گزارش پروژه‌ی یادگیری ماشین

دانشکده مهندسی کامپیوتر - ارائه شده توسط دکتر شریفی زارچی

پائیز 1404

نام و نام خانوادگی اعضا	شماره‌ی دانشجویی اعضا
ستایش حیدری	401104073
امیرحسین شهرابی	401104208
امیرعباس دنیادیده	401104113

فهرست مطالب

این پروژه به بررسی دیتاست ریزش مشتریان در شرکت مخابراتی پرداخته است با هدف درک عواملی که منجر به ترک مشتریان می‌شود و پیش‌بینی اینکه کدام مشتریان احتمال بیشتری برای ترک دارند. پروژه در مراحل مختلف شامل کاوش داده‌ها، تجسم اطلاعات و تحلیل‌های مختلفی انجام شده است تا الگوها و روابط میان داده‌ها شناسایی شوند. تمرکز اصلی بر شناسایی ویژگی‌های کلیدی است که بر ریزش مشتری تأثیر می‌گذارند و ارائه بینش‌هایی برای کاهش ریزش.

فاز اول: تحلیل اکتشافی داده ها (EDA)

دیتاست شامل اطلاعات دقیقی از مشتریان است که شامل ویژگی‌های جمعیت‌شناسی، خدمات مشترک، جزئیات حساب و وضعیت ریزش مشتری است. متغیر هدف در این دیتاست Churn است که نشان می‌دهد آیا مشتری شرکت را ترک کرده است (بله) یا ترک نکرده است (خیر).

دیتاست شامل ۷۰۴۳ ردیف و ۲۱ ستون است که شامل متغیرهای عددی و دسته‌ای است از جمله:

ویژگی‌های جمعیت‌شناسی: gender، SeniorCitizen، Partner، Dependents

خدمات: PhoneService، MultipleLines، InternetService، OnlineSecurity، OnlineBackup، DeviceProtection، TechSupport، StreamingTV، StreamingMovies

اطلاعات حساب: tenure، Contract، PaperlessBilling، PaymentMethod، MonthlyCharges، TotalCharges

هدف: Churn

نوع متغیرها و مدیریت داده های تکراری(Duplicates)

اولین گام در تحلیل، شناسایی و حذف داده‌های تکراری در ردیف‌ها و ستون‌ها بود. این بررسی نشان داد که دیتاست هیچ ردیف یا ستونی تکراری ندارد که این نشان‌دهنده صحت داده‌ها است. ستون‌هایی که بیش از حد منحصر به فرد بودند، مانند customerID از تحلیل حذف شدند زیرا این ستون‌ها اطلاعات مفیدی برای پیش‌بینی ریزش ندارند.

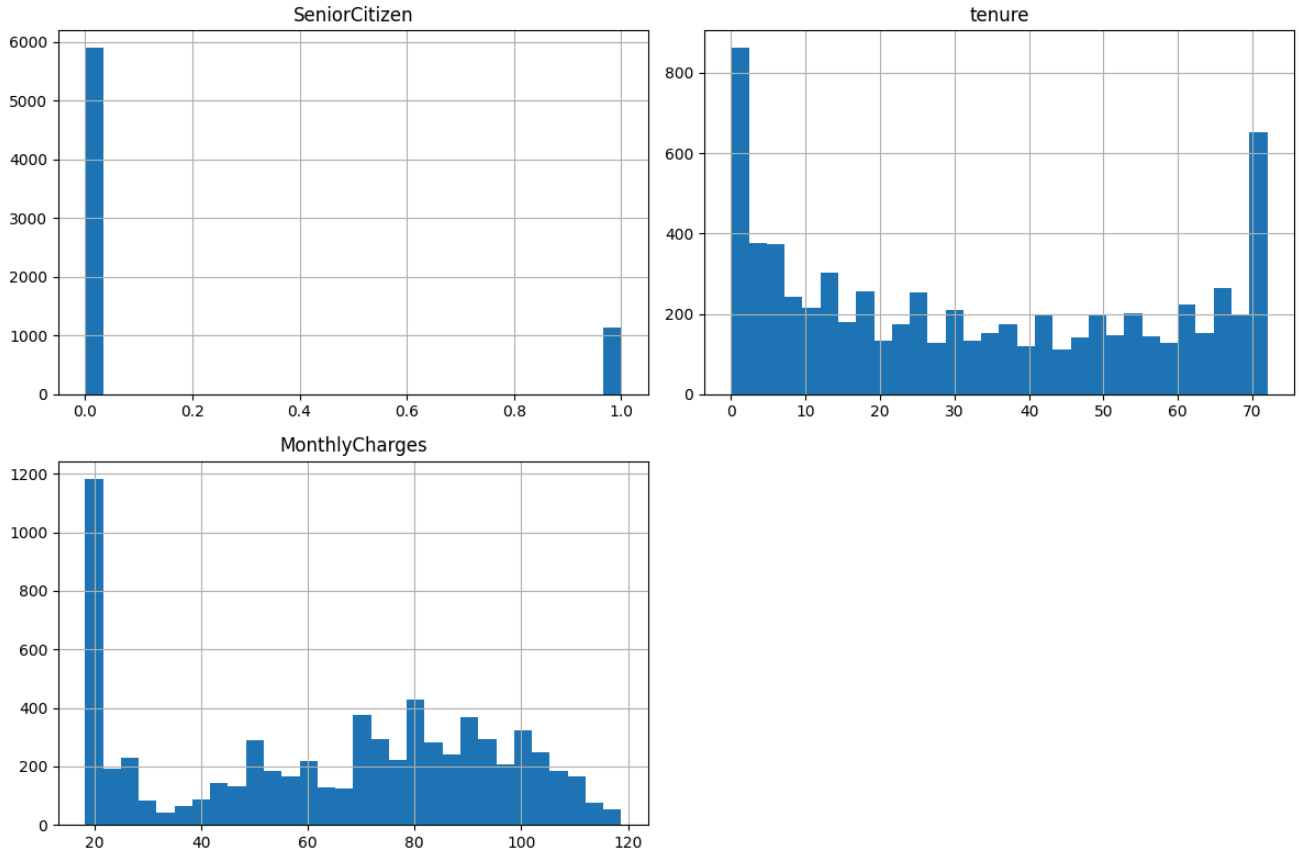
```
No duplicate column names found.
No duplicate columns by content found.
Shape after removing duplicates: (7043, 21)
```

رسم نمودار Histogram برای متغیرهای عددی

tenure :(مدت زمان حضور مشتری در شرکت) توزیعی نسبتاً متمایل به سمت راست دارد که نشان می‌دهد بیشتر مشتریان مدت کمتری در شرکت بوده‌اند، در حالی که تعدادی از مشتریان نیز مدت زمان طولانی‌تری با شرکت بوده‌اند که نشان‌دهنده وفاداری به شرکت است.

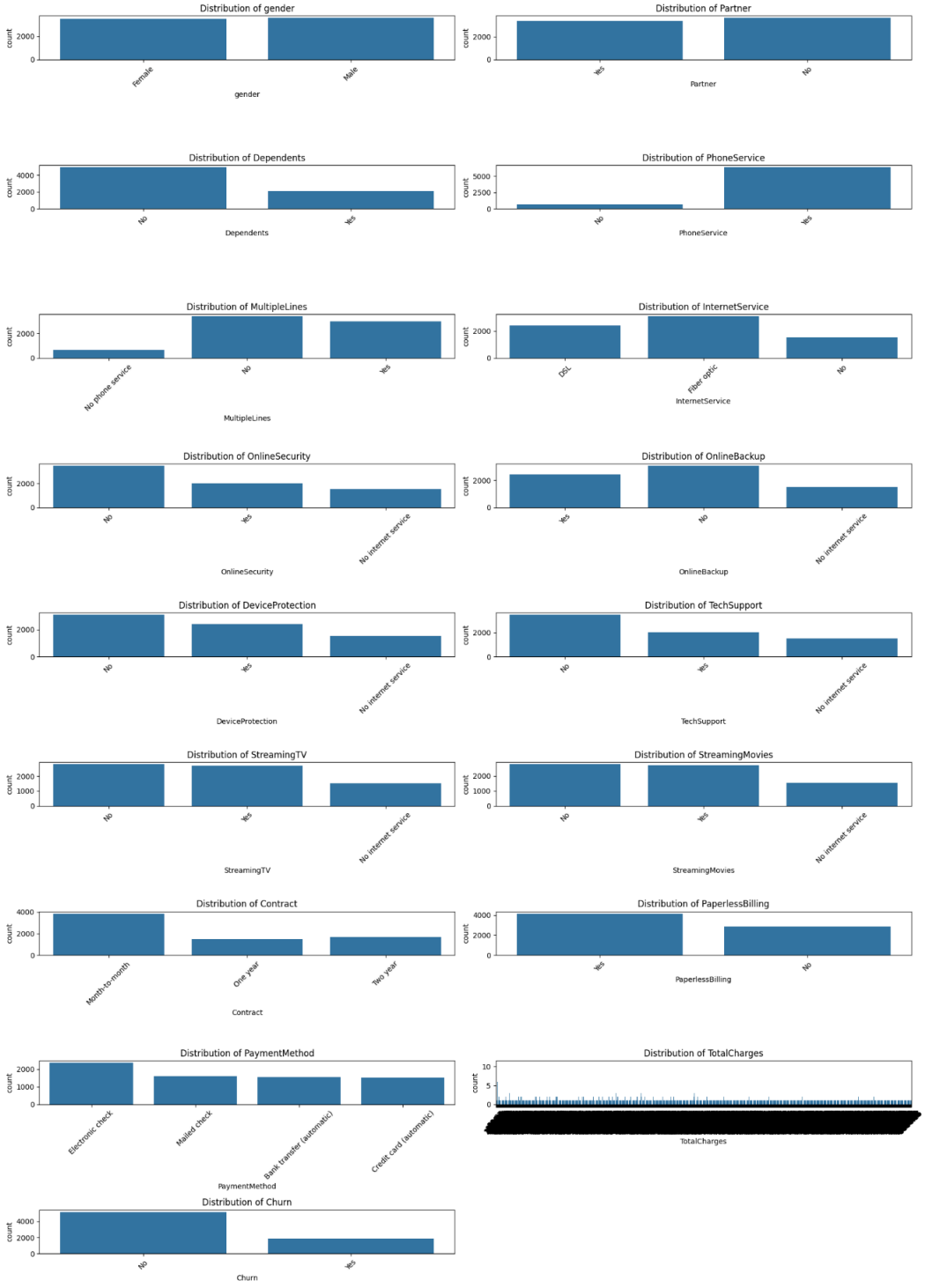
Monthly Charges :(هزینه ماهانه) در محدوده ۲۰ تا ۸۰ دلار متمرکز است و توزیع آن نشان می‌دهد که بیشتر مشتریان هزینه ماهانه کمتری دارند.

Total Charges :(مجموع هزینه‌ها) نیز توزیع متمایلی دارد که نشان‌دهنده هزینه‌های تجمعی مشتریان است. بیشترین مقادیر مربوط به مشتریان با سابقه طولانی است.



رسم نمودار Bar plot برای متغیرهای دسته‌ای

نمودار بارپلات را برای متغیر های دسته ای رسم میکنیم و در ادامه برخی از مهم ترین آن ها را در ادامه تحلیل میکنیم:



gender (جنسیت) :توزیع تقریباً مساوی بین مردان و زنان دارد.

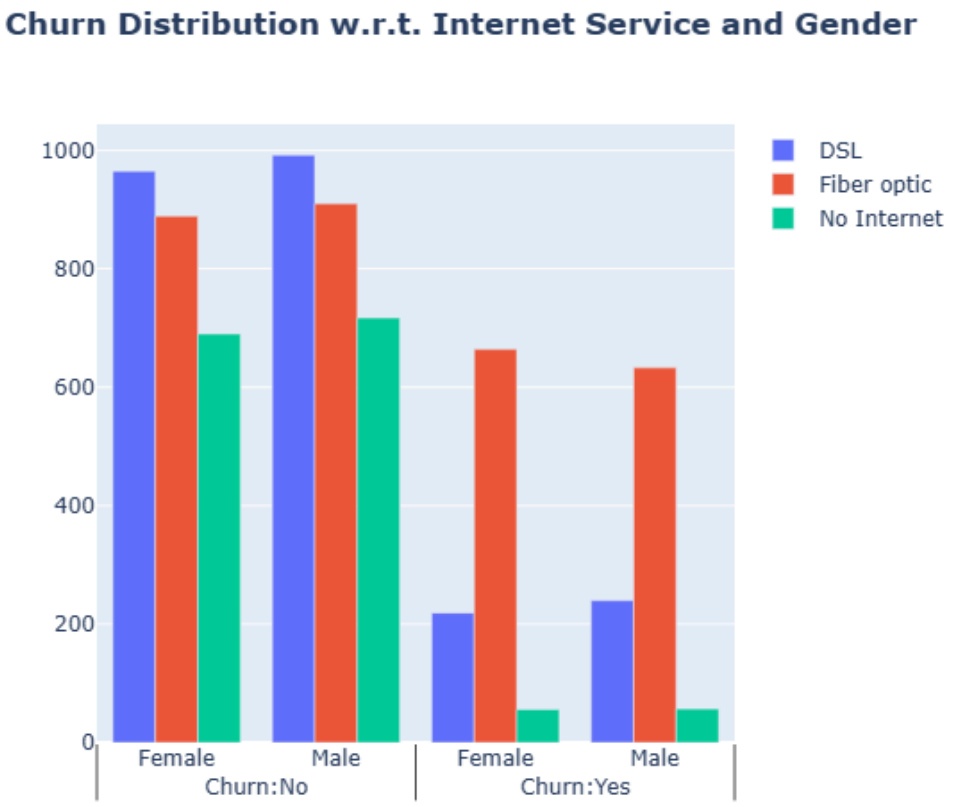
Partner و **Dependents** :نشان می‌دهند که تعداد مشتریانی که شریک یا افراد تحت تکفل ندارند بیشتر است.

Churn (ریزش مشتری): نابرابر است و تعداد مشتریانی که با شرکت مانده‌اند بیشتر از مشتریانی است که ترک کرده‌اند.

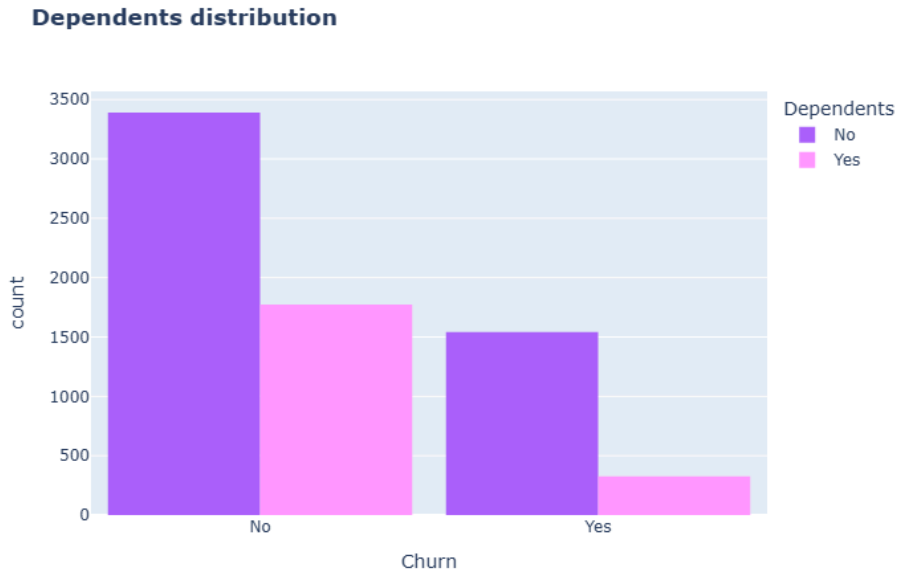
تحلیل توزیع ها

در این بخش از پروژه، هدف اصلی بررسی الگوهای رفتاری مشتریان از طریق تحلیل توزیع ویژگی‌ها و ارتباط آن‌ها با متغیر هدف یعنی ریزش مشتری (Churn) بود. این تحلیل‌ها با استفاده از نمودارهای مختلف انجام شد تا بتوان تفاوت‌های رفتاری میان مشتریان ریزشی و غیرریزش را به‌صورت بصری و قابل درک مشاهده کرد. نتایج به‌دست‌آمده نشان می‌دهد که برخی ویژگی‌ها نقش بسیار مهمی در رفتار مشتریان دارند و می‌توانند به‌عنوان شاخص‌های کلیدی در پیش‌بینی ریزش استفاده شوند.

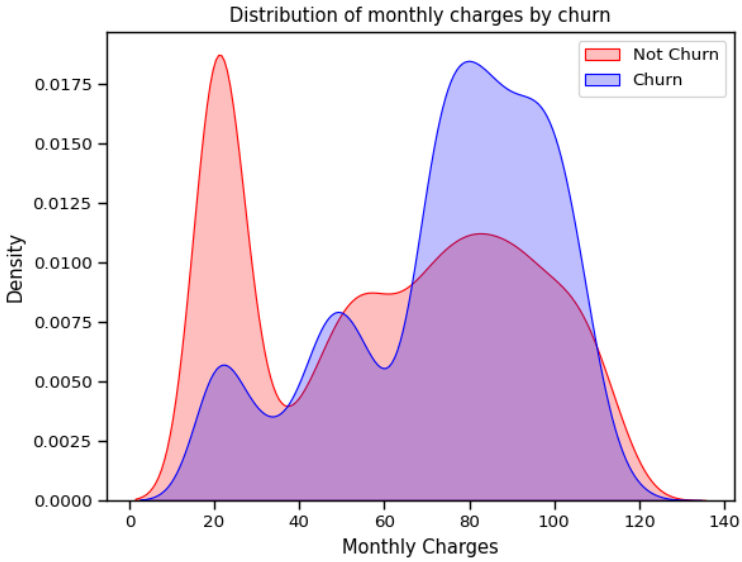
سرویس اینترنت و ریزش: مشتریانی که از Fiber optic استفاده می‌کنند، نرخ ریزش بالاتری دارند. این می‌تواند به دلیل نارضایتی از کیفیت یا قیمت این سرویس باشد.

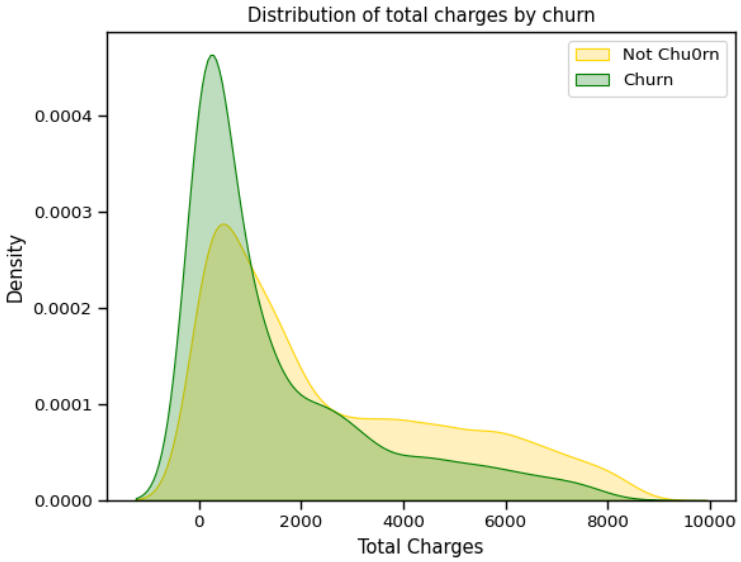


Dependents و ریزش: مشتریانی که افراد تحت تکفل ندارند بیشتر احتمال دارد که ترک کنند. این ممکن است به دلیل داشتن تعهدات مالی کمتر یا انعطاف‌پذیری بیشتر در انتخاب سرویس‌های جدید باشد.

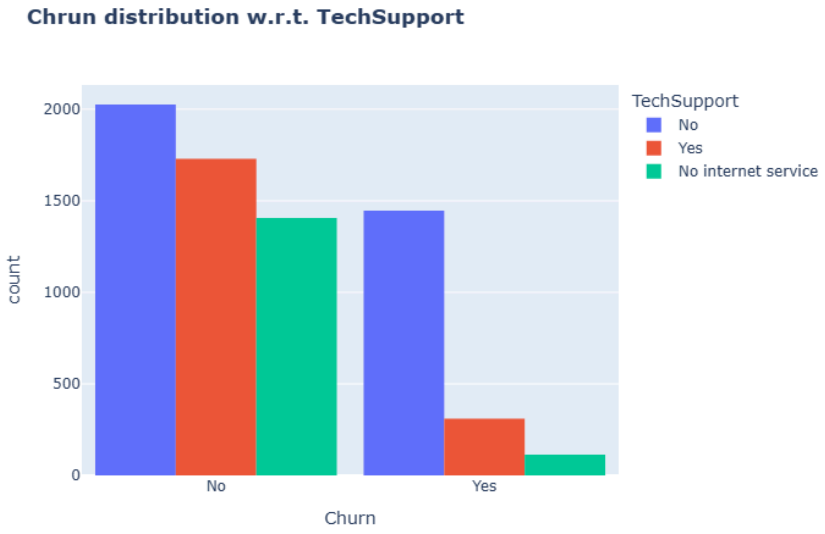


نوع قرارداد: مشتریانی که قرارداد ماهانه دارند، بیشتر احتمال دارد که ترک کنند، در حالی که مشتریانی که قرارداد یک‌ساله یا دو ساله دارند، تمایل کمتری به ریزش نشان می‌دهند.

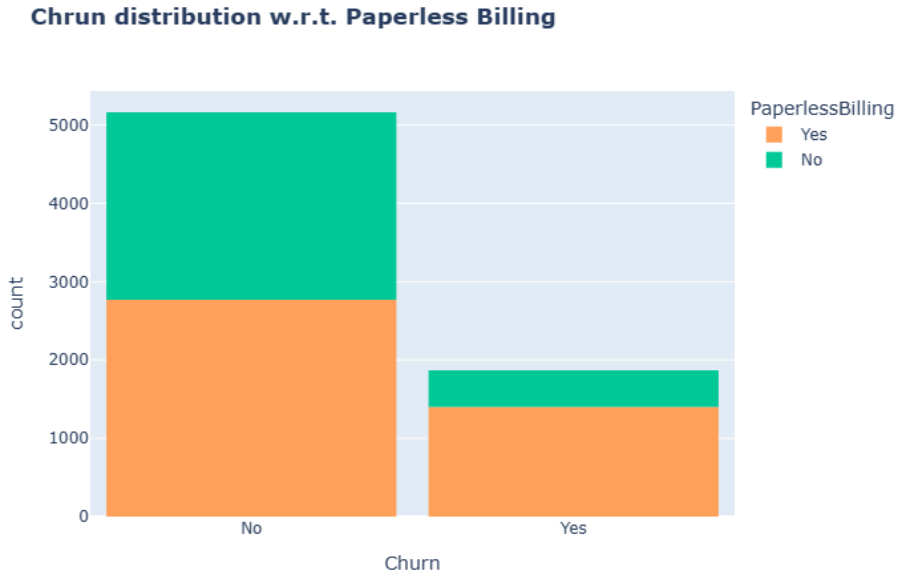




پشتیبانی فنی: مشتریانی که پشتیبانی فنی ندارند، بیشتر احتمال دارد که ترک کنند. این نشان می‌دهد که ارائه پشتیبانی قوی می‌تواند در کاهش ریزش مؤثر باشد.



صورتحساب بدون کاغذ: مشتریانی که صورتحساب الکترونیکی دارند، بیشتر احتمال دارد که با شرکت بمانند.



تحلیل دو متغیره و همبستگی

رسم Heatmap برای بررسی همبستگی متغیر های عددی

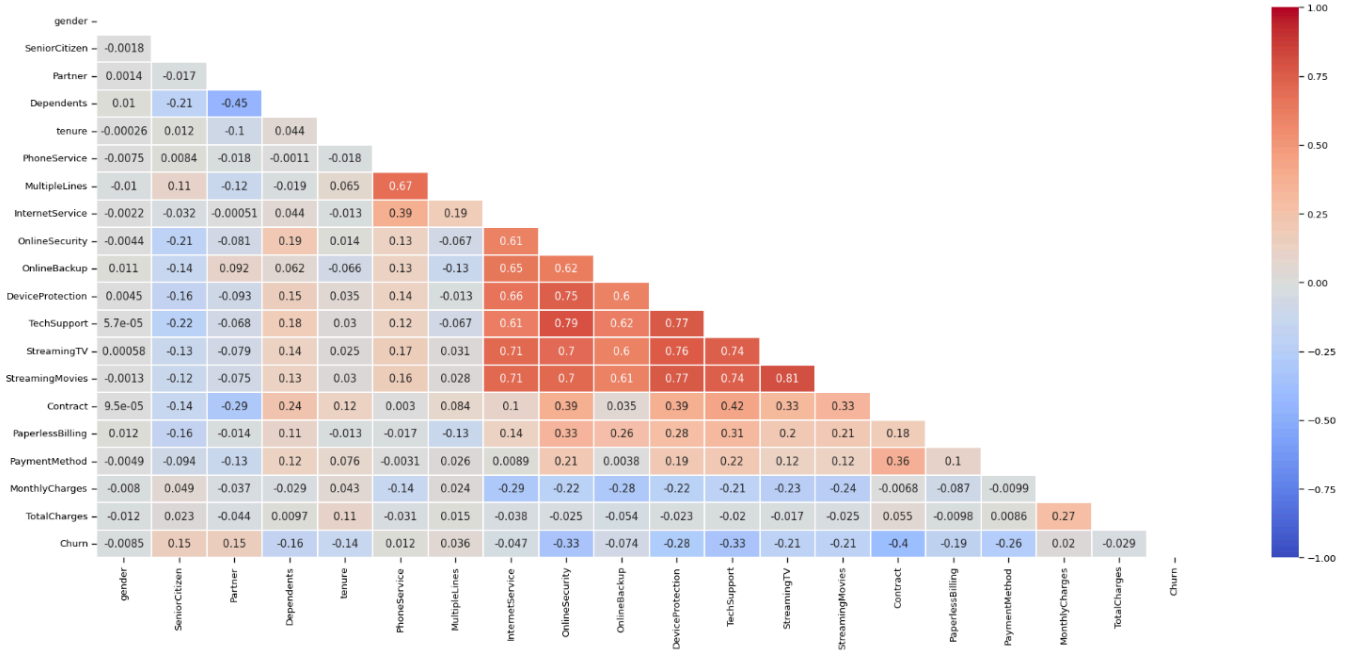
یک نقشه حرارتی برای بررسی همبستگی بین ویژگی‌های عددی ایجاد شد. همبستگی‌ها روابط مهمی را نشان میدهد:

Tenure (مدت زمان حضور) به طور منفی با Churn همبستگی داشت، به این معنی که مشتریانی که مدت طولانی‌تری با شرکت بوده‌اند، کمتر تمایل به ترک دارند.

Total Charges (مجموع هزینه‌ها) نیز با Churn همبستگی منفی نشان داد که مشتریانی که مجموع هزینه بالاتری دارند، احتمال کمتری برای ریزش دارند.

Monthly Charges (هزینه ماهانه) و Churn همبستگی مثبت ضعیفی نشان داد که نشان می‌دهد مشتریان با هزینه‌های بالاتر ممکن است بیشتر تمایل به ترک داشته باشند.

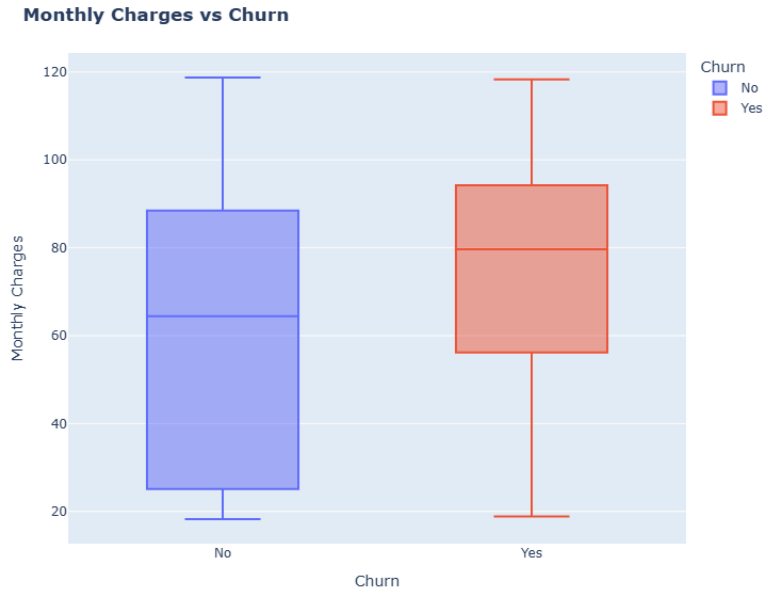
این همبستگی‌ها به شناسایی ویژگی‌های کلیدی کمک کرد که می‌تواند پیش‌بینی ریزش مشتری را بهبود بخشد.



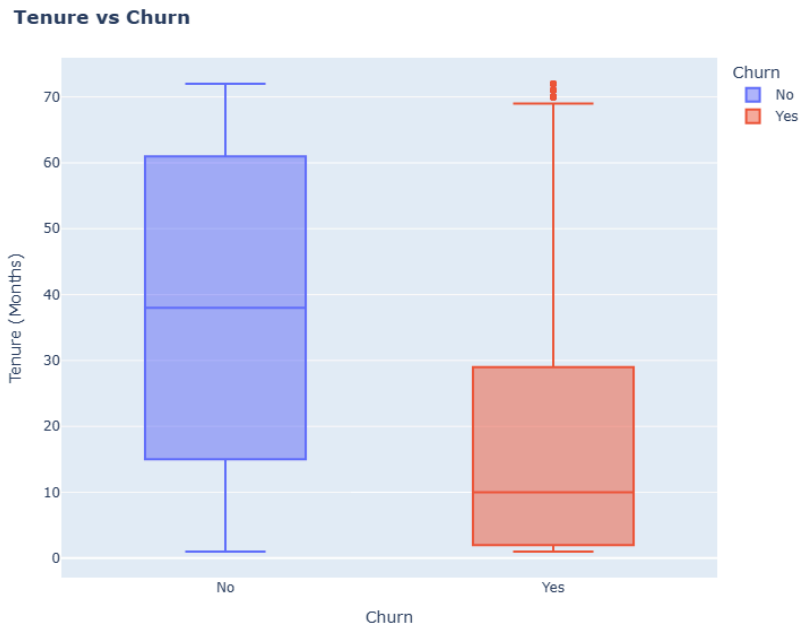
استفاده از نمودارهای مقایسه ای مانند Boxplot برای بررسی رابطه مهم ترین ویژگی های عددی با متغیر هدف(Churn)

تجسمات مختلفی، از جمله هیستوگرام‌ها و نمودارهای جعبه‌ای، برای بررسی روابط بین ویژگی‌های عددی و ریزش مشتری استفاده شد:

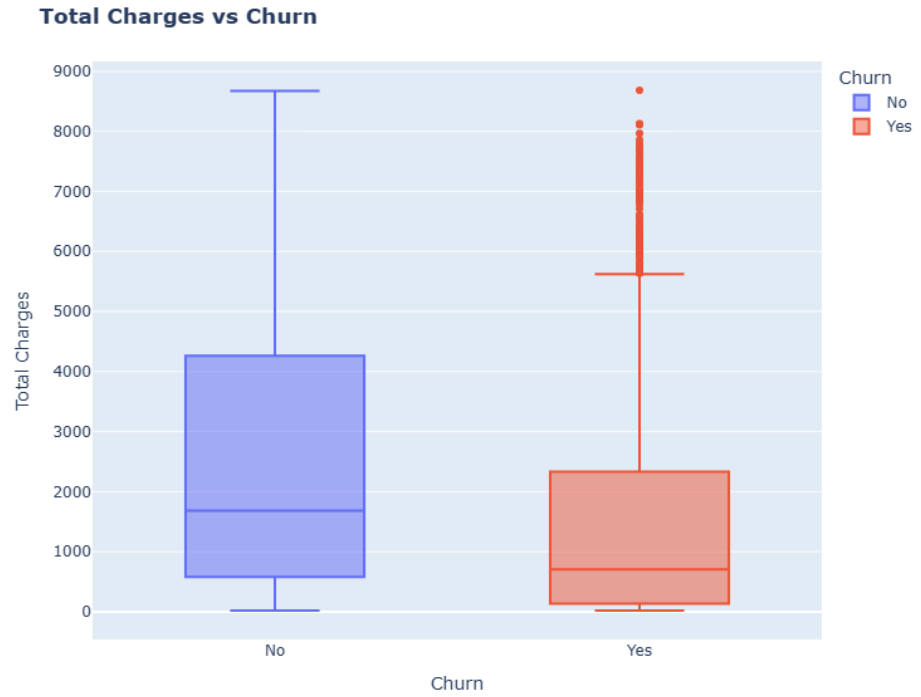
Monthly Charges vs Churn: نمودار جعبه‌ای نشان داد که مشتریانی که هزینه ماهانه بالاتری پرداخت می‌کنند، بیشتر در معرض ریزش هستند. این می‌تواند نشان‌دهنده حساسیت قیمت در میان مشتریان با هزینه‌های بالاتر باشد.



Tenure vs Churn: نمودار جعبه‌ای نشان داد که مشتریانی که مدت زمان بیشتری با شرکت بوده‌اند، احتمال کمتری برای ریزش دارند. این نشان می‌دهد که وفاداری مشتریان با گذشت زمان افزایش می‌یابد.



Total Charges vs Churn: توزیع نشان داد که مشتریانی که مجموع هزینه بالاتری دارند، احتمال کمتری برای ریزش دارند، که با مشاهده مدت زمان حضور مشتریان هماهنگ است.



مشتریان وفادار مدت طولانی‌تری با شرکت هستند: یکی از مهم‌ترین نتایج این تحلیل این است که مشتریانی که مدت بیشتری با شرکت بوده‌اند، احتمال کمتری برای ترک دارند. شرکت باید برای مشتریان جدید برنامه‌هایی طراحی کند تا وفاداری آن‌ها افزایش یابد.

هزینه‌های بالا و ریزش بیشتر: هرچند که هزینه‌های بالا می‌تواند درآمد بیشتری ایجاد کند، اما احتمال ریزش را نیز افزایش می‌دهد. شرکت باید روش‌هایی برای کاهش ریزش در میان مشتریان با هزینه بالا ارائه دهد.

پشتیبانی فنی به عنوان عامل کلیدی: مشتریانی که پشتیبانی فنی ندارند، بیشتر احتمال دارد که ترک کنند. تقویت خدمات پشتیبانی می‌تواند راهبردی مؤثر در کاهش ریزش باشد.

طراحی طرح‌های متناسب با مشتریان مختلف: ارائه طرح‌های ویژه برای مشتریان با نیازهای مختلف مثلاً برای مشتریان با و بدون dependents یا مشتریان جدید و قدیمی می‌تواند باعث کاهش ریزش و افزایش وفاداری شود.

شرکت‌ها می‌توانند با تمرکز بر بهبود این عوامل و ارائه خدمات مناسب، نرخ ریزش مشتری را کاهش دهند و در نهایت وفاداری مشتریان را افزایش دهند.

فاز دوم

در این فاز از پروژه، تمرکز ما بر روی تمیز کردن و آماده‌سازی دیتاست برای مدل‌سازی بود. این مرحله شامل رسیدگی به مقادیر گمشده، رمزگذاری ویژگی‌های دسته‌ای، و مقیاس‌بندی ویژگی‌های عددی بود. هدف از این اقدامات تبدیل دیتاست خام به یک فرم ساختار یافته و پاک شده بود تا برای تحلیل و مدل‌سازی پیش‌بینی آماده شود. در اینجا مراحل مختلف پیش‌پردازش داده‌ها به تفصیل شرح داده شده است.

مدیریت مقادیر گمشده

اولین گام در پیش‌پردازش داده‌ها شناسایی و مدیریت مقادیر گمشده در دیتاست بود. مقادیر گمشده می‌توانند منجر به نتایج مغایر یا خطا در مدل‌های یادگیری ماشین شوند اگر به درستی مدیریت نشوند. پس از بارگذاری دیتاست، بررسی اولیه برای مقادیر گمشده انجام شد.

شناسایی داده‌های گمشده: با بررسی داده‌ها، مشخص شد که ستون TotalCharges دارای ۱۱ مقدار گمشده است. این مشکل به دلیل آن بود که برخی از مشتریان، به ویژه مشتریان جدید، هنوز هزینه‌ای برایشان ثبت نشده است زیرا آن‌ها هنوز فاکتور دریافت نکرده‌اند. سایر ستون‌ها هیچ مقدار گمشده‌ای نداشتند.

imputation

استراتژی جایگذاری مقادیر گمشده

برای جایگذاری مقادیر گمشده از دو روش استفاده شد:

برای ویژگی‌های عددی از میانه (Median) برای جایگذاری مقادیر گمشده استفاده شد. میانه یک روش مقاوم در برابر داده‌های پرت است و برای ستون TotalCharges که ممکن است دارای مقادیر پرت باشد، گزینه مناسبی بود.

برای ویژگی‌های دسته‌ای از پرتکرارترین مقدار (Mode) برای جایگذاری مقادیر گمشده استفاده شد. این روش کمک می‌کند تا توزیع اصلی داده‌ها حفظ شود.

پس از اعمال روش‌های Imputation، تمام مقادیر گمشده پر شدند و دیتاست برای مراحل بعدی آماده شد.

```
Missing values after imputation:
 Series([], dtype: int64)
```

Label Encoding برای ویژگی‌های دودویی

در این مرحله، ویژگی‌های دسته‌ای که فقط دو مقدار یکتای مختلف داشتند، مانند gender، SeniorCitizen، PaperlessBilling، PhoneService، Dependents، Partner و Churn به مقادیر عددی تبدیل شدند. این تبدیل با استفاده از Label Encodingانجام شد که مقادیر متنی مانند Yes و No یا Male و Female را به مقادیر عددی (۰ و ۱) تبدیل می‌کند.

برای هر ستون دودویی، از LabelEncoder موجود در scikit-learn استفاده شد. به عنوان مثال، در ستون gender، مقادیر "Female" به ۰ و "Male" به ۱ تبدیل شد.

پس از انجام Label Encoding، ستون‌های دودویی به مقادیر عددی ۰ و ۱ تبدیل شدند و دیتاست آماده برای مدل‌سازی شد.

	gender	SeniorCitizen	Partner	Dependents	PhoneService	PaperlessBilling	Churn
0	0	0	1	0	0	1	0
1	1	0	0	0	1	0	0
2	1	0	0	0	1	1	1
3	1	0	0	0	0	0	0
4	0	0	0	0	1	1	1

One-Hot Encoding برای ویژگی‌های چندگانه

ویژگی‌های دسته‌ای با بیش از دو مقدار یکتا، مانند MultipleLines، InternetService، OnlineSecurity، OnlineBackup، DeviceProtection، TechSupport، StreamingTV، StreamingMovies، Contract و PaymentMethod به فرمت عددی تبدیل شدند تا بتوانند در مدل‌های یادگیری ماشین استفاده شوند. برای این کار از One-Hot Encoding استفاده شد.

این روش ویژگی‌های دسته‌ای را به چند ستون باینری (۰ و ۱) تبدیل می‌کند. به عنوان مثال، ستون InternetService که شامل مقادیر "Fiber optic"، "DSL" و "No internet service" بود، به سه ستون باینری تبدیل شد که هر ستون نمایانگر یکی از این مقادیر بود.

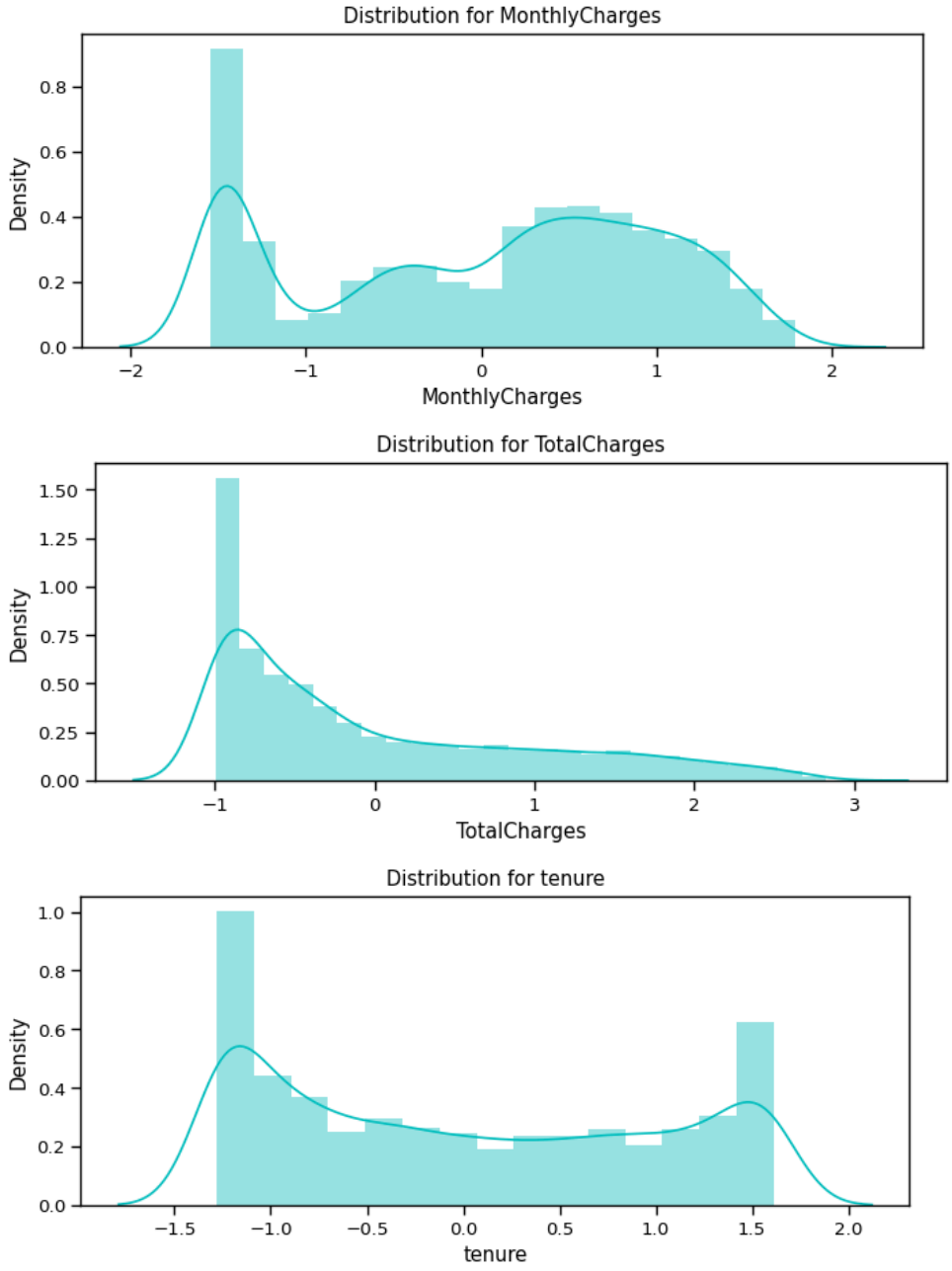
پس از انجام One-Hot Encoding، دیتاست از ۲۱ ستون به ۳۱ ستون افزایش یافت. این افزایش به دلیل ایجاد ستون‌های جدید برای هر دسته از ویژگی‌های چندگانه بود. این تبدیل‌ها موجب شد که تمامی ویژگی‌ها به فرمت عددی تبدیل شوند.

استانداردسازی ویژگی‌های عددی

استانداردسازی ویژگی‌ها یکی از مراحل حیاتی پیش‌پردازش است که برای مدل‌هایی مانند K-Nearest Neighbors (KNN) و Support Vector Machines (SVM) که به مقیاس ویژگی‌ها حساس هستند، ضروری است. در این مرحله از StandardScaler استفاده شد تا تمامی ویژگی‌های عددی مقیاس‌بندی شوند.

برای ویژگی‌های عددی مانند MonthlyCharges، tenure و TotalCharges از StandardScaler برای مقیاس‌بندی استفاده شد. این فرآیند باعث شد که هر ویژگی دارای میانگین صفر و انحراف معیار یک شود.

پس از اعمال استانداردسازی، تمامی ویژگی‌های عددی به مقیاس یکسان درآمدند. این استانداردسازی برای الگوریتم‌های یادگیری ماشین ضروری است تا مدل‌ها از ویژگی‌ها به‌طور یکسان و عادلانه استفاده کنند.



تجسم داده‌ها

توزیع: Monthly Charges از نمودارهای چگالی برای مشاهده توزیع ویژگی MonthlyCharges استفاده شد. این نمودار نشان داد که این ویژگی دارای توزیع چندبخشی است، به‌طوری که مشتریانی که هزینه‌های کمتری پرداخت می‌کنند بیشتر هستند. پس از استانداردسازی، توزیع ویژگی به شکل یکنواخت‌تری درآمد.

نمودار جعبه‌ای: نمودارهای جعبه‌ای برای بررسی رابطه میان ویژگی‌های عددی مانند MonthlyCharges، tenure و TotalCharges با Churn ایجاد شد. این نمودارها نشان دادند که مشتریانی که مدت طولانی‌تری با شرکت بوده‌اند و مجموع هزینه‌های بالاتری دارند، کمتر تمایل به ریزش دارند.

پس از انجام تمام مراحل پیش‌پردازش، دیتاست نهایی شامل ۷۰۳۲ ردیف و ۳۱ ستون بود. ستون‌های جدیدی که از One-Hot Encoding ایجاد شده بودند، به دیتاست افزوده شدند و تمامی ویژگی‌ها به فرمت عددی تبدیل شدند. همچنین، ویژگی‌های عددی به‌طور کامل استاندارد شدند.

این مراحل پیش‌پردازش باعث شدند که دیتاست به‌طور کامل آماده برای مرحله بعدی مدل‌سازی باشد. داده‌ها اکنون به‌طور بهینه‌ای برای استفاده در الگوریتم‌های یادگیری ماشین آماده هستند و درک بهتری از عواملی که بر ریزش مشتری تأثیر دارند، به دست آمده است. این بینش‌ها به مدل‌سازی کمک می‌کنند تا پیش‌بینی‌های دقیق‌تری ارائه دهند و در نهایت به کاهش ریزش مشتری کمک کنند.

فاز سوم: مهندسی و انتخاب ویژگی‌ها

در این مرحله از پروژه، هدف اصلی افزایش قدرت پیش‌بینی مدل از طریق طراحی ویژگی‌های معنادار و انتخاب زیرمجموعه‌ای بهینه از آن‌ها بود. انتخاب صحیح ویژگی‌ها نه‌تنها موجب بهبود عملکرد مدل می‌شود، بلکه تفسیرپذیری نتایج را نیز افزایش می‌دهد و از بیش‌برازش جلوگیری می‌کند. این فاز شامل دو بخش اصلی مهندسی ویژگی و انتخاب ویژگی به دو روش فیلترمحور و مدل‌محور است.

مهندسی ویژگی

در بخش مهندسی ویژگی، سه ویژگی جدید طراحی شد که از نظر رفتاری و اقتصادی قابلیت تفسیر دارند. نخست، متغیر Tenure_Group با هدف مدل‌سازی رفتار غیرخطی ریزش طراحی شد. متغیر اولیه tenure نشان‌دهنده تعداد ماه‌های همکاری مشتری با شرکت است، اما بررسی‌های تجربی نشان می‌دهد احتمال ریزش در ماه‌های ابتدایی عضویت بسیار بالاتر است و سپس با گذشت زمان کاهش می‌یابد. بنابراین tenure به چهار سطح وفاداری شامل مشتری جدید (کمتر از یک سال)، در حال توسعه (۱ تا ۲ سال)، تثبیت‌شده (۲ تا ۴ سال) و وفادار (بیش از ۴ سال) تقسیم شد. ضریب همبستگی این متغیر با Churn برابر با 0.347- به‌دست آمد که نشان می‌دهد با افزایش سطح وفاداری، احتمال ریزش کاهش می‌یابد و این ویژگی قدرت تفکیک مناسبی دارد.

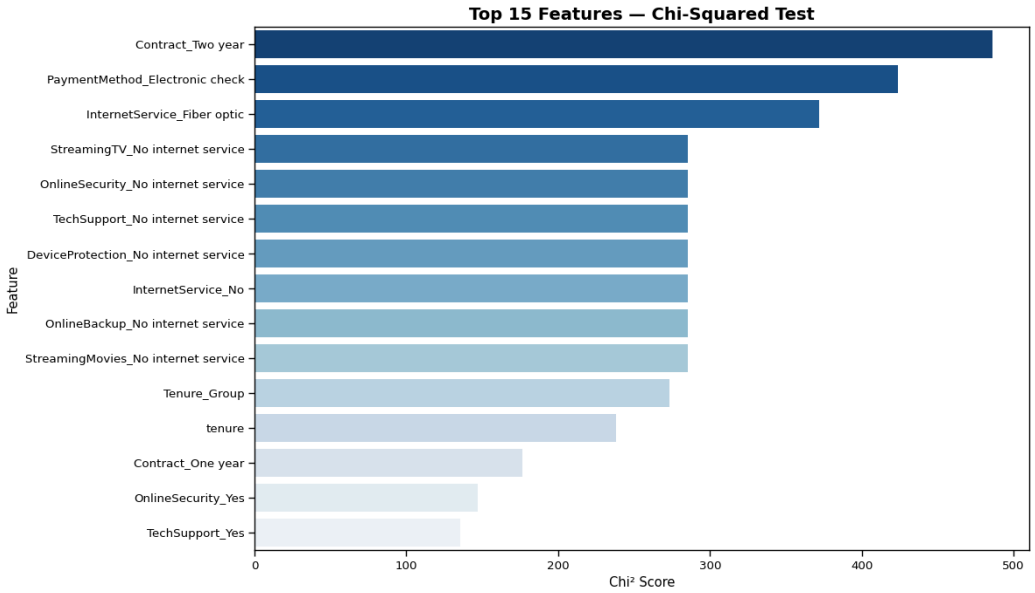
ویژگی دوم، Avg_Monthly_Spend، از تقسیم TotalCharges بر (tenure+1) محاسبه شد تا میانگین پرداخت ماهانه مشتری را نمایش دهد. این متغیر اطلاعات پویاتری نسبت به TotalCharges خام ارائه می‌کند، زیرا شدت مصرف را مستقل از مدت همکاری اندازه‌گیری می‌کند. در بسیاری از موارد، دو مشتری ممکن است TotalCharges مشابهی داشته باشند، اما الگوی پرداخت ماهانه آن‌ها متفاوت باشد. همبستگی مثبت هرچند ضعیف این ویژگی با Churn نشان می‌دهد الگوی هزینه می‌تواند نقش مکملی در پیش‌بینی داشته باشد.

ویژگی سوم، Service_Count، مجموع خدمات افزوده‌ای است که مشتری فعال کرده است. از منظر اقتصاد رفتاری، هرچه تعداد خدمات بیشتر باشد، هزینه جابجایی (Switching Cost) افزایش می‌یابد و در نتیجه احتمال ریزش کاهش می‌یابد. ضریب همبستگی منفی این ویژگی با Churn نیز مؤید همین تحلیل است.

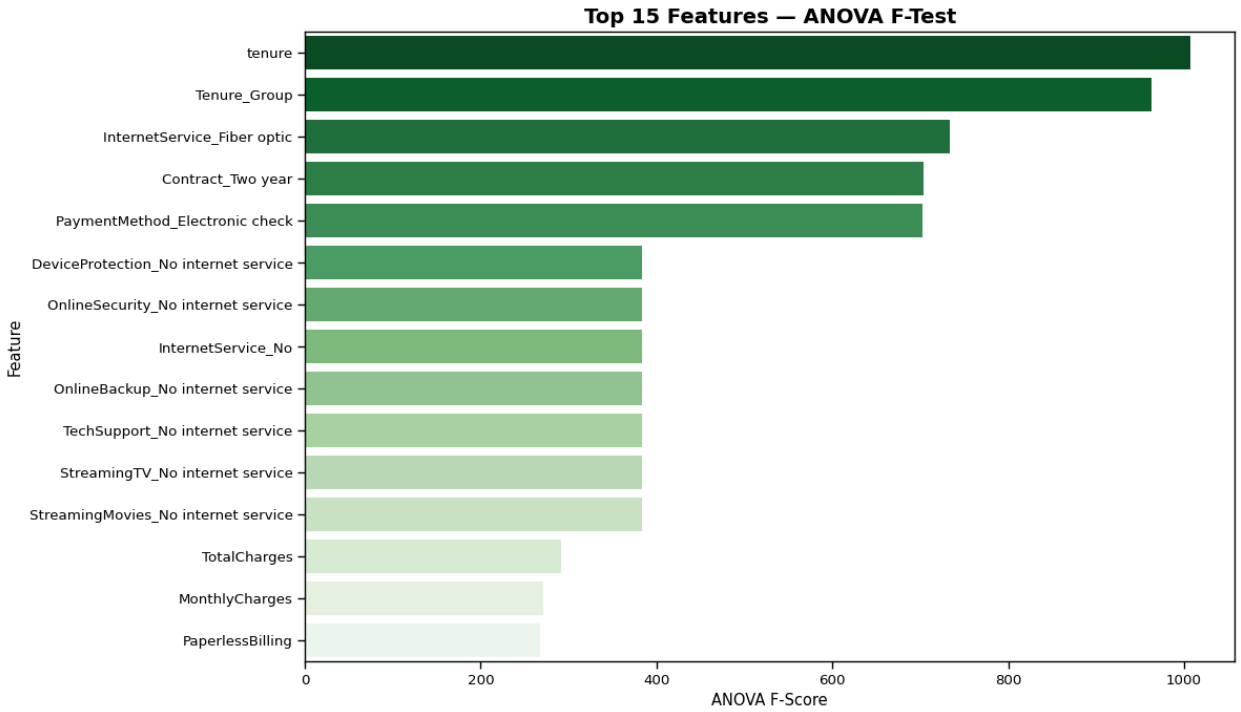
```
Correlation of new features with Churn:
Tenure_Group         -0.347133
Avg_Monthly_Spend     0.070992
Service_Count        -0.069701
Churn                 1.000000
Name: Churn, dtype: float64
```

انتخاب ویژگی فیلترمحور (آزمون مربع کای و ANOVA)

پس از ایجاد ویژگی‌های جدید، مرحله انتخاب ویژگی آغاز شد. در گام نخست از روش‌های فیلترمحور استفاده شد. آزمون Chi-Squared برای ارزیابی وابستگی آماری ویژگی‌های دسته‌ای به متغیر هدف به‌کار رفت و نتایج نشان داد متغیرهایی نظیر InternetService_Fiber، PaymentMethod_Electronic check، Contract_Two year optic و Tenure_Group دارای بیشترین وابستگی آماری و مقادیر p بسیار کوچک (کمتر از 0.05) هستند. این نتایج نشان می‌دهد نوع قرارداد و روش پرداخت نقش تعیین‌کننده‌ای در رفتار ریزش مشتریان دارند.

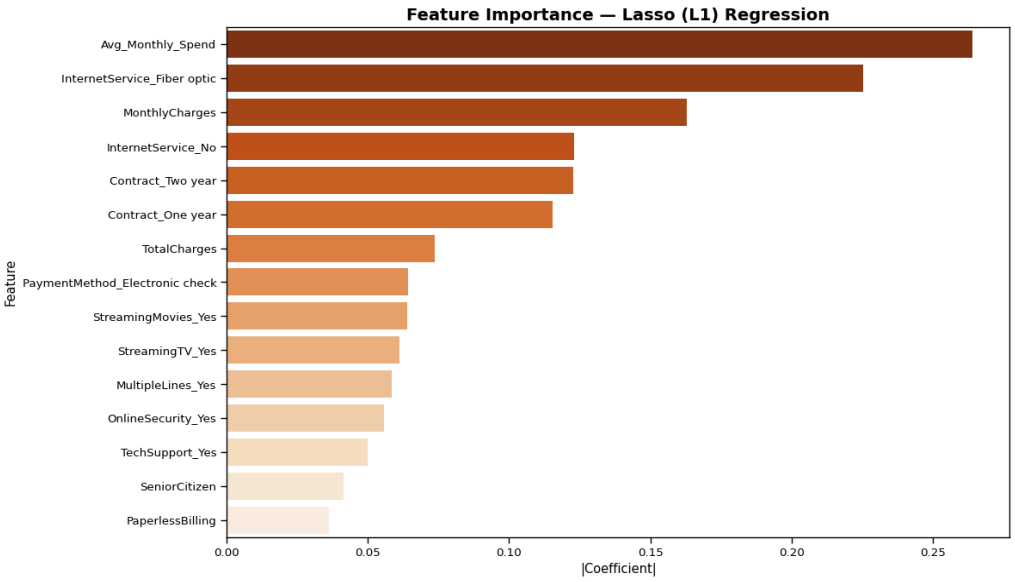


به‌منظور بررسی ویژگی‌های عددی، از آزمون ANOVA F-Test استفاده شد که تفاوت میانگین ویژگی‌ها را بین دو گروه churn و non-churn ارزیابی می‌کند. نتایج نشان داد متغیر tenure با اختلاف قابل‌توجهی بالاترین مقدار آماره F را دارد که بیانگر قدرت تفکیک بسیار زیاد آن است. همچنین InternetService_Fiber، Tenure_Group، Contract_Two year، optic و PaymentMethod_Electronic check نیز از اهمیت بالایی برخوردار بودند. برای افزایش پایداری انتخاب، امتیازات دو آزمون نرمال‌سازی شده و میانگین آن‌ها به‌عنوان معیار ترکیبی در نظر گرفته شد و بر اساس آن ۱۵ ویژگی برتر استخراج گردید. (نمودار در صفحه بعد)

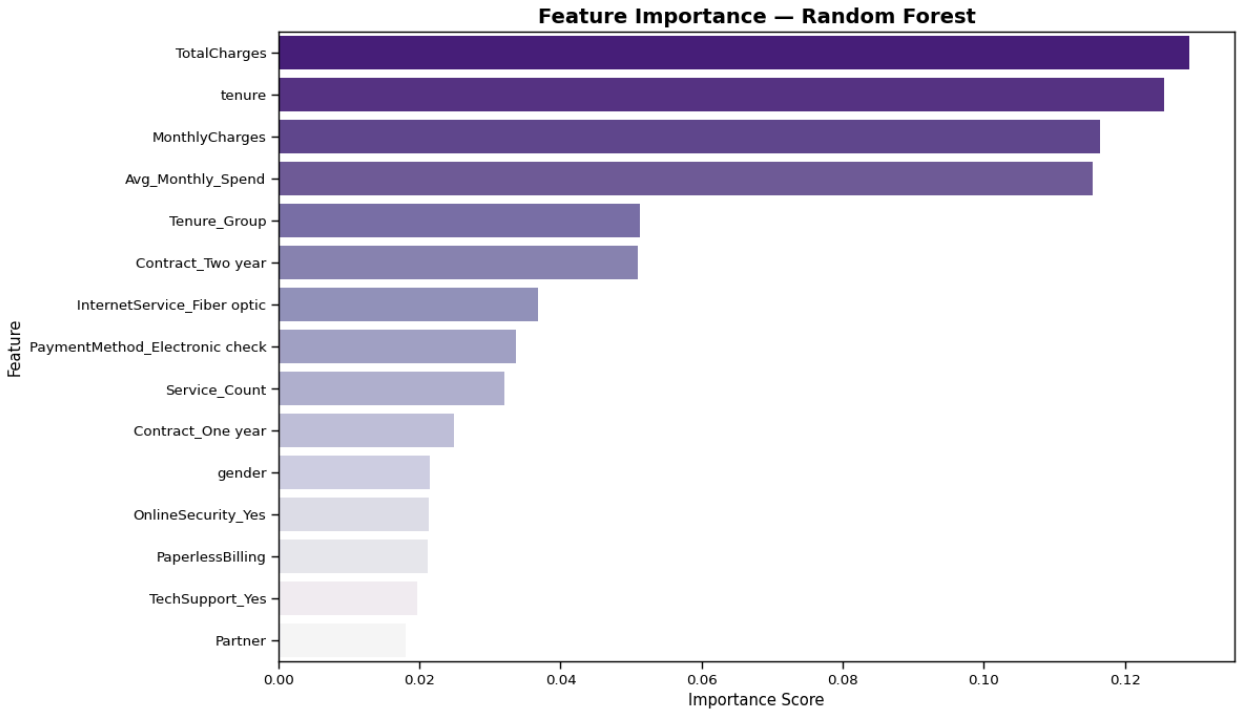


انتخاب ویژگی مدل‌محور (Lasso و Random Forest)

در مرحله بعد از روش‌های مدل‌محور استفاده شد. ابتدا رگرسیون Lasso با جریمه L1 اجرا شد که به‌طور خودکار ضرایب ویژگی‌های کم‌اهمیت را به صفر نزدیک می‌کند و اثر هم‌خطی را کاهش می‌دهد. مقدار بهینه پارامتر تنظیمی α برابر با 0.000188 به‌دست آمد. نتایج نشان داد ویژگی‌هایی مانند Avg_Monthly_Spend، InternetService_Fiber optic، MonthlyCharges و Contract_Two year بیشترین ضرایب غیرصفر را دارند که بیانگر وجود سیگنال پیش‌بینی مستقل در آن‌ها است.



سپس مدل Random Forest به‌عنوان یک روش غیرخطی و مبتنی بر درخت تصمیم اجرا شد. این مدل قادر است تعاملات پیچیده بین ویژگی‌ها را نیز در نظر بگیرد. نتایج اهمیت ویژگی‌ها نشان داد ،tenure ،TotalCharges MonthlyCharges و Avg_Monthly_Spend بالاترین امتیاز اهمیت را دارند. حضور همزمان tenure و Tenure_Group در میان ویژگی‌های مهم نشان می‌دهد که هم اثر خطی و هم اثر غیرخطی سابقه مشتری در پیش‌بینی ریزش نقش دارند.



توجیه نهایی انتخاب ویژگی‌ها

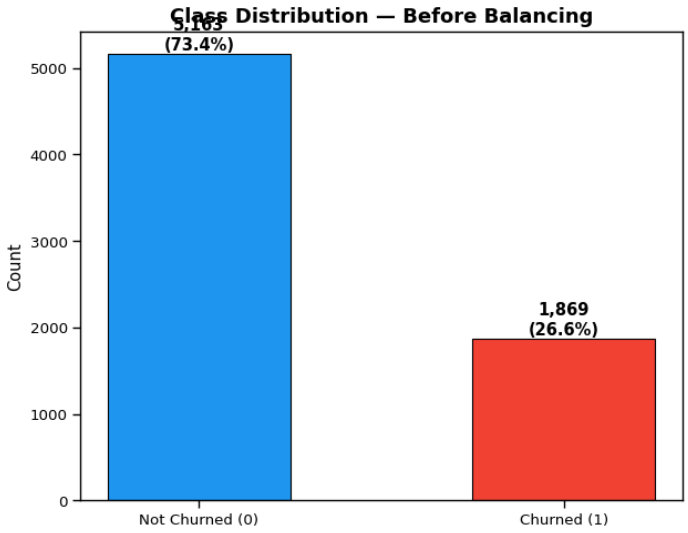
در انتخاب زیرمجموعه نهایی ویژگی‌ها، مبنا صرفاً یک معیار واحد نبوده، بلکه همگرایی شواهد آماری مدنظر قرار گرفته است. نتایج روش فیلترمحور نشان داد که متغیرهای مرتبط با ساختار قرارداد و سابقه مشتری بیشترین وابستگی آماری را با ریزش دارند. به‌طور مشخص، Contract_Two year و Contract_One year در آزمون Chi-Squared دارای مقادیر بسیار بالای آماره و p-value نزدیک به صفر بودند که نشان‌دهنده وابستگی قوی نوع قرارداد به رفتار ریزش است. از سوی دیگر، در آزمون ANOVA، متغیرهای tenure و Tenure_Group بالاترین مقادیر F را کسب کردند (بیش از ۹۰۰)، که بیانگر تفاوت چشمگیر میانگین این ویژگی‌ها در دو گروه churn و non-churn است. این نتایج به‌صورت تجربی تأیید می‌کند که ریزش پدیده‌ای به‌شدت وابسته به چرخه عمر مشتری است. همچنین متغیر PaymentMethod_Electronic check و نوع سرویس اینترنت (InternetService_Fiber optic و InternetService_No) نیز در هر دو آزمون رتبه بالایی داشتند که نشان می‌دهد الگوی پرداخت و زیرساخت سرویس نقش تعیین‌کننده‌ای در تصمیم مشتری برای ماندن یا ترک شرکت دارند.

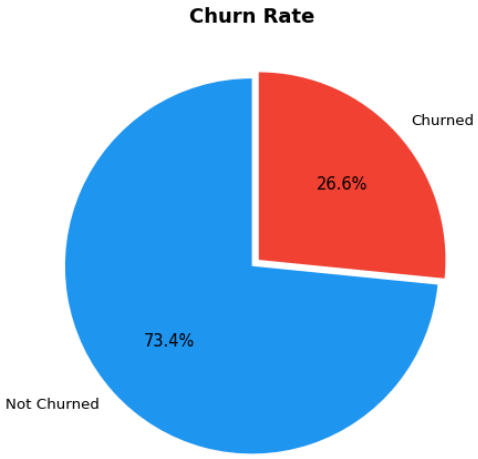
از منظر مدل‌محور نیز یافته‌ها هم‌راستا با نتایج آماری بودند. در رگرسیون Lasso، ویژگی‌هایی نظیر Contract_Two year، MonthlyCharges، Avg_Monthly_Spend و InternetService_Fiber optic ضرایب غیرصفر و نسبتاً بزرگ داشتند که نشان‌دهنده وجود سیگنال پیش‌بینی مستقل حتی پس از اعمال جریمه L1 و حذف هم‌خطی است. به‌طور خاص، حضور همزمان MonthlyCharges، Avg_Monthly_Spend و TotalCharges در میان ویژگی‌های مهم نشان می‌دهد که شدت مصرف و بار مالی مشتری، ابعاد متفاوت اما مکملی از رفتار اقتصادی او را منعکس می‌کنند. در مدل Random Forest نیز متغیرهای tenure، TotalCharges و MonthlyCharges بالاترین اهمیت مبتنی بر کاهش ناخالصی را کسب کردند که بیانگر نقش محوری متغیرهای مرتبط با سابقه و هزینه در روابط غیرخطی و تعاملات بین ویژگی‌ها است. انتخاب نهایی ۱۲ ویژگی بر اساس میانگین رتبه در سه روش مستقل انجام شد تا ویژگی‌هایی حفظ شوند که نه‌تنها در یک چارچوب خاص، بلکه در تحلیل‌های آماری و مدل‌های خطی و غیرخطی به‌طور پایدار اهمیت بالایی دارند. این رویکرد تجمیعی احتمال انتخاب ویژگی‌های کاذب یا وابسته به یک روش خاص را کاهش داده و مجموعه‌ای متوازن، قابل‌تفسیر و از نظر پیش‌بینی قوی را فراهم کرده است.

فاز چهارم: مدلسازی پیشرفته و بهینه‌سازی

بخش مقدماتی: تحلیل متغیر هدف و چالش عدم توازن داده‌ها

پیش از ورود به ارزیابی مدل‌های پیاده‌سازی شده، بررسی دقیق وضعیت توزیع کلاس‌ها در متغیر هدف از اهمیت بالایی برخوردار است؛ چرا که این توزیع، استراتژی ما در انتخاب معیارهای ارزیابی را تعیین می‌کند.





بر اساس تحلیل‌های انجام شده بر روی مجموعه داده خام (پیش از اعمال تکنیک‌های متوازن‌سازی):

وضعیت ریزش مشتریان (دیدگاه تجاری): داده‌ها نشان می‌دهند که از کل جامعه آماری،26.6% از مشتریان (معادل 1,869 نفر) سرویس را ترک کرده‌اند (Churned) و 73.4% (معادل 5,163 نفر) همچنان وفادار باقی مانده‌اند. نرخ ریزش تقریباً 27 درصدی، یک زنگ خطر جدی برای کسب‌وکار محسوب می‌شود و نشان‌دهنده اهمیت حیاتی استقرار یک سیستم پیش‌بینی‌کننده دقیق است.

چالش عدم توازن کلاس‌ها (دیدگاه فنی): نسبت کلاس اکثریت (مشتریان ماندگار) به کلاس اقلیت (مشتریان ریزش‌کرده) برابر با 2.76:12.76 : 12.76:1 است. این عدم توازن (Imbalance) در مجموعه داده به این معناست که:

1. **ناکارآمدی معیار Accuracy:** استفاده از معیار "دقت" یا Accuracy به تنهایی به شدت گمراه‌کننده خواهد بود. یک مدل ساده (Dummy Model) که همیشه خروجی "عدم ریزش" را برگرداند، بدون هیچ یادگیری مفیدی به دقت پایه ∽73% دست می‌یابد.

2. **اهمیت معیارهای تخصصی:** به دلیل هزینه بالای از دست دادن مشتریان (False Negatives - مشتریانی که مدل می‌گوید می‌مانند اما در واقع ریزش می‌کنند)، در فاز ارزیابی مدل‌ها تمرکز اصلی ما باید بر روی معیارهای Recall (پوشش‌دهی مشتریان ریزشی) و F1-Score (میانگین هارمونیک برای ایجاد تعادل) باشد.

نمودارهای میله‌ای و دایره‌ای رسم شده نیز به وضوح این سلطه ساختاری کلاس صفر (Not Churned) را نشان می‌دهند که ضرورت استفاده از تکنیک‌های متوازن‌سازی (Balancing) در مراحل پیش‌پردازش را توجیه می‌کند.

استراتژی متوازن‌سازی داده‌ها: پیاده‌سازی تکنیک پیشرفته SMOTE

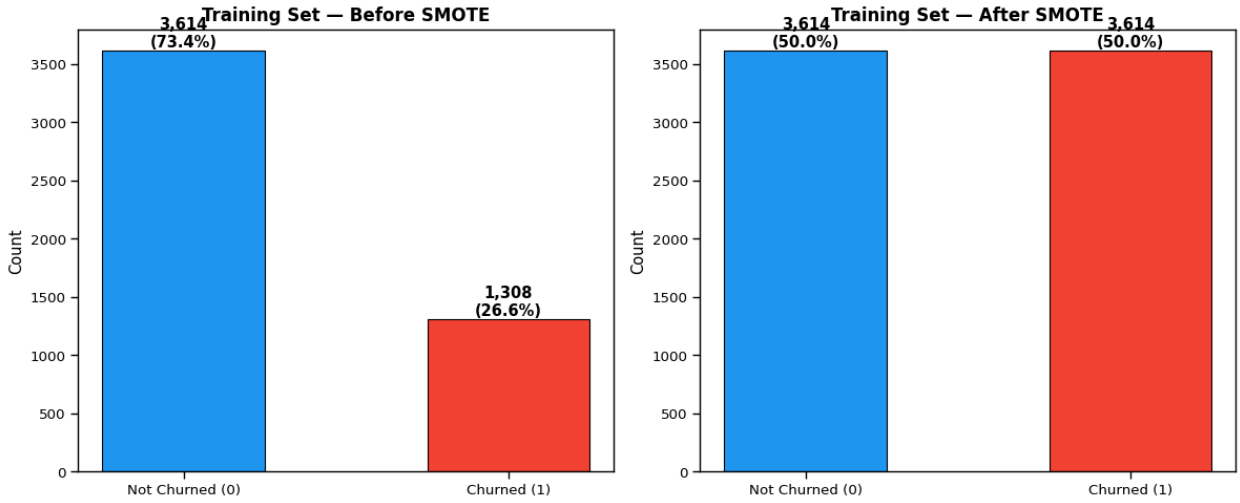
پیرو تحلیل اولیه و مشاهده عدم توازن شدید کلاس‌ها (26.6% در برابر 73.4%)، نیازمند رویکردی ساختاریافته برای جلوگیری از سوگیری (Bias) مدل به سمت کلاس اکثریت (مشتریان ماندگار) بودیم. برای رفع این مشکل، به جای استفاده از روش‌های ساده و پرخطری مانند کپی کردن داده‌های موجود (که منجر به مشکل Overfitting می‌شود) یا حذف داده‌های کلاس اکثریت (که باعث هدررفت اطلاعات ارزشمند می‌شود)، تکنیک SMOTE (Synthetic Minority Over-sampling Technique) به کار گرفته شد.

تحلیل نتایج متوازن‌سازی:

الگوریتم SMOTE با استفاده از رویکرد K-Nearest Neighbors (با تنظیم پارامتر k=5)، فضای بین نمونه‌های کلاس اقلیت را بررسی کرده و داده‌های مصنوعی (Synthetic) جدیدی تولید کرده است. بر اساس خروجی‌ها و نمودارهای رسم شده:

پیش از اعمال SMOTE: مجموعه آموزش دارای 3,614 نمونه از کلاس صفر (ماندگار) و تنها 1,308 نمونه از کلاس یک (ریزش‌کرده) بود.

پس از اعمال SMOTE: با تزریق 2,306 نمونه مصنوعی جدید به کلاس اقلیت، هر دو کلاس دقیقاً به حجم برابر 3,6143,6143,614 نمونه رسیدند (توزیع ایده آل 50:50).



حجم کل داده‌های آموزشی اکنون به 7,228 رکورد ارتقا یافته است که بستری غنی و کاملاً متوازن را برای آموزش الگوریتم‌های یادگیری ماشین فراهم می‌کند. با این پیش‌پردازش موفق، مدل‌های ما اکنون آماده‌اند تا الگوهای رفتاری مشتریانِ در آستانه ریزش را با دقتی برابر با مشتریان وفادار یاد بگیرند و دیگر کلاس اقلیت را نادیده نگیرند.

ارزیابی استراتژی‌های متوازن‌سازی

پیش از پیاده‌سازی و مقایسه مدل‌های پیچیده، ضروری بود تا تأثیر تکنیک‌های متوازن‌سازی بر عملکرد پیش‌بینی‌ها سنجیده شود. برای این منظور، یک مدل پایه رگرسیون منطقی (Logistic Regression) در سه سناریوی مختلف آموزش داده شد:

۱. **Baseline:** آموزش روی داده‌های خام و نامتوازن.

۲. **Class Weight:** جریمه کردن مدل برای اشتباه در کلاس اقلیت (وزن‌دهی متوازن).

۳. **SMOTE:** آموزش روی داده‌های متوازن شده با تولید نمونه‌های مصنوعی.

نتایج به دست آمده بر روی داده‌های آزمون در جدول زیر خلاصه شده است. در مسئله پیش‌بینی ریزش (Churn)، تمرکز اصلی ما بر روی عملکرد مدل در تشخیص کلاس ۱ (مشتریان ریزشی) است.

سناریوی آموزش	Accuracy (دقت کل)	Precision (کلاس ریزش)	Recall (کلاس ریزش)	F1-Score (کلاس ریزش)	ROC-AUC
Baseline	82%	71%	56%	0.63	0.8618
Class Weight	75%	52%	83%	0.64	0.8613
SMOTE	76%	53%	83%	0.64	0.8603

مدل پایه دارای بالاترین(Accuracy 82%) است، اما نگاهی به معیار Recall نشان می‌دهد که این مدل تنها توانسته 56% از مشتریانی که واقعاً قصد ریزش داشته‌اند را شناسایی کند. این یعنی 44% از مشتریان ریزشی (معادل 248 نفر در ماتریس آشفتگی) به عنوان مشتری ماندگار دسته‌بندی شده‌اند (False Negatives). در دنیای تجارت، این یک فاجعه است؛ چرا که کسب‌وکار هیچ اقدام پیشگیرانه‌ای برای حفظ این افراد انجام نخواهد داد و درآمد قطعی از دست می‌رود.

با اعمال روش‌های متوازن‌سازی، معیار Recall جهش چشمگیری پیدا کرده و به 83% می‌رسد. این بدان معناست که مدل اکنون قادر است اکثریت قاطع مشتریان ناراضی را شناسایی کند (افزایش True Positives از 313 به 464 نفر).

همان‌طور که انتظار می‌رفت، افزایش Recall منجر به افت Precision (از 71% به 53%) شده است. به زبان تجاری، مدل ما اکنون هشدارهای کاذب (False Positives) بیشتری می‌دهد؛ یعنی برخی از مشتریان وفادار را نیز در لیست ریزش قرار می‌دهد (حدود 419 نفر در روش SMOTE). با این حال، در مسائل Churn، هزینه ارسال یک کد تخفیف یا تماس با یک مشتری وفادار (خطای نوع اول) بسیار کمتر از هزینه از دست دادن کامل یک مشتری (خطای نوع دوم) است. بنابراین این مصالحه کاملاً منطقی و سودآور است.

در مقایسه دقیق بین Class Weight و SMOTE، مشاهده می‌کنیم که روش SMOTE توانسته با حفظ همان Recall عالی (83%)، تعداد خطاهای False Positive را اندکی کاهش دهد (از 422 به 419) که منجر به بهبود جزئی در(Accuracy 76% در برابر 75%) و(Precision 53% در برابر 52%) شده است.

با توجه به عملکرد برتر، متعادل‌تر و قابل اتکاتر الگوریتم SMOTE در مدیریت خطاهای پرهزینه، مجموعه دادهِ متوازن شده توسط این روش (X_train_smote, y_train_smote) به عنوان ورودی استاندارد برای تمامی مدل‌های یادگیری ماشین در ادامه پروژه در نظر گرفته خواهد شد.

پیاده‌سازی مدل های پایه

نام مدل	Accuracy	Precision	Recall	F1-Score
Logistic Regression	82.27%	71%	56%	0.63
K-Nearest Neighbors	80.80%	69%	50%	0.58
Random Forest	80.85%	69%	50%	0.58
Gradient Boosting	80.52%	66%	55%	0.60

اگرچه الگوریتم‌های Ensemble مانند Voting Classifier دارای معماری قدرتمندی هستند، اما بررسی‌ها نشان داد که استراتژی ترکیب الگوریتم Logistic Regression با داده‌های متوازن‌شده توسط SMOTE در این پروژه بهتر عمل می‌کند.

همانطور که در آزمایشات بخش تنظیم کلاس‌ها اثبات شد، اجرای رگرسیون منطقی بر روی داده‌های SMOTE توانست Recall را از 56% به 83% ارتقا دهد، در حالی که F1-Score نیز در بالاترین سطح خود (0.6427) قرار گرفت. این بدان معناست که مدل نهایی قادر است بیش از ۸۳ درصد از مشتریانی که قصد ترک سیستم را دارند به درستی شناسایی کند.

تنظیم فراپارامترها با استفاده از GridSearchCV

برای یافتن بهینه‌ترین تنظیمات هر الگوریتم، از جستجوی شبکه‌ای (Grid Search) با اعتبارسنجی ۵ مرحله‌ای استفاده شد:

- **بهینه‌سازی Logistic Regression:**

مدل رگرسیون منطقی با مقادیر مختلف جریمه (Penalty) و قدرت منظم‌سازی آزمایش شد. بهترین پارامترهای یافت شده عبارتند از:

```
{'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'}
```

مقدار C=0.1 نشان می‌دهد که مدل به یک منظم‌سازی (Regularization) قوی برای جلوگیری از حفظ کردن داده‌ها نیاز داشته است. این مدل پس از تنظیم، به دقت پایدار 78.36% روی داده‌ها دست یافت.

- **بهینه‌سازی Random Forest:**

برای این مدل تجمعی، یک فضای جستجوی گسترده شامل ۳۶ ترکیب مختلف (با مجموع ۱۸۰ برازش) تعریف شد:

```
n_estimators': [100, 200, 300], 'max_features': ['sqrt','}
'log2'], 'max_depth': [4, 6, 8], 'criterion': ['gini',
{[''entropy
```

بهترین ترکیب یافت شده (`max_depth=8` و `n_estimators=300` با معیار `gini`) امتیاز **F1−Score=0.8229** را بدست آورد. محدود کردن عمق درخت به ۸ باعث جلوگیری از Overfitting شده و افزایش تعداد درختان به ۳۰۰، پایداری مدل را تضمین کرده است.

اعتبارسنجی پیشرفته مدل‌ها

دقت (Accuracy) به تنهایی در داده‌های نامتوازن معیار مناسبی نیست. با توجه به هدف تجاری پروژه (شناسایی مشتریان در حال ریزش)، تمامی مدل‌های بهینه‌شده تحت اعتبارسنجی ۵ مرحله‌ای (Fold CV-5) با تمرکز بر معیار Recall (حساسیت) قرار گرفتند.

جدول زیر میانگین توانایی مدل‌ها در شناسایی مشتریان ریزشی را در ۵ بخش مختلف داده‌ها نشان می‌دهد:

نام مدل الگوریتم	میانگین Recall در ۵ مرحله	انحراف معیار	وضعیت پایداری مدل
Tuned Logistic Regression	(79.4%)	0.0108	عالی (بالاترین حساسیت، کمترین نوسان)
Gradient Boosting	(78.3%)	0.0118	بسیار خوب
Tuned Random Forest	(77.3%)	0.0135	خوب
Voting Classifier	(75.7%)	0.0152	متوسط (نوسان بیشتر بین فولدها)

تمامی مدل‌ها دارای انحراف معیار بسیار پایینی (حدود1%) هستند. این موضوع از نظر علمی ثابت می‌کند که مدل‌های ما به هیچ وجه دچار Overfitting نشده‌اند و روی هر داده جدیدی با همین کیفیت عمل خواهند کرد. مدل رگرسیون منطقی تنظیم‌شده (Tuned LR) با میانگین Recall برابر با 79.40%، رسماً از مدل‌های پیچیده‌تر مبتنی بر درخت (مثل Random Forest و Gradient Boosting) و حتی مدل ترکیبی (Voting) بهتر عمل کرد.

انتخاب مدل نهایی

بر اساس تمامی مستندات فازهای تحلیلی، مدل‌سازی پایه، متوازن‌سازی (SMOTE)، تنظیم فراپارامترها و در نهایت اعتبارسنجی متقاطع، مدل Logistic Regression تنظیم‌شده (Tuned) به عنوان مدل نهایی پروژه انتخاب می‌گردد.

قانون Occam's Razor در یادگیری ماشین بیان می‌کند که اگر دو مدل عملکرد مشابهی دارند، مدل ساده‌تر ارجح است. رگرسیون منطقی یک مدل خطی، بسیار سریع، با قابلیت تفسیرپذیری بالا (Explainable AI) است که نیازی به منابع پردازشی سنگین ندارد، اما توانسته مدل‌های سنگینی مثل Random Forest 300-درختی را در مهم‌ترین معیار تجاری ما (Recall) شکست دهد.