Datametica - Churn

Friday, September 8, 2023 11:28 AM

Step	Description
1. Understanding the Problem	Define the objective: Predict future churn.
1b. Data Collection	Collect data on
	demographics
	plan details
	billing info
	usage details
	feedback
	promotions
	customer support interactions.
2. Data Pre-processing	- Clean the data:
	Handle missing values,
	drop unnecessary columns,
	address duplicate records
	Handle outliers
	Normalize and standardize the data.
	Encode categorical variables
	Engage in feature engineering.
3. Exploratory Data Analysis (EDA)	- Check the
	distribution of the target variable.
	Visualize distributions of features
	Analyze correlations and patterns between churned vs. non-churned users.
4. Model Building	Split the data into training and test sets.
	Implement
	Logistic Regression
	Random Forest,
	XGBoost.
Metrics	Precision

	Recall
	Accuracy
	F1
	TPR
	FPR
	AUC - ROC
5. Hyperparameter Tuning	Adjust parameters for Random Forest and XGBoost
	number of trees
	depth
	learning rate
	add more
	Search for best Hyper-parameters using:
	Grid Search
	Randomized Search.
	Bayesian Optimization
6. Cross Validation	- Validate the model's robustness with
	K-Fold Cross-Validation
	Leave one out
	more
7. Model Deployment (Optional)	Flask API
	Docker
	Kubernetes
	Tableau
8. Continuous Monitoring and Feedback	Monitor the model's real-world performance
	Collect feedback and update the model accordingly.
Tools & Libraries	• Data Cleaning & Pre-processing: Pandas • Visualization: Matplotlib, Seaborn • Modeling: Scikit-learn, XGBoost • Hyperparameter Tuning: GridSearchCV or RandomizedSearchCV from Scikit-learn.

DataFrame	Column Name	Description	Data Type
df_demographics	CustomerID	Unique identifier for each customer	Integer
	Name	Name of the customer	String
	Age	Age of the customer	Integer
	Gender	Gender of the customer (Male/Female/Other)	String
	Region	Geographical region of the customer	String
df_plan	CustomerID	Unique identifier for each customer	Integer
	Plan_Type	Type of telecom plan subscribed (e.g., Basic, Premium)	String
	Plan_Duration	Duration of the plan (e.g., Monthly, Yearly)	String
	Monthly_Cost	Cost of the plan per month	Float
df_billing	CustomerID	Unique identifier for each customer	Integer
	Billing_Date	Date of the monthly bill	Date
	Amount	Amount to be paid	Float
	Payment_Mode	Mode of payment (e.g., Credit Card, Online Transfer)	String
	Payment_Status	Status of the payment (Paid, Due, Overdue)	String
df_usage	CustomerID	Unique identifier for each customer	Integer
	Data_Used	Amount of data used in GB	Float
	Call_Minutes	Total call minutes used	Float
	Messages_Sent	Total number of SMS messages sent	Integer
df_feedback	CustomerID	Unique identifier for each customer	Integer
	Feedback_Date	Date of feedback submission	Date
	Rating	Rating out of 10	Integer
	Comment	Additional comments/feedback	String
df_promotions	CustomerID	Unique identifier for each customer	Integer
	Promotion_Type	Type of promotion (e.g., Discount, Extra Data)	String
	Start_Date	Start date of the promotion	Date
	End_Date	End date of the promotion	Date
df_support	CustomerID	Unique identifier for each customer	Integer
	TicketID	Unique identifier for each support ticket	Integer
	Issue_Type	Type of issue reported	String
	Priority	Priority assigned to the ticket	String
	Resolution_Time	Time taken (in hours) to resolve the issue	Integer

	Resolution_Status	Whether the issue was resolved (Resolved, Pending, Escalated)	String
	Support_Channel	Mode through which the customer raised the issue	String
	Feedback_Score	Score out of 10 given post issue resolution	Integer/N

EDA: General Steps

#	EDA Step	Description
1	Basic Dataset Information	- Check the shape of the DataFrame to get the number of rows and columns. Verify data types of each column. Check for missing values.
2	Descriptive Statistics	Compute measures for quantitative columns, like mean, median, standard deviation, min, and max values.
3	Distribution of Data	- Plot histograms for continuous variables like Age. Plot bar charts for categorical variables like Gender and Region.
4	Outliers Detection	Use boxplots to identify potential outliers, especially for the Age column.
5	Relationships	Explore potential relationships among columns. E.g., if age distribution varies across different regions (Applicable if more demographic-related data becomes available).
6	Unique Values	Check unique categories for columns, especially the categorical ones.
7	Potential Anomalies	Identify any anomalies in the data, such as ages that are unlikely (e.g., above 100 or below 0).
8	Correlations	Check the correlation between multiple numerical columns. (Limited in scope for df_demographics but essential in larger datasets).
9	Value Counts	Get a count of each category in categorical columns like Gender and Region.
10	Final Insights	Summarize the findings at the end of the EDA. Highlight major takeaways, insights, and any anomalies or issues that might need further investigation or cleanup before moving on to data modeling.

EDA: specific to each Data Type

Data Type	EDA Step & Description
Float	1. Descriptive Statistics: Obtain measures like mean,

	median, standard deviation, min, and max.
	2. Distribution Visualization: Plot histograms or kernel density plots to view the distribution.
	3. Box Plots: For identifying potential outliers and understanding the spread & skewness.
	4. Check for Missing Values: Identify and sum any NaN or null values.
	5. Relationship with Target Variable: Use scatter plots or group-by means to see how the float variable changes relative to the target.
Object (Categorical)	1. Value Counts: Get frequency counts of categories.
	2. Visualize Frequency Distributions: Bar charts or count plots to view frequency distribution.
	3. Relationship with Target Variable: Use bar charts with hue as the target variable (like 'Churn') to see the relationship.
	4. Check for Missing Values: Identify and sum any NaN or null values.
	5. Analyze Unique Categories: Count unique values/categories in each categorical column.
Text	1. Text Length Analysis: Study the distribution of lengths of text entries.
	2. Common Word Analysis: Identify frequently occurring words using word clouds or frequency distributions.
	3. Check for Missing Values: Identify and sum any NaN or null text entries.
	4. Relationship with Target Variable: If possible, group texts by categories of interest and study their relationship with the target.
	5. Text Preprocessing: Tokenization, stemming, and removal of stop words for further analysis or model building.
Int	1. Descriptive Statistics: Obtain measures like mean, median, standard deviation, min, and max.
	2. Distribution Visualization: Plot histograms to view the distribution, especially since int values are discrete.
	3. Box Plots: For identifying potential outliers and

	understanding the spread & skewness.
	4. Check for Missing Values: Identify and sum any NaN or null values.
	5. Value Counts: For discrete int variables, getting the frequency of each value can be insightful.
	6. Relationship with Target Variable: Use scatter plots, group-by means, or bar charts (for fewer unique int values) to see how the int variable relates to the target.

Types of models used:

- Logistic Regression
- Decision Trees
- Random Forest
- XGBoost

Metrics:

- Accuracy
- F1
- Precision
- Recall
- TPR
- FPR
- AUC - ROC
- Type 1 Error
- Type 2 Error
- Statistical Power
- Specificity
- Sensitivity

Human in the Loop

When the accuracy is low, we routed the model prediction and the corresponding features to a dedicated HUMAN in the loop.

- Human in the Loop is RULE BASED
- Logistic Regression is ML Based
- So, we have a HYBRID model: RULE based + ML based

==========

END

==========