عنوان پروژه : مدل‌سازی پیش‌بینانه صنعتی برای ریزش مشتریان تلکام

۱ .تعریف مسئله

هدف از این پروژه، توسعه یک مدل پیش‌بینانه با استفاده از مجموعه داده "Telco Customer Churn"است تا تعیین کند آیا یک مشتری بر اساس ویژگی‌هایی مانند مدت زمان حضور(Tenure) ، هزینه‌های ماهانه، نوع قرارداد و روش پرداخت، سرویس خود را لغو می‌کند) ریزش یا (Churn یا خیر.

۲ .معیارهای موفقیت

برای اینکه پروژه موفقیت‌آمیز تلقی شود، مدل نهایی باید حداقل به معیارهای زیر دست یابد:

- **دقت :(Accuracy) حداقل ۸۵٪**
- **امتیاز :(F1-score) F1 حداقل ۰.۸۰**

این پروژه مطابق با مسیر شماره ۲ (یادگیری ماشین صنعتی) پیش می‌رود و الگوریتم‌های کلاسیک مانند Random Forestو XGBoost را برای شناسایی کارآمدترین راهکار جهت کاهش ترک خدمت مشتریان، مقایسه خواهد کرد.

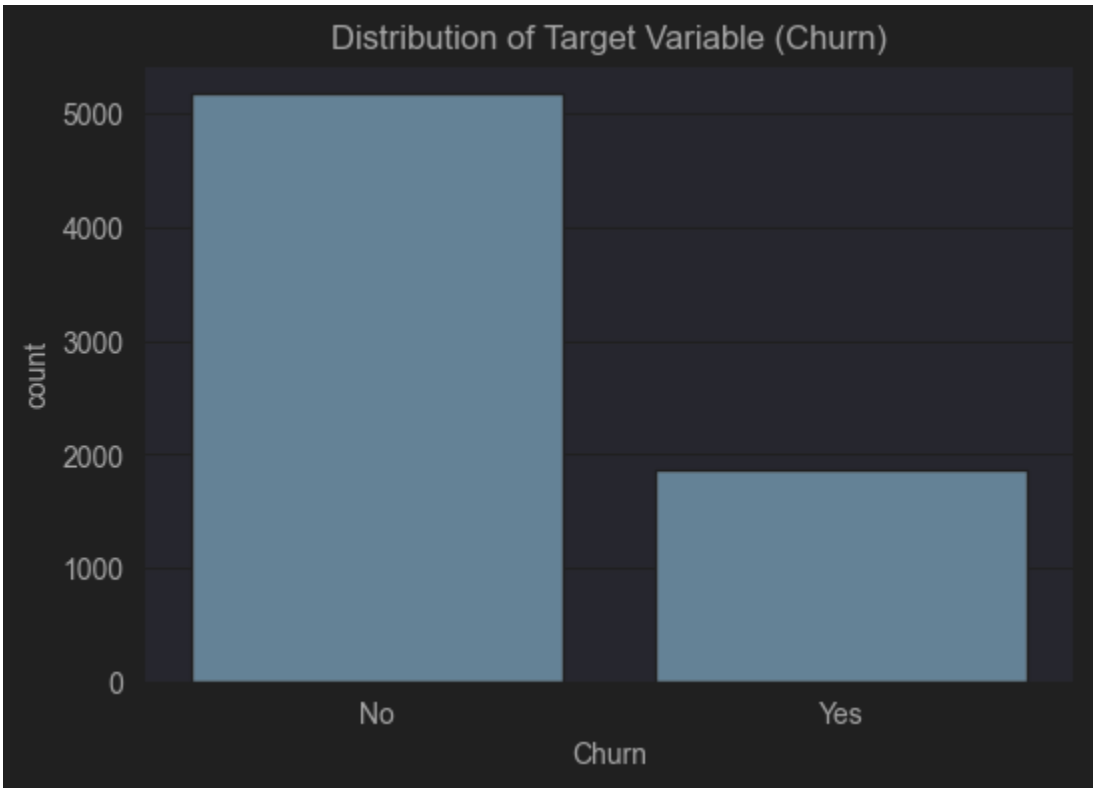
۳ .خلاصه داده‌ها :

جزئیات	مقدار
منبع مجموعه داده	Telco Customer Churn (مجموعه داده صنعتی)
تعداد نمونه‌ها	۷,۰۴۳
تعداد ویژگی‌ها	۲۰ (ورودی) + ۱ (هدف)
متغیر هدف	Churn (بله/خیر)

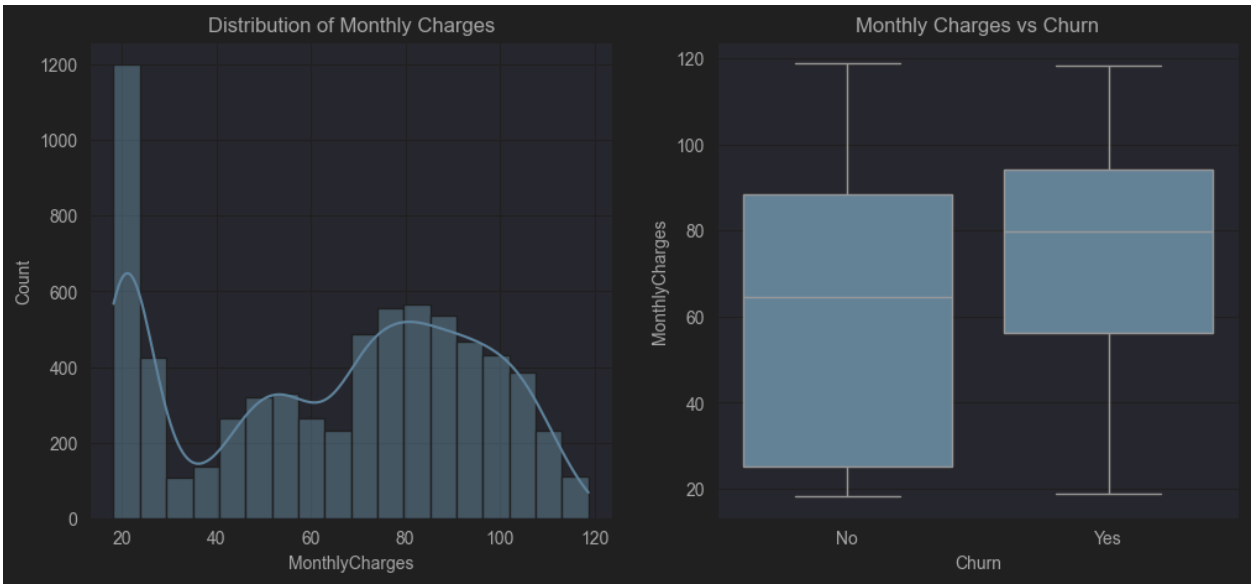
۴ .تحلیل اکتشافی داده‌ها (EDA) و بصری‌سازی

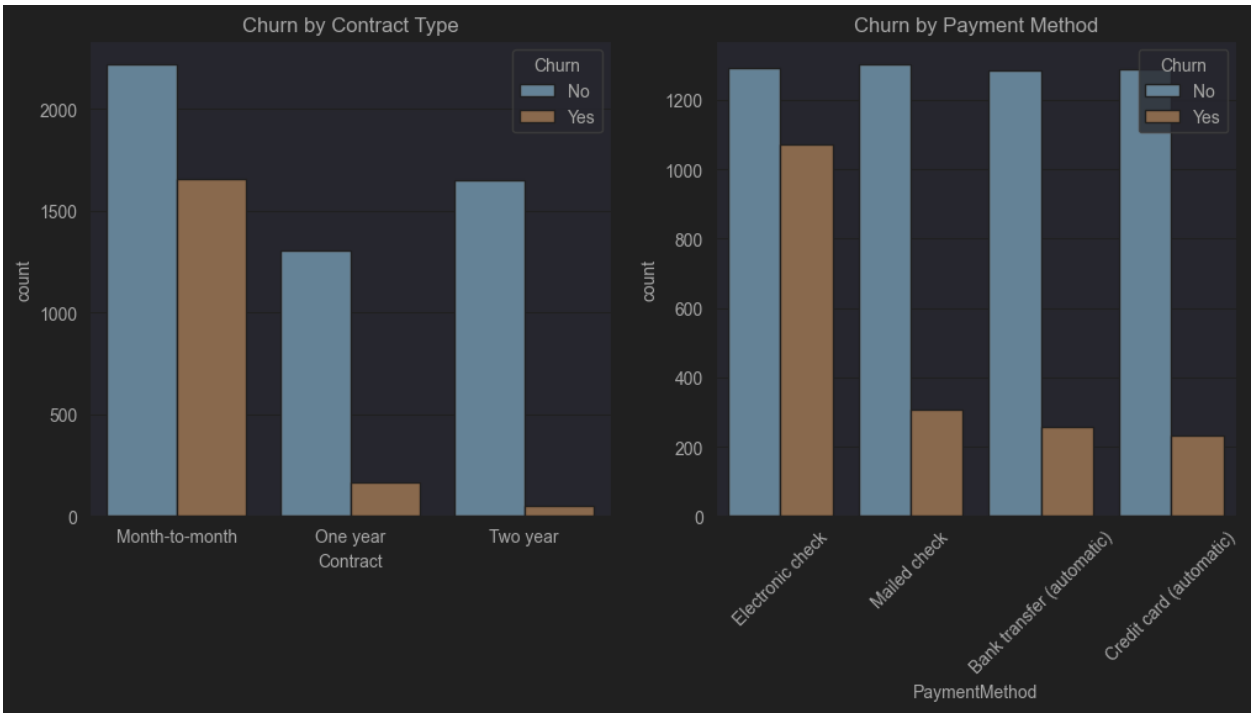
در این مرحله، نمودارهای زیر برای درک بهتر داده‌ها ترسیم شده‌اند:

- **نمودار ۱ :** توزیع متغیر هدف) نمودار میله‌ای Churn در مقابل.(No Churn
- **نمودار ۲ :** ماتریس همبستگی (نقشه حرارتی ویژگی‌های عددی).
- **نمودار ۳ :** توزیع هزینه‌های ماهانه) هیستوگرام به همراه خط.(KDE
- **نمودار ۴ :** مقایسه هزینه‌های ماهانه در مقابل ریزش (نمودار جعبه‌ای).
- **نمودار ۵ :** ریزش بر اساس نوع قرارداد (مقایسه قراردادهای ماهانه، یک‌ساله و غیره).
- **نمودار ۶ :** ریزش بر اساس روش پرداخت (بررسی روش‌هایی مانند چک الکترونیکی، چک پستی و غیره).




Correlation Matrix
SeniorCitizen
tenure
MonthlyCharges
1
0.017
0.22
0.017
1
0.25
0.22
0.25
1
SeniorCitizen
tenure
MonthlyCharges
1.0
0.8
0.6
0.4
0.2


Distribution of Monthly Charges
Count
MonthlyCharges
0
200
400
600
800
1000
1200
20
40
60
80
100
120
Monthly Charges vs Churn
MonthlyCharges
Churn
No
Yes
20
40
60
80
100
120



۵ .پیش‌پردازش داده‌ها

۵.۱ .پاکسازی و مهندسی ویژگی‌ها

برای آماده‌سازی داده‌های خام جهت استفاده در مدل، مراحل زیر انجام شد:

- **تبدیل عددی :** ستون TotalCharges از نوع متنی (Object) به نوع اعشاری (Float)تبدیل شد.
- **مدیریت مقادیر خالی :** ردیف‌های دارای رشته‌های خالی به عنوان مقادیر Null در نظر گرفته شده و با استفاده از "میانه (Median) "ستون مربوطه پر شدند.
- **حذف ویژگی‌های زائد :** ستون customerID به دلیل اینکه یک شناسه منحصربه‌فرد است و قدرت پیش‌بینی ندارد، حذف شد.
- **کدگذاری متغیرهای دسته‌ای :** از روش **One-Hot Encoding**برای تبدیل متغیرهای متنی (مانند نوع قرارداد و روش پرداخت) به ستون‌های باینری استفاده شد.

- **کدگذاری هدف :** متغیر هدف (Churn) با استفاده از Label Encoder به فرمت باینری (۰ برای خیر و ۱ برای بله) تبدیل شد.

۵.۲ .مقیاس‌بندی و تقسیم‌بندی داده‌ها

- **استانداردسازی : برای جلوگیری از تسلط ویژگی‌های با بازه بزرگتر مانند (TotalCharges) بر مدل، از StandardScaler برای تمامی ستون‌های عددی استفاده شد.**
- **تقسیم‌بندی داده‌ها : داده‌ها به دو بخش ۷۰٪ برای آموزش و ۳۰٪ برای تست تقسیم شدند.**
- **نمونه‌برداری طبقه‌بندی شده (Stratified Sampling) : این روش برای حفظ توازن توزیع "ریزش" در هر دو مجموعه آموزش و تست به کار گرفته شد.**

۶ .انتخاب مدل پایه (Baseline) و نتایج

مطابق با الزامات مسیر صنعتی، مدل رگرسیون لجستیک (Logistic Regression) به عنوان مدل پایه انتخاب شد.

علت انتخاب : تفسیرپذیری بالا و کارایی مناسب در وظایف طبقه‌بندی باینری با داده‌های جدولی استاندارد.

عملکرد مدل پایه :

معیار	امتیاز
دقت (Accuracy)	۸۰.۸۸٪
امتیاز F1 (F1-Score)	۰.۶۰۸۵

تحلیل نتایج :

- دقت (Accuracy) : مدل پایه به دقت ۸۰.۸۸٪ دست یافت که نقطه شروع خوبی است، اما در حال حاضر کمتر از هدف ۸۵٪ تعیین شده برای پروژه است.
- دقت (Precision) و فراخوانی (Recall) : مدل برای مشتریانی که ریزش ندارند دقت بالاتری (۰.۸۵) نسبت به مشتریان دارای ریزش (۰.۶۷) نشان می‌دهد؛ این موضوع بیانگر نیاز به دسته‌بندی بهتر کلاس اقلیت یعنی "Churn" در مرحله بعدی است.
- ماتریس آشفتگی (Confusion Matrix) : بصری‌سازی داده‌ها تأیید می‌کند که اگرچه مدل در شناسایی مشتریانی که می‌مانند خوب عمل می‌کند، اما تقریباً ۴۴٪ از ریزش‌کنندگان واقعی را از دست می‌دهد (با فراخوانی یا Recall معادل ۰.۵۶).
-

۷ .برنامه آزمایشات برای فاز دوم

برای رسیدن به هدف ۸۵٪ دقت، گام‌های زیر در فاز بعدی برداشته خواهد شد:

1. مقایسه الگوریتم‌ها : پیاده‌سازی و مقایسه Random Forestو .XGBoost
2. تنظیم ابرپارامترها : بهینه‌سازی پارامترهایی مانند تعداد درخت‌ها، عمق درخت و نرخ یادگیری.
3. مهندسی ویژگی پیشرفته : آزمایش تعامل بین ویژگی‌ها (مانند رابطه مدت حضور و هزینه‌های ماهانه).
4. مدیریت عدم توازن داده‌ها : بررسی تکنیک‌هایی مانند SMOTEیا تنظیم وزن کلاس‌ها برای بهبود امتیاز