|  |
| --- |
| به‌نام خدا  **فاز شناخت و فهم مسئله‌ی پروژه­ی درس مبانی داده‌کاوی**  **1401-02** |
| **اعضای گروه:** |
| سارا سلطانی گردفرامرزی |
| **عنوان پروژه:** |
| دسته ‌بندی برای پیش‌بینی پرداخت وام |
| **الف – مشکل چیست؟ (توضیح: چرا این مسئله یا داده تعریف شده است در واقع اگر این پروژه انجام نمی شد چه مشکلی، حل نشده باقی می ماند.)** |
| اگر مشتری وام های خود را به موقع پرداخت نکند بانک ها با ضرر زیادی مواجه میشوند که این مسئله در بلند مدت میتواند بر اقتصاد کشور تاثیرگذار باشد. در این مسئله به ویژگی‌های مختلفی مانند مبلغ وام، موجودی و غیره نگاه می‌کنیم تا پیش‌بینی کنیم که آیا یک فرد وام خود را پرداخت خواهد کرد یا خیر تا بتوان تا حدودی جلوی خسارت وارده به بانک ها را گرفت. |
| **ب-سوال داده کاوی و معیار ارزیابی آن چیست؟ (توضیح: قرار است چه مقدار(ستون یا اطلاعات) را پیش بینی (تخمین یا برآورد) کنید؟ و قصد دارید از چه اقلام اطلاعاتی برای این پیش بینی استفاده کنید؟معیار ارزیابی چیست و ایده آن چقدر است؟)** |
| مسئله‌ی ما باینری هست و باید در نهایت مشخص کنیم آیا به فرد با ویژگی‌های مشخص وام بدهیم یا نه.  اگر مسئله‌ی خود را LogisticRegression در نظر بگیریم به عنوان تابع هزینه می‌توانیم از Mean Squared Error و یا  Log lossاستفاده کنیم.  اگر بخواهیم مدل را براساس Decision Tree آموزش دهیم می‌توانیم از معیارهای Gini یا Entropy استفاده کنیم.  اگر از svm برای آموزش استفاده کنیم از معیارهای kernel (similarity) function استفاده می‌کنیم. |
| **ج- مشخصات دیتاست و ویژگی­های موجود در آن** |
| دیتاست ترین شامل 35 فیچر یا ستون بوده (که یکی از این ستون ها متعلق به برچسب یا کلاس مد نظر است و مابقی فیچرها در مرحله پیش پردازش جهت انتخاب برای ترین مدل استفاده میشوند). همچنین به تعداد 67463 سطر یا رکورد داریم که نظیر هر کدام از فیچر ها مقدار دارند. این فیچر ها از تایپ هایی همچون int64 ، float64 ، object بوده که هیچکدام از رکوردهای آن میس شده یا نال نیستند.  دیتاست تست نیز همچون دیتاست ترین است با این تفاوت که نظیر برچسب یا کلاس مدنظر مقداری نداشته و نال میباشد تا با استفاده از مدل های آموزش دیده این مقادیر پیش بینی شوند.  این فیچرها عبارتند از: مبلغ وام ، مبلغ وامی که تامین شده ، مبلغ وام تایید شده توسط سرمایه گذاران ، مدت وام ، شماره دسته نمایندگان ، نرخ بهره ، نمره داده شده توسط بانک ، مدت زمان استخدام مشتری ، وضعیت مالکیت خانه ، وضعیت تایید ، برنامه پرداخت وام ، عنوان وام ، نسبت کل بازپرداخت ماهانه بدهی نماینده تقسیم بر درآمد ماهانه ، تعداد بیش از 30 روز بزهکاری در 2 سال گذشته ، تعداد کل درخواست ها در 6 ماه گذشته ، تعداد حساب های باز مشتری ، تعداد سوابق بد عمومی فرد ، کل موجودی گردان اعتباری ، تعداد کل حساب های موجود فرد ، وضعیت وام (در انتظار یا پرداخت شده) ، کل سود دریافتی تا به امروز ، کل هزینه تاخیر دریافت شده تا به امروز ، تعداد حساب های معوق ، تک نفره یا چند نفره بودن مشتری ، کل موجودی جاری از همه حساب ها ، کل سقف اعتبار گردان |
| **د- فعالیت پیش رو و بیان مختصر ایده تیم** |
| ابتدا از روش های بصری سازی داده ها استفاده کرده تا روابط بین فیچرهای مختلف بررسی شود (مثلا از نمودار heatmap استفاده کرده تا همبستگی فیچرها به یکدیگر ارزیابی شود.) که این روش به ما کمک میکند تا فیچرهای اضافی برای ترین کردن مدل را حذف کنیم(مثلا اگر دو فیچر خیلی به یکدیگر همبسته بودند میتوان در صورت نیاز یک فیچر را حذف کرده تا عمل ترین ساده تر شود.) همچنین در این مرحله باید داده های categorical , nominal , ordinal… تشخیص داده شده تا رفتار متمایزی با هر کدام شود.  در مرحله بعد باید عملیات پیش پردازش روی داده ها انجام شود. میتوان داده ها را در صورت نیاز اسکیل کرده یا استاندارد کنیم. در صورت نیاز میتوان داده های پیوسته را با روش های sampling , binning …. گسسته کرد. در این مرحله باید داده های پرت یا میس شده را شناسایی کرده و برخورد مناسبی با آنها داشته باشیم.(با بررسی های انجام شده دیتاست مورد نظر داده ی میس شده ای نداشته و همچنین برای داده های پرت با روش های Capping & Flooring یا بهرگیری از میانه یا میانگین داده ها میتوان مقدار جایگزین در نظر گرفت.)  همچنین در این مرحله میتوان ستون های بیهوده را حذف کرد (مثلا میتوان تعداد مقادیر یکتا در هر فیچر را بررسی کرده تا اگر فیچری نظیر هر رکورد یک مقدار یکتا دارد در ترین کردن مدل استفاده نشود).  در مرحله پیش پردازش مهم است که اگر ستون های انتخابی برای ترین کردن مدل عددی نبوده (nominal , ordinal… ) باید با استفاده از روش های encoding تبدیل به داده ی عددی شوند تا بتوان برای ترین مدل از آنها استفاده کرد.  مرحله اصلی آموزش مدل های مختلف است. با استفاده از اطلاعات موجود و فیچرهایی که در مرحله پیش پردازش انتخاب شدند با الگوریتم های مختلفی همچون( درخت تصمیم ، svm ، Random Forest...) میتوان مدل های مختلف را آموزش داد.  در نهایت داده های تست را به مدل های آموزش دیده میدهیم تا پیش بینی مدنظر خود را انجام دهند و عملکرد هر کدام را به صورت جداگانه بررسی میکنیم تا در نهایت با استفاده از میار های ارزیابی و مقایسه دقت هر مدل بهترین مدل انتخاب شود. |