

**پروژه درس داده کاوی**

**پیش بینی احتمال رویگردانی مشتری در شرکت ملک رادار با استفاده ماشین بردار پشتیبان**

**Customer Churn Prediction using Support Vector Machine**

استاد راهنما: دکتر پرند

دانشجو: حمیدرضا فیروزه(400422147(

تابستان 1401

# مقدمه

شرکت ملک‌رادار، یک شرکت ارائه دهنده خدمات به صورت نرم‌افزار(Software as a Service- SaaS) در بستر ابری است. مشتریان هدف این شرکت مشاورین املاک فعال در سطح استان‌های تهران، البرز، خراسان، مازندران، گیلان، گرگان و فارس هستند.

سرویس ارائه شده توسط این شرکت بر اساس اشتراک‌های مدت‌دار به مشتریان ارائه شده و مشتری بعد از اتمام مدت اشتراک خود می‌تواند با تمدید آن، استفاده از سرویس‌های این شرکت را ادامه دهد.

یکی از ویژگی‌های اصلی این شرکت، ماهیت دورکاری آن است و در مدت 6 سال فعالیت آن، تمامی افراد به صورت ریموت و با استفاده از ابزارهایی مثل مایکروسافت تیمز، آژور دواپس و ...، به توسعه و عملیاتی شدن فعالیت‌های آن ادامه داده اند.

در حال حاضر در این شرکت چندین پروژه هوش مصنوعی در بخش‌های مختلف تعریف شده و تیم هوش مصنوعی شرکت به صورت موازی بر روی این پروژه‌ها در حال فعالیت است.

# تعریف مساله

همانطور که قبلا توضیح داده شد، مدل درآمدی شرکت براساس فروش اشتراک‌های مدت‌دار برای ارائه خدمات به مشاورین املاک تعریف شده است. یکی از مسائل مهمی‌ که این کسب و کار با آن دست و پنجه نرم می کرد، نرخ پایین بازگشت مشتری برای تمدید اشتراک بود. به همین دلیل شرکت تصمیم گرفت دلایل عدم تمدید رو پیدا کرده و در صدد رفع آنها قدم بردارد. دلایل اصلی عدم نرخ بازگشت مشتری شامل مواردی مانند تجربه بد خرید، تجربه بد استفاده، عدم رضایت از پشتیبانی، تغییر شغل توسط مشاور و موارد دیگری از این قبیل بود.

حال با اینکه این عوامل شناسایی شد، تشخیص اینکه کدام کاربر این مشکلات رو تجربه می‌کند کار سختی بود. یک راه تماس با تمام کاربران و درک مشکلات آنها بود که هزینه این تعداد تماس عملا غیر ممکن بود. به همین دلیل این پشینهاد مطرح شد که اگر ما قبل از اتمام اشتراک یک کاربر بتوانیم متوجه شویم قرار است از ما رویگردانی کرده و دیگر تمدید نکند، آنگاه در فقط با این مشتری‌های تماس گرفته و مشکل آنها را شناسایی و حل می‌کنیم.

برای یافتن مشتری‌ای که احتمال بالایی برای رویگردانی دارد، یکی از راه حل‌هایی که امروز استفاده می‌شود، روش‌های کلاسیفیکیشن است. از این روی، طی جلساتی که با افراد کلیدی شرکت داشتم، پروژه بهینه سازی نرخ بازگشت مشتری با عنوان CRO(Customer Retention Optimization) تعریف شد.

# هدف پروژه CRO

هدف این پروژه بهینه سازی نرخ بازگشت مشتری با استفاده از داده‌های در دسترس برای هر کاربر بود. خروجی این پروژه لیستی از مشتریانی بود که احتمالا برای خرید دوباره از سیستم ما اقدام نمی کنند. ما به دنبال این بودیم که 20 روز قبل از اتمام اشتراک این افراد این موضوع را بفهمیم و بر همین اساس با این افراد تماس گرفته و دلایل این رویگردانی احتمالی را درک کنیم.

# محدودیت‌های پروژه

بعد از شروع پروژه یکی از محدودیت های اصلی که با ان ربرو شدیم عدم در دسترس بودن تمامی اطلاعات رفتاری کاربر در طول دوره اشتراک آن بود. به طور کلی پروژه‌های CRO نیاز به اطلاعاتی پایه‌ای از رفتار کاربر درو طول دوره استفاده از سرویس دارند، ولی با توجه به محدودیت‌های بودجه در تهیه سرورهای مورد نیاز برای جمع‌آوری این دیتا، تنها بخشی از این داده‌ها در دسترس بود و ما بیشتر داده‌های رفتار خرید کاربر را در دسترس داشتیم تا داده‌های رفتار مصرف. به همین دلیل یکی از پیش بینی‌های اولیه ما این بود که احتمالا امتیاز نهایی مدل ما آن چنان بالا نخواهد بود و برای رفع این مشکل باید راه حل مناسبی پیدا کنیم که در بخش نهایی این گزارش این موضوع توضیح داده خواهد شد.

# ویژگی‌های دیتاست

بردار ویژگی‌های دیتاست شامل موارد زیر است.

| ویژگی | تایپ | توضیحات | ردیف |
| --- | --- | --- | --- |
| NormalMobileNo | int64 | شماره موبایل کاربر | 1 |
| InvoiceDate | object | تاریخ ثبت فاکتور | 2 |
| PaymentDate | object | تاریخ پرداخت | 3 |
| ExpireDate | object | تاریخ انقضای سرویس | 4 |
| TresholdDate | object | تاریخی که باید متوجه شیم مشتری قرار است رویگردانی کند یا تمدید کند | 5 |
| RegisterDate | object | تاریخ ثبت نام در سیستم | 6 |
| ServiceAmount | int64 | تعداد روز‌های سرویس خریداری شده | 7 |
| ServiceType | object | نوع سرویس خریداری شده | 8 |
| DurationBetweenRegisterationAndFirstPurchase | float64 | فاصله بین ثبت نام تا اولین خرید | 9 |
| PurchaseLabel | object | برچسب مشتری | 10 |
| PurchaseCountAdver | int64 | تعداد دفعات خرید سرویس تا کنون | 11 |
| PurchaseCountCustomer | int64 | تعداد دفعات خرید سرویس جایگزین تا کنون | 12 |
| PaymentTryCount | int64 | تعداد دفعات تلاش برای پرداخت | 13 |
| InvoiceSumOfTotalFee | int64 | مجموع فاکتور ها تاکنون | 14 |
| InvoiceSumOfTotalPrice | int64 | مجموع خرید بعد از کسر تخفیف | 15 |
| InvoiceSumOfTotalDiscount | int64 | مجموع تخفیف دریافتی تا کنون | 16 |
| InvoiceSumOfAdverFee | int64 | مجموع فاکتور سرویس اصلی تا کنون | 17 |
| InvoiceSumOfAdverPrice | int64 | مجموع پرداخت سرویس اصلی تا کنون | 18 |
| InvoiceSumOfAdverDiscount | int64 | مجموع تخفیف سرویس اصلی تا کنون | 19 |
| InvoiceSumOfCustomerFee | int64 | مجموع فاکتور سرویس جایگزین تا کنون | 20 |
| InvoiceSumOfCustomerPrice | int64 | مجموع پرداخت سرویس جایگزین تا کنون | 21 |
| InvoiceSumOfCustomerDiscount | int64 | مجموع تخفیف سرویس جایگزین تا کنون | 22 |
| PaymentDurationsMax | float64 | ماکسیم فاصله زمانی بین اتمام سرویس و تمدید بعدی | 23 |
| PaymentDurationsMin | float64 | مینیم فاصله زمانی بین اتمام سرویس و تمدید بعدی | 24 |
| PaymentDurationsAvg | float64 | میانگین فاصله زمانی بین اتمام سرویس و تمدید بعدی | 25 |
| FolderCount | float64 | تعداد زونکن های ایجاد شده | 26 |
| FlowAgents | object | کارشناس فروش تماس گرفته با کاربر تا کنون | 27 |
| FlowAgentsCount | int64 | تعداد کارشناسان مختلف تماس گرفته | 28 |
| FlowAgentsSaleOtherCount | int64 | تعداد کارشناسان فروش تماس گرفته | 29 |
| FlowAgentsSaleAziziCount | int64 | تعداد تماس کارشناس عزیزی با کاربر | 30 |
| FlowAgentsSalePeymanCount | int64 | تعداد تماس کارشناس پیمان با کاربر | 31 |
| FlowAgentsSaleJabbarpourCount | int64 | تعداد تماس کارشناس جبارپور با کاربر | 32 |
| FlowAgentsSaleBarghiCount | int64 | تعداد تماس کارشناس برقی با کاربر | 33 |
| HasSatisfactionCall | object | ایا تماس رضایت بعد از خرید داشته | 34 |
| SupportCallCount | int64 | تعدا تماس های پشتیبانی با کاربر | 35 |
| FlowCount | int64 | تعدادکانال های تماس متنوعی که کاربر در آن قرار داشته | 36 |
| City | object | شهر محل فعالیت کاربر | 37 |

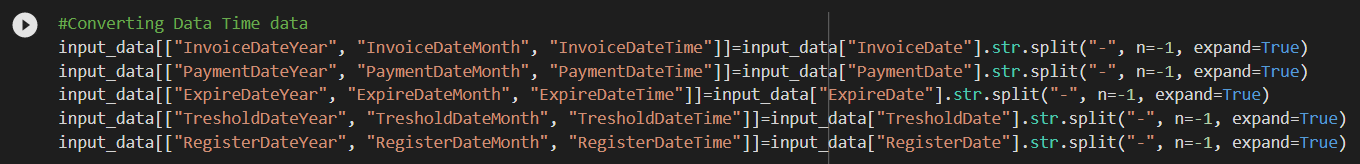
فیچر شماره 10 در رواقع همان لیبل ما از مشتریان است، اگر کاربر از سیستم رویگردانی کند برچسب Churn و اگر تمدید کند، برچسب Retention خواهد خورد.

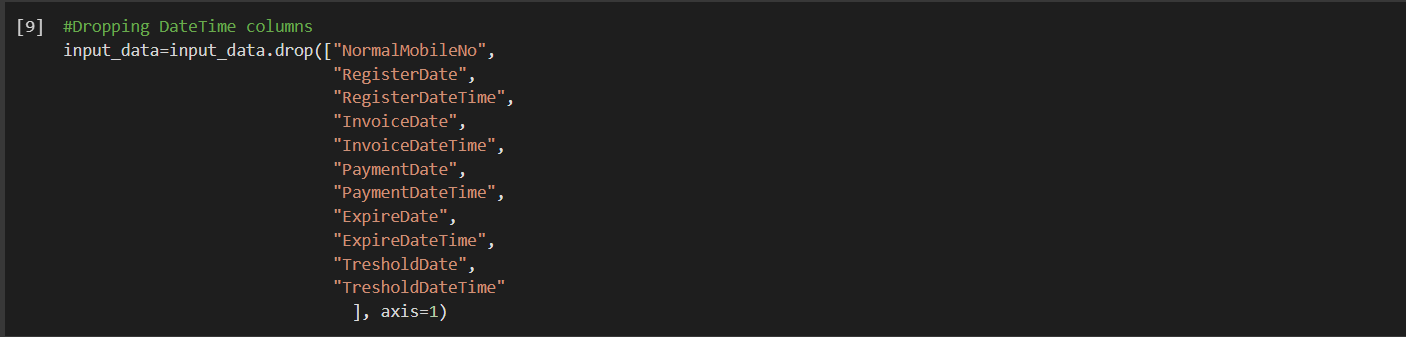
در فیچرهای 11 تا 22 در مورد سرویس اصلی و جایگزین صحبت شده است. سرویس اصلی به نام adver، خدمت اصلی شرکت بوده و هدف اصلی هم پیش بینی تمدید این سرویس است. سرویس جایگزین، در واقع سرویس ثانویه شرکت بوده و برای تعداد کمی از مشاورین کاربرد دارد. با این حال نحوه استفاده از این سرویس هم بر روی رفتار تمید تاثیر گذار خواهد بود.

# پیش پردازش داده ها:

1. تبدیل تاریخ

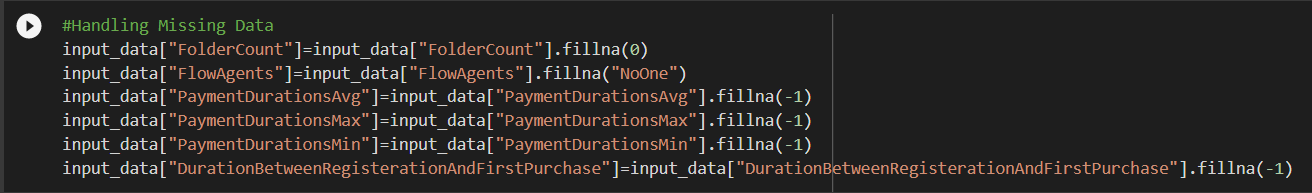
داده‌های تاریخ زمانی، فرمت مناسبی برای استفاده ندارند. به همین دلیل این داده‌ها به دو ستون داده سال و ماه تبدیل خواهند شد.

سپس داده اصلی حذف می‌شود.

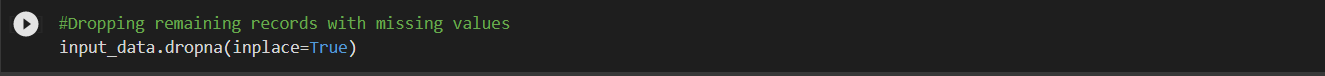


2-حذف داده‌های گمشده

برای داده‌های گمشده از روش ساده جایگزینی اسفتاده شده است. بر اساس دستورالعمل زیر این داده‌ها پر می‌شوند.



در صورتی که بعد از فرایند بالا همچنان داده گمشده‌ای وجود داشته باشد آن رکورد به طور کامل حذف خواهد شد.



# آماده سازی پایپلاین مدل یادگیری ماشین

ابتدا متدهای برای هریک از متغیرهای کتگوریکال، اردیتنال و عددی ترنسفورمر مورد نظر را تعریف می‌کنیم.



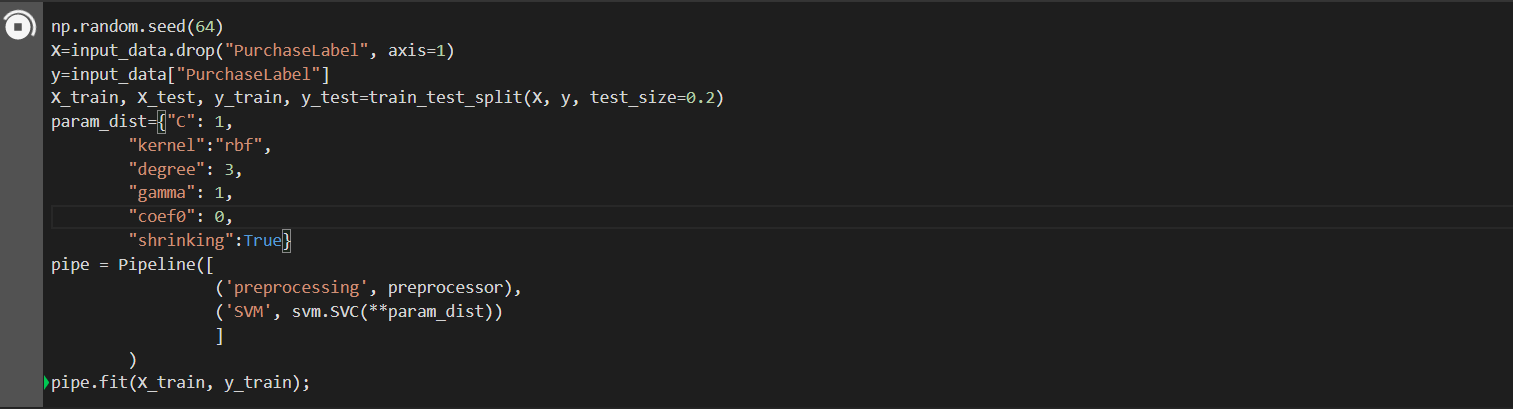
حال با استفاده از متد کی برای Support Vector Classification تعریف شده مدل مدنظر را آموزش خواهیم داد.

ابتدا هسته اعداد تصادفی را ثابت می‌کنیم تا نتایج قابل مقایسه باشند.

سپس داده‌های X و y رو بر اساس داده‌های ورودی تعریف کرده و داده‌ها را بر اساس نسبت 20 به 80 به دو دسته تست وترین تقسیم می‎‌کنیم.

حال لیست هایپرپارامترهای SVC را تعریف می‌کنیم. با توجه به اینکه این پروژه نقش آموزشی داشته ما لیست کوچکی از هایپرپارامترها رو تعریف کرده و تا ترکیب بهینه این لیست کوچک را در مراحل بعد بدست آوریم.

در پایان هم مدل را فیت کرده و به بررسی و تحیلل آن خواهیم پرداخت.

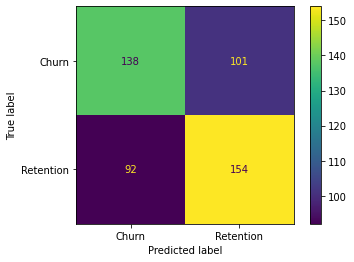


بعد از فیت شدن مدل اسکور آنرا محاسبه می کنیم. که عددی برابر 60% خواهد بود.

در فرایند اجرای پروژه مدل‌های متفاوتی تست شد و در نهایت بیشترین امتیاز با استفاده از مدل XGBoost با مقدار 64% محاسبه شد. دلیل اصلی اینکه امتیاز خیلی بالایی برای این مدل بدست نیامد، این است که داده‌های استفاده شده برای مدل اکثرا مربوط به رفتار خرید بوده اند و برای محاسبه این نرخ رویگردانی داده های مصرف کاربر از اهمیت بالاتری برخوردا بوده و بهتر است این داده‌ها استفاده شوند.

# حل مساله امتیاز پایین مدل SVC

ابتدا باید درک درستی از نحوه پیش بینی مدل ساخته شده داشته باشیم. به همین دلیل کانفیوژن ماتریس مدل ساخته شده برای داده‌های تست را رسم می‌کنیم



نکته مهم برای این مساله از دست ندادن افرادی است که احتمال رویگردانی دارند. در واقع هرچه احتمال این افراد کمتر باشند، آنگاه مساله شرایط بهتری دارد. در واقع ما به دنبال این هستیم که آن 101 نفری که در باکس آبی رنگ هستند و به اشتباه لیبل retenetion برای آنها پیش بینی شده را کم کنیم. طی جلساتی که با تیم‌های پشتیبانی شرکت برقرار شد، اولین راهکار برای افراد با لیبل Churn تماس تلفنی بوده، و با توجه به اینکه ظرفیت تماس برای تیم پشتیبانی بالا بود، این امکان وجود داشت که ما ترش هلد مد نظر برای پیش بینی لیبل را از 0.5 به اعداد دیگری تغییر دهیم. اگرچه این کار باعث می‌شد که افرادی که واقعا قصد تمدید داشته باشند، لیبل Churn دریافت کنند ولی برای مساله ما موضوع چندان با اهمیتی نبود، چراکه حتی تماس با این افراد منجر به تغییر نظر آنها نخواهد شد، بلکه این سبک از پشتیبانی حتی می توانست آنها ار در تصمیم خرید دوباره، مصمم‌تر نماید. تنها موردی که نگرانی ایجاد می‌کرد، هزینه تماس با افرادی بود که واقعا نیاز به تماس نداشتند ولی همانطور که گفته شد، ظرفیت تیم پشتیبانی برای این تماس‌ها کاملا آزاد بوده و این مساله از دغدغه‌های این پروژه خارج شده بود.