I. **Cover Page:**

Enterprise Cloud Computing and Big Data (BUDT737)

Project Title: **Explanatory Ecommerce Churn Analysis.**

Team Members: Hetvi Shah

Prashant Goswami

Rohit Abbireddi

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	**Signature**
1	Hetvi Shah	Hetvi Shah
2	Prashant Goswami	Prashant Goswami
3	Rohit Abbireddi	Rohit Abbireddi

II. **Executive Summary**

This comprehensive e-commerce analysis project is centered on unraveling the intricate dynamics of customer churn, employing advanced analytics tools such as PySpark, SQL queries, machine learning algorithms, and innovative techniques like GraphFrames and KMeans clustering. Our primary focus is to understand the factors influencing customer churn, with an emphasis on younger customer segments, the impact of gender and brand on churn, and the significance of personalized customer interactions.

Throughout the project, we delved into feature engineering, introducing the novel "PurchaseToAgeRatio" metric to capture nuanced relationships between age and purchasing behavior. This addition contributes significantly to our churn analysis, providing a more comprehensive understanding of customer behavior. Additionally, the segmentation of churned customers based on gender and brand unveils specific groups that are more prone to churn, offering valuable insights for targeted retention strategies.

One of the innovative aspects of our study is the application of GraphFrames for network analysis and KMeans clustering to segment customers based on demographics and purchasing behavior. These methodologies enable us to identify influential customers whose churn may have cascading effects on broader customer networks. This approach goes beyond traditional churn analysis, offering strategic insights for customer retention and resource allocation.

In conclusion, our research pioneers innovative approaches to customer churn analysis in the e-commerce domain, emphasizing the importance of personalized strategies and understanding network dynamics. The overarching goal is to foster sustained business growth through improved customer satisfaction, personalized marketing strategies, and strategic resource allocation, ultimately reducing churn and ensuring long-term business success.

III. **Data Description**

Data Source:

The data has been manually created for the purpose of e-commerce customer churn analysis.

Variables:

1. CustomerId: A unique identifier for each customer (numerical).
2. Name: The name of the customer (categorical).
3. Age: The age of the customer (numerical).
4. Gender: The gender of the customer (categorical).
5. State: The state where the customer resides (categorical).
6. Brand: The brand of products purchased by the customer (categorical).
7. TotalPurchase: The total amount of purchases made by the customer (numerical).
8. Churn: A binary variable indicating whether the customer has churned (left the platform) or not (0 for no churn, 1 for churn) (categorical).

Sample Size and Number of Variables:

The dataset contains 620 observations (customers) and 8 variables.

Sample Observations:

CustomerId	Name	Age	Gender	State	Brand	TotalPurchase	Churn
101	Liam Taylor	29	Male	California	Nike	190.75	0
102	Logan Robinson	34	Female	New York	Adidas	480.3	0
103	Amy Wilson	32	Female	Florida	Samsung	700.8	0
104	Joseph Davis	28	Male	California	Nike	180.25	0
105	Sophia Wilson	41	Female	Florida	Samsung	820.75	0
106	Liam Wilson	26	Male	Texas	Apple	1000.5	0
107	Noah Martin	37	Female	New York	Adidas	850.3	1
108	Noah Martin	37	Female	New York	Adidas	710.25	1

Interest in the Data:

This dataset is of interest as it provides information about customers' demographics, purchasing behavior, and churn status in an e-commerce platform. Such data can be used to analyze customer behavior, identify patterns, and develop strategies to minimize customer churn, which is crucial for the growth and success of an e-commerce business.

Since the data has been manually created, there is no specific website or reference to provide. Features like age, gender, total purchase gives an idea about the contribution of these factors in affecting the customer churn. We can find out which factors plays major role in the churning of customers.

III. Research Questions

1. **What are the key factors influencing customer churn in the e-commerce platform, and how can they be identified and understood?**

- To gain insights into customer behaviors, preferences, and interaction patterns by leveraging various data analysis techniques.
- Contributes to a deeper understanding of the factors influencing churn and aids in identifying key features that impact churn prediction.

2. **How does age impact customer purchasing behavior and retention in the e-commerce domain, and what insights can be derived from the introduction of the "PurchaseToAgeRatio" feature?**

- To investigate the relationship between customer age, purchasing behavior, and retention in the e-commerce domain.
- To utilize the "PurchaseToAgeRatio" feature as a key metric to gain nuanced insights into how age influences purchasing behavior. Explores whether certain age groups exhibit distinct patterns in terms of transactional activity and long-term engagement with the platform.

3. **To what extent does gender play a role in customer churn, and how can categorical variables like gender be leveraged to uncover valuable insights into customer behavior and retention?**

- To assess the impact of gender on customer churn within the e-commerce platform.
- To explore the utilization of categorical variables, particularly gender, as a key factor in uncovering valuable insights into customer behavior and retention. c. Investigate whether specific gender groups exhibit distinctive patterns in terms of churn behavior and engagement with the platform.

4. **What is the significance of implementing personalized strategies and customer segmentation in reducing churn, and how can feature engineering for categorical variables contribute to this understanding?**

- To investigate how personalized strategies and segmentation contribute to improved customer retention, satisfaction, and overall business performance.
- To assess the impact of feature engineering on the creation of meaningful segments, emphasizing the role of categorical variables in tailoring strategies to specific customer needs.
- To explore the role of feature engineering for categorical variables in enhancing the understanding of customer behavior and preferences.

5. **What analytical methods, including GraphFrames for network analysis and KMeans clustering, can be employed to effectively identify influential customers whose churn may have cascading effects on the broader customer network?**

- To assess the effectiveness of these analytical methods in pinpointing influential customers whose churn could potentially impact the broader customer network.
- To explore the application of GraphFrames for network analysis as a method to identify influential customers within the e-commerce platform.
- Investigate the use of KMeans clustering to segment customers based on demographics and purchasing behavior.

IV. Methodology

1. **Data Preprocessing and Feature Engineering**:

- Missing value imputation: The Age and TotalPurchase columns had missing values, which were imputed using the median and mean values, respectively. Imputation is necessary to handle missing data and ensure that all observations can be included in the analysis.
- Feature creation: A new feature 'PurchaseToAgeRatio' was created by dividing the TotalPurchase by Age. This feature can provide insights into the spending patterns of customers across different age groups.
- Categorical variable encoding: Categorical variables like Gender, State, and Brand were encoded using techniques like StringIndexer and OneHotEncoder. This step is crucial for machine learning algorithms to understand and process categorical data effectively.

2. **Supervised Machine Learning:**

- Logistic Regression, Decision Tree, and Random Forest classifiers were used for predicting customer churn. These algorithms learn from the provided features (e.g., Age, Gender, Brand, TotalPurchase) to identify patterns and make predictions about whether a customer is likely to churn or not.
- The data was split into training and testing sets, and the models were trained on the training data.
- The performance of the models was evaluated using metrics like accuracy and the BinaryClassificationEvaluator.

3. **Graph Analysis:**

- The customer data was represented as a graph, where nodes represent customers, and edges represent relationships or interactions between them.
- The PageRank algorithm was applied to identify the most influential customers within the network. Customers with higher PageRank scores are considered more influential, as they may have more connections or are connected to other influential customers.
- This analysis can help the e-commerce platform identify key customers to focus their retention efforts on, as losing these influential customers could potentially lead to increased churn among other customers in their network.

4. **Unsupervised Machine Learning (Clustering):**

- K-Means clustering was applied to segment customers into distinct groups based on their features (e.g., Age, Gender, TotalPurchase).
- The clustering results were visualized using a scatter plot, which can help identify and understand the characteristics of different customer segments.
- Customer segmentation allows the e-commerce platform to tailor their marketing strategies, product offerings, and retention efforts to cater to the specific needs and preferences of each segment, leading to improved customer satisfaction and loyalty.

These techniques were chosen because they address different aspects of customer churn analysis and provide valuable insights into customer behavior, patterns, and potential strategies for customer retention. The combination of supervised and unsupervised machine learning techniques, along with graph analysis, offers a comprehensive approach to understanding and mitigating customer churn in the e-commerce domain.

V. Results and Finding

1. Key factors influencing customer churn:

- Through feature engineering, supervised learning models, and unsupervised clustering techniques, several factors were identified as influential in predicting customer churn, such as age, gender, brand preference, and total purchase amount.
- The PageRank algorithm highlighted influential customers within the network, whose churn could potentially lead to increased churn among their connected customers.

id	Name	Age	Gender	State	Brand	TotalPurchase	Churn	pagerank
125	Benjamin White	24	Male	Texas	Apple	1050.0	1	3.3846420462244735
109	Ava Adams	36	Female	Florida	Samsung	860.75	0	2.6060987613269737
114	Emma Harris	25	Male	Texas	Apple	1000.5	0	2.456269355951818
130	Liam Davis	36	Female	Florida	Samsung	860.75	1	1.8836090922795936
126	Sophia Turner	22	Male	Texas	Apple	950.0	0	1.825724705925734
117	Aria Clark	32	Female	New York	Adidas	380.3	0	1.7880610988517378
146	Mia Davis	26	Male	Texas	Apple	1100.0	0	1.6794861570185222
131	Owen Adams	34	Male	California	Nike	200.5	1	1.622573184275564
116	Mia Taylor	27	Male	California	Nike	190.75	0	1.5556362891068178
127	Liam Davis	24	Male	Texas	Apple	1050.0	0	1.5132356007415657
120	Liam Taylor	29	Male	California	Nike	190.75	0	1.4060088148160572
136	Mason Martin	28	Male	California	Nike	180.25	1	1.371931014633654
128	Mason Clark	37	Female	New York	Adidas	490.8	1	1.2266535475479785
102	Logan Robinson	34	Female	New York	Adidas	480.3	0	1.1403360494010872
115	Noah Adams	30	Female	Florida	Samsung	870.3	0	1.0656238087850602
144	Henry Davis	26	Male	Texas	Apple	1100.0	0	1.047633829189543
103	Amy Wilson	32	Female	Florida	Samsung	700.8	0	1.0236970299028605
124	Robert White	28	Male	California	Nike	180.5	1	1.0031493124457456
119	Liam Davis	25	Male	Texas	Apple	1050.0	0	0.9814021482228067
112	Mason Harris	34	Female	Florida	Samsung	870.3	1	0.9722622811270385

2. Impact of age on customer purchasing behavior and retention:

- The introduction of the "PurchaseToAgeRatio" feature provided insights into the spending patterns across different age groups.
- The clustering analysis revealed distinct customer segments based on age and total purchase, suggesting that certain age groups exhibit unique purchasing behaviors and engagement levels with the platform.

3. Role of gender in customer churn:

- The supervised learning models, which included gender as a feature, demonstrated the significance of gender in predicting customer churn.
- The SQL query grouping churned customers by brand and gender revealed patterns in churn rates across different gender and brand combinations, highlighting potential challenges in retaining specific segments.

Brand	Gender	CustomerCount
Adidas	Female	41
Adidas	Male	17
Apple	Female	13
Apple	Male	43
Nike	Female	11
Nike	Male	36
Samsung	Female	41
Samsung	Male	16

4. **Significance of personalized strategies and customer segmentation:**

- The K-Means clustering analysis effectively segmented customers based on features like age, gender, and total purchase, enabling the e-commerce platform to tailor strategies and offerings to specific customer segments.
- Feature engineering for categorical variables, such as gender and brand, played a crucial role in creating meaningful customer segments and understanding their unique behaviors and preferences.

5. **Identifying influential customers and cascading effects:**

- The GraphFrames implementation, utilizing the PageRank algorithm, successfully identified influential customers within the network based on their connections and interactions.
- The analysis revealed that retaining these influential customers could potentially prevent other customers from churning, reducing overall churn rates.
- The K-Means clustering analysis further supported the identification of distinct customer segments, which could be leveraged to understand the potential cascading effects of churn within specific segments.

These results and findings provide valuable insights into customer behavior, preferences, and interaction patterns within the e-commerce platform. By identifying key factors influencing churn, understanding the impact of age and gender, and recognizing the significance of personalized strategies and influential customers, the platform can develop targeted retention strategies, optimize resource allocation, and enhance overall customer satisfaction and loyalty.

VI. **Conclusion**

In conclusion, our comprehensive e-commerce analysis project utilizes advanced analytical techniques to address the critical issue of customer churn. Employing PySpark, SQL queries, and cutting-edge machine learning and graph analysis, we gained profound insights into customer behaviors, preferences, and interactions.

The project highlights the pivotal role of age in shaping purchasing behavior and retention, the influence of gender on churn dynamics, and the strategic significance of personalized strategies and customer segmentation. The introduction of the "PurchaseToAgeRatio" feature added a valuable dimension to our dataset, offering nuanced insights into customer engagement.

The innovative application of GraphFrames for network analysis and KMeans clustering provides a strategic advantage by identifying influential customers, crucial for predicting potential cascading effects of their churn. These findings empower businesses to formulate targeted retention strategies, optimize resource allocation, and drive sustained growth.

In summary, our project not only identifies churn-prone customer attributes but also equips e-commerce platform with actionable intelligence. The overarching objective is to elevate customer satisfaction, tailor marketing strategies, and strategically allocate resources, ultimately reducing churn and ensuring long-term success in the dynamic e-commerce landscape.