Spare it

Tuesday, August 22, 2023 9:16 PM

Waste Management Pipeline

Part 1: Instance Segmentation

- Computer Vision
- Pre-trained model

Part 1: Report Generation

- LLM based summary generation
- Prompt Engineering
  - Zero Shot
  - One - shot
  - Few -shot
- Fine - Tuning
  - Full fine tuning
  - Parameter Efficient Fine Tuning
  - LORA: Low Rank Adaptation
  - QLoRA
- Complete Pre - training
- Issues
  - Hallucinations
  - Bad language
  - Outrageous Responses
- LLM ideals:
  - Helpful
  - Honesty
  - Harmless

- RLHF
- Metrics
  - ROUGE
  - BLEU
  - HELM

Datametica_1

Tuesday, August 22, 2023 9:17 PM

Automatic Ticket classification using Machine Learning and Deep Learning

• Word embeddings: Cleaned customer texts using Stemming, Lemmatization, tag removal and other methods, and created word vectors using TF-IDF, word2vec, FastText and Glove to create features for Supervised classification.
• Machine Learning models: Utilized Naïve Bayes, Logistic Regression, CatBoost, and XGBoost models to classify customer tickets into 5 levels of severity and urgency. Added Human in the loop to handle exceptions and improved accuracy to 85%.
• Deep Learning models: Implemented RNN and LSTM models for features with more than a year of data. Applied techniques like Regularization, Early stopping, and Skip connections and achieved an accuracy of 91% AUC.
• Dashboards using Tableau: Designed real-time dashboards to monitor trends and anomalies and track metrics and KPIs and communicated insights and recommendations to stakeholders

Resources:
Automatic Ticket Classification | Kaggle

Steps:

1. Dataset collection
2. Feature categories
    a. Categorical variables
    b. Numerical
    c. Text data
    d. other meta data
3. Text cleaning
    a. Stop words
    b. domain specific words
4. Feature Extraction
    a. Categorical
    b. Numerical
    c. EDA
    d. Text Processing:

        i. TF - IDF
        ii. Word2Vec
        iii. Word embeddings
5. Dimensionality Reduction
    a. PCA
    b. SVD
6. ML Based classification algorithms
    a. Naïve Bayes
    b. DT, RF
    c. XGBoost
7. Improvement
    a. Hyper-parameters
    b. Cross Validation
    c. Regularization
8. Human in the Loop
9. Accuracy
10. More data after an year
11. Deep Learning
    a. RNN
    b. LSTM
12. Accuracy
13. Deployment
14. End point
15. Iteration

Topic modeling to classify customer complaints from various feedback channels.

Datasets and Text Analytics: Collected customer concerns from customer support channel, social media, and audio transcripts and conducted Text analysis using word distributions, N-gram analysis, sparsity visualization, TF-IDF and other meta data.

Topic Modeling using Latent Dirichlet Allocation and Latent Semantic Analysis: Performed Topic modeling with unsupervised classification algorithms like LSA, LDA and Non-negative matrix factorization to find the best model that balances speed, scalability, flexibility, and parameter sensitivity.

Metrics and Visualization: Used PCA for dimensionality reduction and t-SNE to visualize 2D and 3D interpretation of complaint clusters. Attained a perplexity score of 180 and a Topic coherence score of 0.87 after multiple iterations.

Datametica_Churn_prediction

Tuesday, August 22, 2023 9:20 PM

Proactive Churn Prediction: Predictive Analytics for Churn Reduction

Data wrangling and Feature Engineering: Created data pre-processing pipelines to perform cleaning, outlier handling, normalization, and standardization, that utilizes customer features related to Demographics, device type, plan type, billing methods, promotions, and usage patterns.

Machine Learning models: Developed supervised classification models like Logistic Regression, Random Forest and XGBoost to classify potential churn profiles to retain them with promotion policies. Enhanced the model performance through Hyper-parameter tuning and Cross validation to achieve an F1 score of 93% and reduced the churn rate by 34%

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.

Capgemini_Fraud

Tuesday, August 22, 2023 9:17 PM

Project Outline

#	Stage	Key Activities	Example Techniques/Tools/Methods
1	Problem Statement	Define the problem that needs to be solved.	Identify potentially fraudulent and circular transactions based on historical data, user history, and profiles.
2	Data Collection	Gather all the necessary data.	Transaction Details, User History, User Profiles
3	EDA	Conduct preliminary data analysis.	Missing Values, Distribution Analysis, Correlation Analysis, Trend Analysis
4	Data Preprocessing	Prepare the data for model training.	Handling Missing Values, Data Normalization, Data Transformation, Encoding Categorical Variables
5	Feature Engineering	Create new features or modify existing ones to improve model performance.	Time-Based Features, Statistical Features, Text Features (if applicable)
6	Model Selection	Choose appropriate machine learning models.	K-means for Cold Start Problems, Isolation Forest
7	Training	Train the selected models on the preprocessed data.	Use K-means to classify new transactions initially, then train Isolation Forest.
8	Evaluation	Assess model performance using various metrics.	Metrics: F1-score, Precision, Recall, AUC-ROC; Cross-Validation: K-Fold
9	Improvement	Fine-tune the model for better performance.	Hyperparameter Tuning, Feature Selection
10	Deployment	Deploy the model to a live environment once satisfactory performance is achieved.	Deploy once accuracy reaches 92% or better

What are the types of Fraud that can happen with a financial firm using Finacle, and some examples

Functionality	Type of Fraudulent Activity	Description
Core Banking	Account Takeover	Unauthorized users gaining control over a legitimate user's account.
	Fake Account Creation	Creating accounts with fraudulent identification to funnel illicit funds.
	Internal Fraud	Bank employees manipulating data for personal gains.
Online Banking	Phishing Attacks	Using fake websites to gain confidential user info.
	Man-in-the-Middle Attacks	Unauthorized interception of user data during a transaction.
	Transaction Reversal Fraud	A user maliciously reverses a completed online transaction.
Mobile Banking	SIM Swap Fraud	Unauthorized users gaining control over a user's mobile SIM and receiving OTPs.
	Mobile Malware	Malware on a user's device capturing sensitive banking details.
	QR Code Tampering	Scanning a tampered QR code which redirects payment to a fraudulent account.
Payments	Duplicate Payments	The same transaction gets processed multiple times fraudulently.
	Salary Fraud	Manipulating the payroll to disburse salaries to fake or terminated employee accounts.
	Cross-border Fraud	Illicit funds being transferred to offshore accounts to avoid detection and taxation.
Common Types	Circular Transactions	Quick back-and-forth transferring of money between accounts to obfuscate the money trail.
	Mismatched Locations/IP	Login or transaction requests coming from locations not usually associated with the account holder.
	Anomalies in Transaction Frequency/Amount	Unusually large or frequent transactions that deviate from normal behavior patterns.

Dataset with this features are used

Feature Category	Feature	Description	Data Type
User Demographics	Age	Age of the user	Numerical
	Gender	Gender of the user (Male, Female, Other)	Categorical
	Location	Location where the user resides (Urban, Suburban, Rural)	Categorical
	Marital_Status	Marital status (Single, Married, Divorced)	Categorical
User Profile	Employment	Employment status (Employed, Unemployed, Self-Employed)	Categorical
	AccountCreationDate	Date when the user account was created	Date
	SavingsAmount	Amount of money saved in the user's account	Numerical
User Transaction History	AvgTransactionValue	Average value of transactions	Numerical
	TransactionCountLast30Days	Number of transactions in the last 30 days	Numerical
	AvgTransactionPerMonth	Average number of transactions per month	Numerical
	AvgTransactionAmount	Average amount per transaction	Numerical
Device and Application Usage (Supply Side)	DeviceType	Type of device used for transactions (Mobile, Desktop, Tablet)	Categorical
	AppUsageTime	Time spent on the application in minutes	Numerical
	AppUsageDuration	Number of days the user has used the app	Numerical
	Browser	Browser used for transactions	Categorical
	ChatbotUsage	Whether the user has interacted with the chatbot	Binary (0 or 1)
App/Software (Demand Side)	ActiveUsers	Number of users currently active on the app	Numerical
	DailyNewUsers	Number of new users acquired daily	Numerical
	AverageSessionL	Average time a user spends per	Numerical

	ength	session on the app	
	MonthlyRevenue	Total revenue generated by the app in a month	Numerical
Label	IsFraud	Whether the transaction is fraudulent or not	Binary (0 or 1)

Additional Features for Improvement

Time-Series Features	Trends over time, seasonality components, time-series analysis methods
Geographical Features	More granular geographical data like city, latitude, longitude
Network Features	Relationships between users, who sends money to whom
Boolean Flags	Multi-factor authentication status, login from multiple locations, etc.
Text-based Features	Transaction descriptions or notes, processed using natural language processing
Device-Related Features	Operating system, screen resolution, device model

Types of Algorithms used:

Category	Algorithm	Description
Statistical Methods		
	Z-Score Method	Computes Z-scores and flags points beyond a certain threshold as anomalies.
	MAD (Median Absolute Deviation)	Similar to Z-score but based on median, making it more robust to outliers.
Distance-Based Methods		
	K-Nearest Neighbors (K-NN)	Flags observations as anomalies if their distance to their k nearest neighbors

		exceeds a predefined threshold.
	DBSCAN	Identifies low-density regions as anomalies.
	LOF (Local Outlier Factor)	Considers the local density deviation of a data point with respect to its neighbors.
Ensemble Methods		
	Random Forest for Anomaly Detection	Trained to perform anomaly detection by fitting it to the inliers and using it to score both inliers and outliers.
	Gradient Boosting for Anomaly Detection	Similar to the Random Forest but using Gradient Boosting Trees.
	Isolation Forest	An ensemble-based method specifically designed for anomaly detection. Isolates anomalies instead of constructing normal profiles.

- Numerical Features
- Categorical Features

- Z-scores can be applied to only numerical features
- But the numerical features are supposed to follow a NORMAL distribution
  - How can we check for normality?
    - Visually
      - Histogram
      - QQ plot
      - Box Plot
      - IQR plot
      - KDE plot
    - Formula based
      - Shapiro-Wilk Test
      - Kolmogorov-Smirnov Test
      - Anderson-Darling Test
      - D'Agostino and Pearson's Test

        - Skewness and Kurtosis
- If not, we should try converting them to NORMAL using Data Transformation techniques
    - Log Transformation
    - Square Root Transformation
    - Box-Cox Transformation
    - Yeo-Johnson Transformation
    - Z-score Transformation
    - Quantile Transformation
    - Min-Max Scaling
- What if we cannot transform them at any cost? Should we ignore them?
    - We cannot achieve 100% normality
    - Also, NOT all methods need NORMALITY
- Methods that ASSUME normality:

Method/Technique	Purpose/Usage	Assumes Normality Of
Parametric Tests		
t-Test	Used to compare the means of two groups.	Data (but robust to violations with large samples)
ANOVA	Used to compare the means of more than two groups.	Data (but robust to violations with large samples)
Linear Regression	Predictive modeling, especially for continuous outcomes.	Residuals
Pearson's Correlation	To measure the strength and direction of the relationship between two variables.	Each of the two variables
Quality Control		
Control Charts	Monitoring the quality of processes.	Data Points
Other Methods		
Principal Component Analysis	Dimensionality reduction, feature extraction, data visualization.	Original Variables
Discriminant Analysis	Classifying a set of observations into predefined classes.	Each group
Statistical Intervals		
Confidence	To estimate the range in which a population	For means and other

Intervals	parameter lies with a certain level of confidence.	statistics (not strictly required)
Prediction Intervals	In regression analysis, to predict an individual outcome within a range.	Residuals (but can be relaxed with large samples)
	AppUsageDuration	Number of days the user has used the app
	Browser	Browser used for transactions
	ChatbotUsage	Whether the user has interacted with the chatbot

- Methods that does NOT ASSUME normality

Method/Technique	Purpose/Usage	Assumes Normality Of
Non-Parametric Tests		
Mann-Whitney U Test	Used to compare the distributions of two groups.	No
Kruskal-Wallis H Test	Used to compare the distributions of more than two groups.	No
Spearman's Rank Correlation	Measures the strength and direction of association between two ranked variables.	No
Wilcoxon Signed-Rank Test	Used to compare two related samples, paired samples, or repeated measurements on a single sample.	No
Machine Learning Methods		
Decision Trees	Classification and Regression	No
k-Nearest Neighbors	Classification and Regression	No
Random Forest	Classification and Regression	No
Gradient Boosting Machines	Classification and Regression	No
Clustering Methods		
k-Means Clustering	Partitioning a dataset into clusters based on similarity	No (but sensitive to outliers)
Hierarchical Clustering	Produces a tree of clusters, useful	No

	for understanding hierarchical relationships.	
DBSCAN	Density-based clustering algorithm	No
Other Methods		
Chi-Square Test	Used to test the independence of two categorical variables.	No (but assumes large sample sizes for accuracy)
Fisher's Exact Test	Used to examine the significance of the association between two categorical variables (for small samples).	No
Bootstrap Methods	Used to estimate the sampling distribution of an estimator by resampling with replacement.	No

Capgemini_AB_testing

Tuesday, September 5, 2023 12:40 PM

Project Plan Outline for A/B Testing to Enhance Consumer Experience on Finacle by Infosys

Introduction

- Briefly introduce the Finacle platform and the significance of enhancing the consumer experience.
- Outline the key KPIs you aim to improve: engagement and satisfaction scores.

Objectives

1. To address user complaints regarding automated customer support.
2. To improve dashboard design and response times.
3. To optimize security protocols to enhance user trust and safety.

KPIs to Monitor

1. Customer satisfaction score (CSAT)
2. User engagement metrics (time spent on the platform, number of interactions, etc.)
3. Security incidents reported
4. Response times for customer support and dashboard functionalities

Timeline

- Phase 1: Planning (Week 1)
- Phase 2: Development and Pre-Test (Weeks 2-4)
- Phase 3: A/B Testing (Weeks 5-8)
- Phase 4: Analysis and Implementation (Weeks 9-10)
- Phase 5: Monitoring and Feedback (Weeks 11-12)

Phase 1: Planning

1. Team formation and roles
2. Set baseline metrics for KPIs
3. Choose A/B testing tools and software

Phase 2: Development and Pre-Test

1. Create alternative solutions (Version A and Version B) for the automated customer support system.
2. Develop alternative dashboard designs.
3. Plan for security protocol changes.

Phase 3: A/B Testing

1. Roll out Version A to a selected group of users.
2. Roll out Version B to a different selected group of users.
3. Monitor real-time data and ensure everything is recorded properly.

Phase 4: Analysis and Implementation

1. Analyze the results of the A/B tests.
2. Decide which versions to implement based on KPI improvement.
3. Implement the winning versions.

Phase 5: Monitoring and Feedback

1. Monitor KPIs to ensure that they have improved as expected.
2. Open channels for user feedback for further iterations.
3. Make ad-hoc adjustments based on real-time data and user feedback.

Budget and Resources

- Provide an estimate of the budget for the A/B tests.
- List the human and material resources needed.

Risk Assessment

- Outline possible risks, such as negative user feedback, and have contingency plans in place.

Conclusion

- Summarize the project plan and emphasize its importance for improving the user experience on Finacle.

Module 1

Check if Quick summary button is helping the users.

Problem Definition

Problem: Whether the absence of a "Quick Summary View" is limiting user experience and efficiency.

Hypothesis: Introducing a "Quick Summary View" will improve user experience and increase user engagement.

Metrics

Metric	Description
Primary Metric	User engagement rate with the Quick Summary • click-through rate • time spent on the summary
Secondary Metrics	• Change in overall dashboard usage time • Customer Satisfaction Score (post-interaction) • Number of help requests or queries related to the account summary • Conversion rate (if the summary leads to other actions like transactions)

Experiment Design

	Control Group (A)	Treatment Group (B)
Users	Users who see the original dashboard without the Quick Summary View	Users who see the dashboard with the Quick Summary View

Sample Size

Term	Definition
Baseline conversion rate	The current rate of the metric you're trying to improve. For instance, if 5 out of 100 users currently engage with a section of the dashboard, your baseline engagement rate is 5%.
Minimum detectable effect (MDE)	The smallest effect size you want to be able to detect. For example, if you want to see at least a 2% improvement in engagement, then your MDE is 2%.
Statistical Power	Typically set at 0.8, this is the probability of detecting an effect if there is one.
Significance Level (alpha)	Typically set at 0.05, this is the probability of detecting an effect that isn't there (false positive).
Sample Size Calculation	Online calculators are available to input these values and get the required sample size.

Duration of the Test

Factors to Consider for Test Duration	Description
User Traffic	If your platform has 1,000 daily active users, and you need 20,000 users for your test, then you'll need to run the test for at least 20 days.
Business Cycle	If there's a weekly pattern (e.g., more usage on weekdays vs. weekends), you'd want the test to run for a full week or multiple weeks to capture the entire cycle.
Seasonality	Avoid running tests during atypical times of the year, such as holidays or special events, unless the test specifically pertains to those periods.

For Finacle for InfoSys

Factor	Description
Sample Size	If you expect a small change (e.g., a 1-2% increase in engagement), you might need a larger sample—maybe tens of thousands or even more users in each group (Control & Treatment).
Duration	With a large user base, you could reach this sample size in a relatively short period, perhaps a couple of weeks. However, for capturing weekly patterns, consider running the test for at least 2-3 weeks.

Execute the Test: Randomly assign users to either the control or treatment group. Ensure that the assignment is indeed random to prevent biased results.

Data Collection :
Collect data on the aforementioned metrics for both groups.

Analysis using Python
Code with Jupyter notebook

Conclusion
Based on the p-value, determine if the "Quick Summary View" had a statistically significant positive impact on user engagement.

Enhancements:
Feedback Mechanism: Allow users to provide feedback on the new feature.

Segment Analysis: Maybe the feature is particularly beneficial for a specific user segment (e.g., business accounts, frequent users).

Iterate: Use feedback and data to improve the feature and perhaps run another A/B test after making refinements.

Reporting
Prepare a detailed report highlighting the setup, findings, statistical significances, and qualitative feedback, if any.

Rollout
If successful, consider rolling out the feature to all users, and

monitor its performance in the live environment.

infoSys

Tuesday, August 22, 2023 9:17 PM

resume

Wednesday, August 23, 2023 4:47 PM

		Emphasis
	Spare - IT	CV and LLMs
	Datametica	NLP - Deep Learning RNN, LSTM, Transformers General Data Science project XGBoost Tableau Non-technical people
	Cap Gemini	Data Analysis
	Info Sys	SQL, Backend

Data Engineer - Resume

Thursday, August 24, 2023 12:23 AM

- Data Migration
- ETL Pipeline Development
- Infrastructure and Setup
- Data Ingestion
- Advanced Data processing
- Data Storage
- Optimization
- Initial Reporting
- Analytics

Old	source	destination	tools used
	MySQL	Netezza	Python based scripts to call utilities that transfers data from MySQL to BigQuery
	Redshift	PySpark	
	Oracle		
	SQL Server		create record that contains source and destination and data transformation logic in PostgreSQL
			Write a python script Invoke Utility to run record
New	source		Tools used
	MySQL	GCP	• Google Cloud Storage • BigQuery • Composer • CloudSQL • Stack Driver Logs
	Query Optimization		

	using PRUNING with partitioning and Clustering columns		

One of the saddest lessons of history is this: If we've been bamboozled long enough, we tend to reject any evidence of the bamboozle. We're no longer interested in finding out the truth. The bamboozle has captured us. It's simply

too painful to acknowledge, even to ourselves, that we've been taken. Once you give a charlatan power over you, you almost never get it back.
~ Carl Sagan

NLP basics:

1. Text normalization
   a. Stop words removal
   b. Tokenization
   c. Stemming
   d. Lemmatization
   e. Case - Folding
2. Text standardization
   a. Jargon removal - defaulter, bail-out
   b. Removing grammatical mistakes
3. Part-of-Speech tagging
4. Named Entity Recognition
5. Dependency parsing
6. Count Vectorizer – Term Frequency -Inverse Document Frequency
7. Word embeddings – Word Vectorization – Word Vectors.
   a. Word2Vec
   b. GloVe
   c. FastText - Facebook
8. *Contextualized* word embeddings (ELMo, BERT, GPT)
9. Sentiment analysis
10. Text classification

==========Summary==========

Friday, September 8, 2023 8:44 AM

capGemini_AB_testing

Friday, September 8, 2023 8:44 AM

1	Problem Definition	• Quick Summary Page • Color and UI change • Promotions • Present Old data when System is down
2	Metrics	
3	Experiment Design	
4	Sample Size	
5	Duration of the Test	
6	Test Execution	
7	Data Collection	
8	Analysis using Python	
9	Conclusion	
10	Reporting	
11	Rollout	
12		
13		
14		
15		

Capgemini_Fraud

Friday, September 8, 2023 8:52 AM

1	Problem Definition	
2	Data Collection	
3	EDA	
4	Data Preprocessing	
5	Feature Engineering	
6	Model Selection	
7	Training	
8	Evaluation	
9	Improvement	
10	Deployment	
11		
12		
13		
14		
15		

Friday, September 8, 2023 4:06 PM

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.

Datametica - Topic Modeling

Friday, September 8, 2023 3:51 PM

Step	Description
Problem Definition	Topic modeling on customer feedback can be immensely valuable, allowing you to discover underlying patterns and themes in your customer complaints, which can help drive process improvements. I'll guide you through the steps on how to approach this:
1. Data Collection	Gathering data from various sources - Customer support channels • (e.g., ticketing system) - • Web scraping tools for social media: • Beautiful Soup, • Scrapy - • Tweepy for Twitter - • Speech-to-text tools: • Google's Speech-to-Text API, • IBM's Watson

2. Text Preprocessing	Cleaning and preparing the text data for analysis -
	• Text normalization
	  • Stop words removal
	  • Tokenization
	  • Stemming
	  • Lemmatization
	  • Case - Folding
	• Text standardization
	  • Jargon removal - defaulter, bail-out
	  • Removing grammatical mistakes
	• Part-of-Speech tagging
	• Named Entity Recognition
	• Dependency parsing

3. Text Analysis similar to EDA	Analyzing Word Distributions and Relationships:
	• Word Distributions:
	  • Refers to the frequency or probability distribution of words appearing in a text corpus.
	  • Often visualized using histograms or bar plots to identify the most common words or phrases in a given dataset.
	• N-gram Analysis:
	  • N-grams are continuous sequences of n items (words, characters, etc.) from a text.
	  • Helps in understanding context and capturing phrases. E.g., "New York" would be a 2-gram (bigram).
	• Sparsity Visualization:
	  • Often related to the term-document matrix, where many entries are zero because a large vocabulary results in many words not appearing in individual documents.
	  • Visualization can be done using a matrix plot, highlighting which terms appear in

	which documents.

Count Vectorization	• Count Vectorizer:
	  • Converts a collection of text documents to a matrix of word/token counts.
	  • The result is a sparse matrix where rows are documents and columns represent word counts.
	• TF-IDF (Term Frequency-Inverse Document Frequency):
	  • Similar to the Count Vectorizer, but instead of raw counts, the matrix represents the weighted count where weights are determined by the importance of a word to a document relative to its frequency across all documents.

Word Embeddings	• A form of representing words as dense vectors such that words with similar meanings are close to each other in the vector space.
	• Result in dense vectors, typically with hundreds of dimensions.
	• Word2Vec
	  ○ Developed by Google, it captures semantic relationships between words. Has two architectures: Skip-gram and Continuous Bag of Words (CBOW).
	• GloVe (Global Vectors for Word Representation)
	  ○ Developed by Stanford, it focuses on word co-occurrence statistics.
	• FastText
	  ○ Developed by Facebook, it represents words as bags of character n-grams, allowing it to handle out-of-vocabulary words.

Contextualized Word Embeddings	More advanced than traditional word embeddings.
	They consider the context around a word, allowing the same word to have different vectors based on its surrounding words.
	• a. ELMo (Embeddings from Language Models):
	  ○ Developed by AllenNLP, it uses character-based word representations and bidirectional LSTMs.
	• b. BERT (Bidirectional Encoder Representations from Transformers)
	  ○ Developed by Google, it's a transformer-based model pre-trained on a large corpus, known for its bidirectionality.
	• c. GPT (Generative Pre-trained Transformer)
	  ○ Developed by OpenAI, it's a transformer model trained to predict the next word in a sequence.

4. Topic Modeling	Identifying underlying topics in the data • LSA: Latent Semantic Analysis (LSA) • SVD: Decompose term-document matrix using Singular Value Decomposition • LDA: Latent Dirichlet Allocation (LDA): Assign topics to documents and words • NNMF: Non-Negative Matrix Factorization (NMF) - Sklearn for NMF, TruncatedSVD Gensim for LDA
5. Visualization & Interpretation	Visual representation of topics - • pyLDAvis for LDA visualization - • Word Clouds for each topic
6. Evaluation	Assessing the quality of the topics - • Examine the distinctness of topics - • Topic coherence scores (using Gensim)

7. Application and Business Impact	• Utilizing the topics in practical scenarios - • Classify new complaints - • Route complaints to relevant departments - • Drive business improvements based on topics

How does LDA work Latent Dirichlet Allocation

- In traditional LDA, the model operates on a document-term matrix, which can be either BoW or TF-IDF representations. However, LDA doesn't directly work with dense embeddings like word2vec.

- If you wanted to leverage word embeddings, you'd likely look into other topic modeling or document clustering methods that can work in the dense vector space.

#	Concept	Description
1	Document	A piece of text. In the context of LDA, a document can be an article, a paragraph, or any text segment.
2	Corpus	A collection of documents.
3	Topic	A collection of words that frequently appear together and represent a particular theme or subject.
4	Dirichlet Distribution	A family of multivariate probability distributions parameterized by a vector of positive reals. LDA uses the Dirichlet distribution for the creation of topics and documents.
5	Bag of Words (BoW)	A representation of text where each document is represented as a vector of word frequencies, disregarding the order of words.
6	Latent Variables	Variables that are not directly observed but inferred from mathematical models. In LDA, topics are latent variables.
7	Gibbs Sampling	A Markov chain Monte Carlo (MCMC) technique for obtaining a sequence of samples from a multivariate distribution, especially when direct sampling is challenging.
8	Hyperparameters (Alpha and Beta)	Parameters of the Dirichlet distribution. Alpha affects document-topic density and Beta affects topic-word density.
9	Posterior Distribution	The probability distribution of an unknown quantity, treated as a random variable, conditional on the evidence obtained from an experiment or survey.
10	Topic Coherence	A measure used to judge the quality of the topics produced by a topic model. It scores a single topic by measuring the degree of semantic similarity between high scoring words.

Metrics for LDA model

#	Evaluation Metric	Description
1	Perplexity	A measure of how well a probability distribution predicts a sample. Lower perplexity value indicates better generalization to unseen documents. However, lower perplexity might not always correlate with human interpretability.
2	Topic Coherence	Measures the degree of semantic similarity between high scoring words within a topic. Models with higher coherence scores generally produce more interpretable topics.
3	Human Judgement	Involves having humans, especially domain experts, interpret and validate topics. It could also involve crowdsourcing feedback.
4	Visual Inspection with pyLDAvis	Visualization tools that help visually interpret and validate the quality and separability of topics.
5	Model Stability	Tests the robustness of the topics by checking if the same topics are found across multiple samples (subsamples or bootstraps) of data.
6	Topic Diversity	Measures the lexical diversity of topics. A model that produces topics with overlapping terms isn't producing distinct enough topics.
7	Document Classification	If labeled data is available, the topic distributions can be used as features for document classification. Good classification performance indicates the topics capture meaningful patterns.
8	Document Retrieval or Recommendation	Topics can be used as features to build retrieval or recommendation systems. Improved recommendation or retrieval quality indicates the topics have meaningful structure.

==========

How does LSA work: Latent Semantic Analysis

- For Latent Semantic Analysis (LSA), we typically use Term-Document matrices. These matrices can be generated using either the Count Vectorization method or the TF-IDF Vectorization method.
- Word embeddings like Word2Vec, GloVe, etc., are not directly used in LSA. Instead, LSA works on the principle of singular value decomposition (SVD) of the Term-Document matrix.

#	Concept/Component	Explanation/Details
1	Term-Document Matrix (TDM)	A matrix representation of a corpus where rows represent terms and columns represent documents. Entries are typically raw term frequencies or TF-IDF scores.
2	Singular Value Decomposition (SVD)	A matrix factorization method. When applied to the TDM, it breaks it down into three matrices: ﷺ (term-topic matrix), ΣΣ (diagonal matrix of singular values), and ﷺT (document-topic matrix).
3	Dimensionality Reduction	In LSA, the goal is to capture the underlying structure by reducing the number of topics (latent factors). This is done by keeping only the top ﷺ largest singular values in ΣΣ and discarding the rest.
4	Latent Topics	The reduced matrices ﷺ and ﷺT capture 'latent' topics in the data. Each row in ﷺ and each column in ﷺT can be viewed as a topic represented as a distribution over terms or documents, respectively.
5	Semantic Space	The space in which documents and terms are mapped after reduction. In this space, the cosine similarity between vectors can be used to measure the semantic similarity between terms or documents.
6	Noise Reduction	By keeping only the top ﷺ singular values, LSA effectively filters out noise, helping in capturing the more significant patterns or structures in the data.
7	Term & Document Similarity	In the reduced semantic space, cosine similarity between term vectors or document vectors can be used to find similar terms or documents.
8	Challenges	1. The method is linear, so it might not capture complex relationships. 2. LSA does not have a probabilistic foundation, unlike LDA. 3. SVD can be computationally expensive for very large matrices.
9	Applications	Document retrieval, information retrieval, query disambiguation, and other text mining tasks.
10	Comparisons	LSA focuses on reducing dimensionality to detect latent patterns. LDA, in contrast, uses a probabilistic approach to model topic distributions.

==========

Advanced methods for TOPIC MODELING

LDA and LSA are generally based on Count Vectorizer kind of DISCRETE word embeddings
For continuous word embeddings like Word2Vec, which are DENSE vector representations of words there are other methods like:

Below, Doc2Vec is directly related to Word2Vec

#	Method	Description
1	Document Embedding	Extends word2vec to represent entire documents. Documents can then be clustered using algorithms like KMeans.

	(Doc2Vec)	
2	Clustering on Embeddings	Convert each document into a vector by averaging its word vectors. Then apply clustering algorithms such as KMeans or DBSCAN. Clusters represent topics.
3	BERTopic	Uses sentence transformers to create embeddings, followed by UMAP for dimensionality reduction and HDBSCAN for clustering.
4	Top2Vec	Combines Doc2Vec, word2vec, or universal sentence encoder embeddings with UMAP for dimensionality reduction. Automatically identifies topics.
5	LDA2Vec	Hybrid of word embeddings and LDA, designed to learn topic vectors in the same embedding space as word vectors.
6	Neural Variational Document Model (NVDM)	Neural approach combining variational autoencoders (VAE) and neural networks to model documents as topic mixtures.
7	NMF on Embeddings	Uses Non-negative Matrix Factorization on dense document embeddings derived from word embeddings.

==========

How does LDA work: Linear Discriminant Analysis

- This is NOT used for Topic Modeling
- This is intended for Dimensionality Reduction and Classification
- This is a Supervised model

#	Concept/Component	Explanation/Details
1	Objective	To find the linear combination of features that best separate two or more classes in a dataset.
2	Assumption	Assumes that the independent variables are normally distributed and that the classes have identical covariance matrices.
3	Between-Class Variance	Variance between the mean of different classes. LDA aims to maximize this to ensure that classes are well separated.
4	Within-Class Variance	Variance within individual classes. LDA aims to minimize this to ensure that members of the same class are close to each other.
5	Fisher's Linear Discriminant	A criterion that seeks to maximize the ratio of between-class variance to the within-class variance, thereby ensuring maximum separability.
6	Eigenvalues & Eigenvectors	Upon computing the between-class scatter matrix and the within-class scatter matrix, the goal is to determine the eigenvectors and eigenvalues which dictate the new feature space.
7	Dimensionality Reduction	LDA can be used for dimensionality reduction by projecting data onto a lower-dimensional space. The number of dimensions is at most $C-1$ where C is the

		number of classes.
8	Multiclass LDA	While originally developed for two classes, LDA can be generalized to more than two classes.
9	Applications	Pattern recognition, feature extraction, and dimensionality reduction in fields like face recognition and predictive modeling.
10	Challenges	Assumes linear separability, requires normally distributed classes, and equal class covariance. Sensitive to outliers.

==========
==========

++

Friday, September 8, 2023 4:06 PM

Spare IT - Instance Segmentation

Friday, September 8, 2023 4:06 PM

Spare IT - LLM Report Summary

Friday, September 8, 2023 4:06 PM

Friday, September 8, 2023 4:07 PM