10. استفاده از RNN ها علی الخصوص مدل حافظه دار ترتیبی LSTM
11. استفاده از روش های Bagging و Ensamble Learning برای افزایش دقت مدل ها(در این روش ها با استفاده از چندین مدل یکسان یا متفاوت (ترجیحا یکسان) و تقسیم دادگان به بخش های مختلف و آموزش دادن تعدادی مدل برای هر یک از این بخش ها و ویژگی های جدا شده با استفاده از روش های مختلف Sampling مثل Bootstrapping افزایش دقت و قابلیت اتکا و Robustness را در هر یک از مدل ها داشتیم )

هر یک از مدل های ذکر شده در فوق با دو دیدگاه روش کلاسه بندی شده و روش پیش بینی پیوسته مقادیر اجرا و ارزیابی شده اند که در ادامه به صورت جزئی تر به نتایج و روند اجرای هر یک می‌پردازیم.

1. **میزان تناسب الگوریتم انتخاب شده با کاربرد هدف گذاری شده**

چالش‌ها و مشکلات موجود بر سر حل مساله:

با توجه به انواع مشتریان حقیقی، حقوقی، تجاری و ... و نیز تنوع حساب‌های مشتریان (جاری، قرض الحسنه، کوتاه مدت و بلند مدت)، نیاز است تا مدل‌های مختلفی با توجه به تنوع یاد شده پیاده‌سازی گردد.

از طرفی پردازش حجم بالای داده‌ها و تیز تعداد زیاد ویژگی‌های استخراج شده (حدود 80 ویژگی) یکی دیگر از چالش‌هایی است که ممکن است هر پروژه تحلیل داده‌ای را با مشکل مواجه کند.

مساله مهمتر اما، تاثیر عوامل بیرونی بر رفتار مالی مشتریان، تصمیمات آنی مشتریان، پیشنهادات سایر رقبا به مشتریان بمنظور سرمایه گذاری است. بمنظور غلبه بر این چالش سعی شده است که رابطه و میزان تاثیر پذیری مشتری از عاومل بیرونی را با لحاظ کردن پارامترهایی چون تورم، CPI، نرخ طلا و شاخص‌هایی چون GDP مدل کرد اما در هر صورت ممکن است مدل نتواند برخی تصمیمات آنی مشتری ناشی از حوادثی غیر مترقبه را پیش‌بینی و مدل کند که البته با این فرض که این تصمیمات بصورت مساوی بر ورود و خروج نقدینگی در بانک موثرند، می‌توان از آن چشمپوشی کرد.

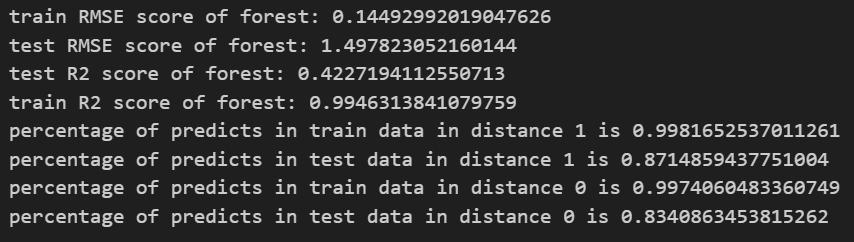
از طرفی ممکن است بانک بوسیله بازاریابی‌ها و مذاکرات و برگزای کمپین‌های مختلف موجب جذب مشتریان و منابع جدید بپردازد که این امر هم قابل پیش‌بینی برای مدل نیست که بانک قرار است در آینده چنین رفتاری را در دستور کار خود داشته باشد.

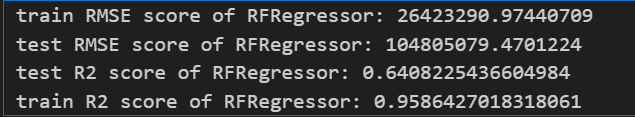
1. **مدل‌های اجرا شده و مدل پیشنهادی**

حال باید مدل توضیح داده شود. مدل‌های مختلفی به تجربه اجرا شد که هر کدام تشریح می‌شود.  
1-8 مدل جنگل تصادفی (Random Forest) RF:

الگوریتم یادگیری گروهی (Ensemble Learning Algorithm) است که چندین درخت تصمیم را برای پیش‌بینی ترکیب می‌کند. هر درخت تصمیم در جنگل تصادفی بر روی یک زیر مجموعه تصادفی از داده‌های آموزشی که با استفاده از Bootstrapping جدا شده است و یک زیر مجموعه تصادفی از فیچرها آموزش داده می‌شود. سپس خروجی جنگل تصادفی با تجمیع پیش‌بینی‌های همه درخت‌های تصمیم تعیین می‌شود. این رویکرد به جنگل تصادفی اجازه می‌دهد تا بسیار دقیق و مقاوم در برابر مشکل بیش‌برازش عمل کند. همچنین این روش در مقابل موارد نویز یا ویژگی‌های گمراه کننده مقاومت بسیار خوبی را دارد.

نتایج





2-8- مدل XGBoost:

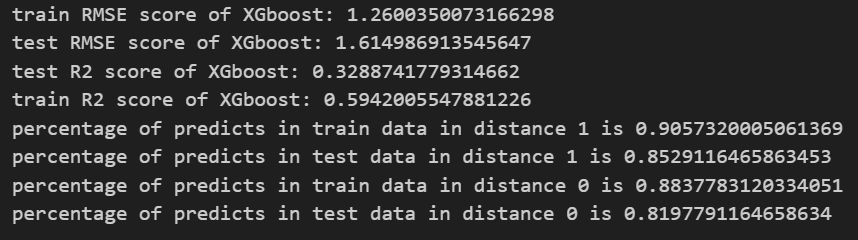
الگوریتم XGBoost یک الگوریتم یادگیری باناظر است که به‌خاطر قابلیت‌هایش در تنظیم خودکار، سرعت بالا در آموزش و قدرت پیش‌بینی‌اش، به‌عنوان یکی از قدرتمندترین و مؤثرترین ابزارها در حوزه یادگیری ماشین شناخته می‌شود و در طیف گسترده‌ای از کاربردهای عملی از جمله تجزیه و تحلیل داده‌های مالی استفاده می‌شود.

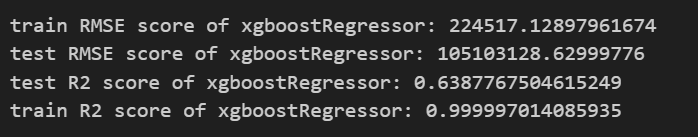
مزایای XGBoost با توجه به مساله پیش رو:

* دقت بالا (با بهینه‌سازی درخت‌های تصمیم و استفاده از تکنیک‌های پیشرفته، توانایی پیش‌بینی فوق‌العاده‌ای را ارائه می‌کند)
* کارایی محاسباتی (با استفاده از پردازش موازی و تکنیک‌های بهینه‌سازی، قابلیت اجرای کارآمد را حتی روی داده‌های بزرگ دارد)
* انعطاف‌پذیری (طیف گسترده‌ای از پارامترهای تنظیم‌شدنی دارد که امکان شخصی‌سازی آن را برای انواع مختلف داده‌ها و مسائل فراهم می‌کند)
* مدیریت داده‌های ناقص (با مجهزبودن به تکنیک‌های پیشرفته مدیریت داده‌های ناقص، به‌طور هوشمندانه، این داده‌ها را مدیریت می‌کند و از تأثیر منفی آن‌ها بر دقت پیش‌بینی جلوگیری می‌کند)
* رگولاریزاسیون (با استفاده از انواع مختلف رگولاریزاسیون تعادل میان انعطاف‌پذیری و دقت مدل را حفظ می‌کند)

نتایج:

نتایج مربوط به داده‌های کلاسه‌بندی شده:





3-7- مدل MLP:

اینم باید توضیح بدیم که چیه.

1-3-7- مزایای MLP با توجه به مساله پیش رو:

مدل خوبی است

2-3-7- معایب و مشکلات MLP با توجه به مساله پیش رو:

حافظه ندارد.

3-3-7- آماده‌سازی داده‌ها و برچسب گذاری داده‌ها:

مثلا PCA زدیم. نرمال کردیم به روش‌های زیر و ...

همچنین یک بار اومدیم برچسب‌ها رو به 5 کلاس تقسیم کردیم و یک بار هم رگرسیون زدیم.

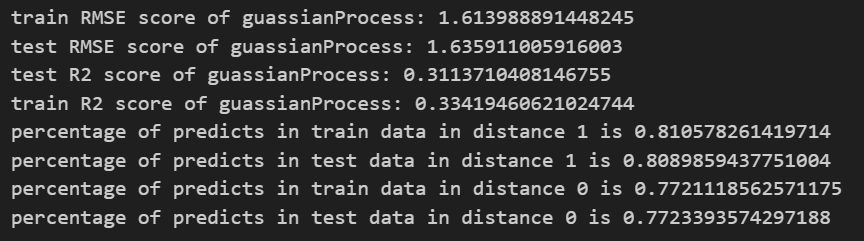
4-3-7- نتایج:

کانفیوژن ماتریکس، ROC و AUC و سایر معیارهای مربوطه با توجه به هر دو نوع برچسب زنی.

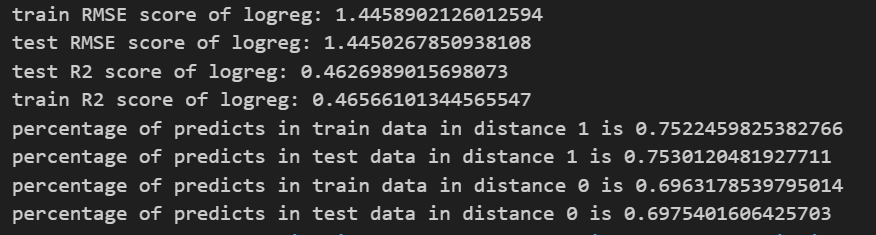
و باقی مدل‌ها مانند KNN، SVM، LR و MR و ...

4-8: مدل GaussianProcess:

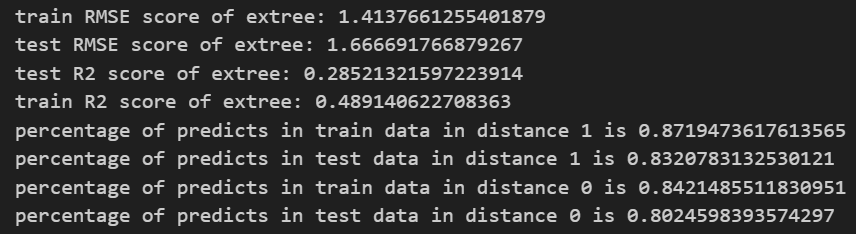
نتایج:



8-5: Logistic Regression:

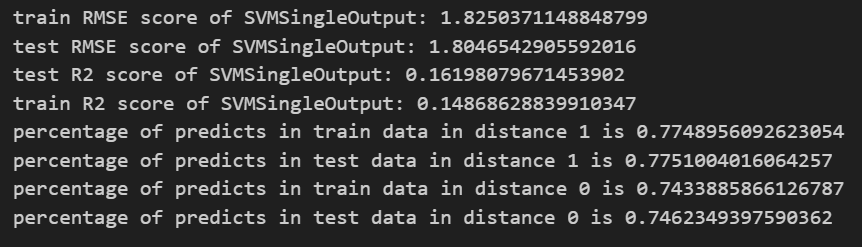


6-8: ExtraTree:

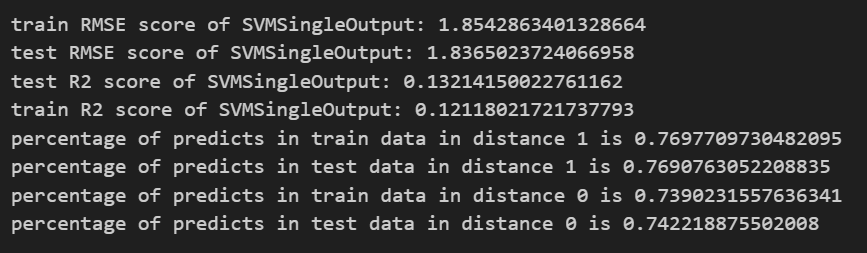


8-7: مدل SVM

1. با استفاده از کرنل rbf:

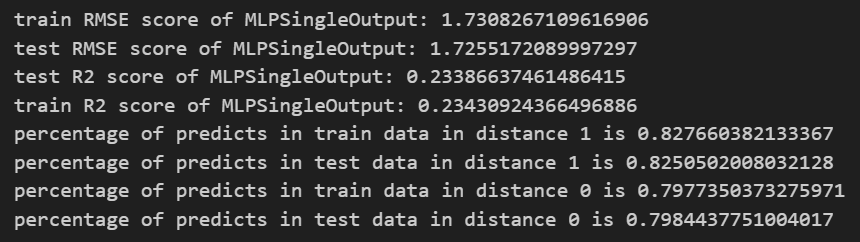


1. با استفاده از کرنل linear:

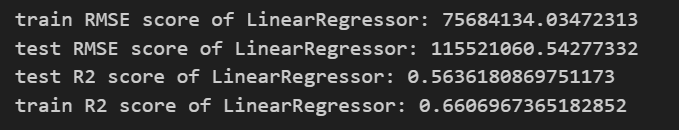


:

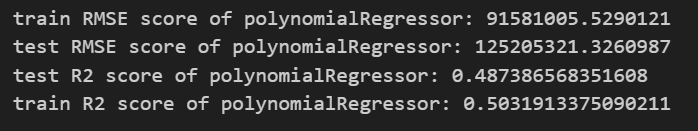
8-8: مدل MLP:



9-8: مدل LinearRegression:



10-8: Polynomial Regression:



مدل پیشنهادی:

مدل LSTM رو رفتیم.

اینجا هم باید مثل مدل‌های قبلی مزایا و معایب و ... تا نتایج نهایی رو بگیم.

1. **نتیجه گیری**

ما خیلی خوبیم.