International Research Journal of Management and Commerce
ISSN: (2348-9766)
Impact Factor 7.098 Volume 12, Issue 02, Feb 2025
©Association of Academic Researchers and Faculties (AARF)
www.aarf.asia, **Email** : editoraarf@gmail.com

IDENTIFYING CUSTOMER/VISITOR CHURN THROUGH PREDICTIVE ANALYTICS

DR. NITVE DNYANDEV LAXMAN
HOD, Department of Commerce
PVG's College of Science and Commerce
Shivdarshan, Parvati Pune

DR. S. M. GAIKWAD
Principal- Pvg's College of
Science and Commerce
Shivdarshan, Parvati Pune

ABSTRACT:

In the fiercely competitive landscape of E-commerce, retaining customers is paramount for sustainable growth and success. This research focuses on utilizing predictive analytics techniques to identify and understand customer churn patterns in an E-commerce retail platform. By analyzing transactional and interaction data, predictive models are developed to forecast customers who are likely to disengage from the platform in the subsequent month. The insights derived from this analysis empower the company to proactively address churn risks, enhance customer experience, and implement targeted retention strategies. Through the application of advanced machine learning algorithms and data-driven decision-making, this project aims to provide actionable insights for reducing churn rates and fostering long-term customer relationships in the dynamic E-commerce ecosystem.

OBJECTIVES of Study

1) Data Gathering:

Data gathering involves collecting relevant information from various sources, such as databases, APIs, surveys, or web scraping, to create datasets for analysis. This process includes identifying sources, extracting or collecting data, and organizing it for further analysis. The quality and comprehensiveness of the gathered data are essential for accurate insights and decision-making.

2) Data Preparation:

Data preparation is the process of cleaning, transforming, and structuring raw data to make it suitable for analysis. It involves handling missing values, removing duplicates, standardizing formats, and encoding categorical variables. Effective data preparation ensures that the dataset is accurate, complete, and formatted correctly, enabling efficient analysis and modelling.

3) **Data Exploration:**

Data exploration involves analysing and visualizing datasets to gain insights and understand underlying patterns. It includes summarizing key statistics, identifying trends, and detecting outliers or anomalies. Techniques such as histograms, scatter plots, and correlation matrices are commonly used to explore relationships between variables. Data exploration aids in hypothesis generation, feature selection, and understanding the data's characteristics before proceeding to more advanced analyses or modelling.

1. **Data Wrangling:**

Data wrangling, also known as data munging, is the process of cleaning, structuring, and transforming raw data into a format suitable for analysis. It involves tasks such as handling

International Research Journal of Management and Commerce
ISSN: (2348-9766)
Impact Factor 7.098 Volume 12, Issue 02, Feb 2025
©Association of Academic Researchers and Faculties (AARF)
www.aarf.asia, **Email** : editoraarf@gmail.com

missing values, removing duplicates, standardizing formats, and dealing with inconsistencies in the data. Additionally, data wrangling often includes merging or joining multiple datasets and creating new variables or features to enhance analysis. Data wrangling is a crucial step in the data preparation process, ensuring that the data is accurate, complete, and properly formatted for analysis or modelling purposes.

2. Implementing Machine Learning Algorithms:

In implementing machine learning techniques, two fundamental methods are utilized: classification and clustering.

a) Classification: Classification involves categorizing data into predefined classes or labels based on features. The Decision Tree classifier is a popular algorithm for classification tasks, where it recursively splits the data into branches based on feature values, ultimately leading to leaf nodes representing class labels. This method is particularly useful in scenarios where the relationship between features and classes can be represented as a hierarchical structure, allowing for interpretable and intuitive decision-making.

b) Clustering, on the other hand, group's similar data points into clusters without predefined class labels. K Means is a widely used clustering algorithm that partitions data points into 'k' clusters based on similarity, aiming to minimize intra-cluster variance. It iteratively assigns data points to the nearest centroid and updates the centroids until convergence. K Means is suitable for exploratory data analysis, customer segmentation, and anomaly detection, providing insights into the inherent structure of the data.

SCOPE OF STUDY:

The scope of this study encompasses several key areas within the domain of predictive analytics and customer churn prediction for the E-Commerce retail platform. Specifically, the study will focus on:

1) **Data Gathering:** This phase involves downloading the dataset from Kaggle and converting it into a data frame. Additionally, it includes confirming the data conversion accuracy by displaying initial and final records, as well as summarizing the dataset's statistical information.

2) Data Preparation: Here, the dataset will be inspected for duplicate entries, missing data, and inconsistencies. Techniques will be applied to address these issues, such as removing duplicates, imputing or removing missing data, and resolving inconsistencies.

3) Data Exploration: Various plots, including scatter plots and exploratory data analysis (EDA) visuals, will be generated to gain insights into the relationships between different attributes and the target variable.

4) Data Wrangling: Univariate filters will be applied to the data, and observations will be reported based on the findings.

5) Implementing Machine Learning Techniques: Two machine learning techniques, namely Decision Tree classification and K Means clustering, will be implemented to predict customer churn and segment the data, respectively.

6) Conclusion: The study will conclude by summarizing the findings and providing insights into customer churn prediction and potential strategies for reducing churn rate and enhancing customer experience.

International Research Journal of Management and Commerce
ISSN: (2348-9766)
Impact Factor 7.098 Volume 12, Issue 02, Feb 2025
©Association of Academic Researchers and Faculties (AARF)
www.aarf.asia, **Email** : editoraarf@gmail.com

RESEARCH METHODOLOGY:

Sample Size:

The sample size for the classification & clustering models is 49,358 records and consisting of 49 features.

SAMPLING METHODS:

Classification models without oversampling techniques & model performance metrics

1. Decision Tree
2. K-Means

Decision Tree

A Decision Tree is a machine learning algorithm that models decisions through a tree-like structure. It recursively splits data based on features, with each internal node representing a decision point and each leaf node providing a final outcome or prediction. The algorithm selects splits by optimizing criteria like Gini impurity for classification or mean squared error for regression. Decision Trees are interpretable, handle both numerical and categorical data, and implicitly perform feature selection. However, they can be prone to overfitting and may not capture complex relationships. Pruning and ensemble methods like Random Forests mitigate these issues.

K-Means

K-Means is a widely applied clustering algorithm renowned for its simplicity and effectiveness in grouping data points into clusters. As an unsupervised learning technique, K-Means doesn't require labeled data for training, making it particularly useful in scenarios where the underlying patterns are not explicitly known. The algorithm operates by iteratively assigning each data point to the nearest cluster centroid based on a specified distance metric, typically the Euclidean distance. Subsequently, it recalculates the centroids by computing the mean of all data points assigned to each cluster. This process repeats until convergence, where centroids stabilize, or a predefined stopping criterion is met. Despite its simplicity, K-Means can effectively partition datasets into clusters, aiding in various tasks such as customer segmentation, image compression, and anomaly detection. However, it's worth noting that K-Means is sensitive to initial centroid placement and may converge to local optima, necessitating careful consideration of initialization techniques and multiple runs with different starting points to enhance robustness.

OBSERVATIONS:

Decision tree Classifier:

From the Decision Tree classifier performance results, we can deduce:

1) An accuracy of 0.784 indicates that around 78.4% of the predictions made by the model are correct.

2) A precision of 0.917 means that when the model predicts a customer as not churned, it is correct about 91.7% of the time.

3) A recall of 0.832 indicates that the model is able to correctly identify about 83.2% of the actual retained customers.

4) A high F1 score of 0.873 or 87.3% indicates a better balance between precision and

recall.
5) An AUC-ROC score of 0.577 indicates the model's ability to discriminate between churned and not churned customers.
6) The confusion matrix provides a breakdown of true positive, true negative, false positive, and false negative predictions.

K-Means Clustering:

From the Decision Tree classifier performance results, we can deduce:
1) The code is finding the optimal number of clusters (K) using the silhouette score. Silhouette score measures how close each sample in one cluster is to the samples in the neighboring clusters, and it provides an indication of the quality of the clustering. The K value having maximum silhouette score is selected.
2) A Silhouette Score of 0.6025 indicates that the clusters have a moderate degree of separation and cohesion. While this score suggests that there is some level of distinction between the clusters, it's not a very high score indicating that the clustering might not be as well-defined as desired.
3) We tried using higher K values for performing clustering, but the highest silhouette score was obtained by using K = 2.
4) We have a target class Defined for our data which can be used to compare the cluster labels to analyse the model health.

LIMITATIONS OF THE STUDY

1) Data Quality: The effectiveness of predictive analytics heavily relies on the quality and completeness of the dataset. Limitations in data quality, such as missing values, inaccuracies, or biases, may affect the accuracy and reliability of the predictive models and insights derived from the analysis.

2) Generalization: The findings and recommendations derived from this study may be specific to the dataset and conditions under which the data was collected. Generalizing the results to different contexts or time periods may not always be appropriate and could lead to misinterpretation or ineffective strategies.

3) Assumptions and Simplifications: Predictive analytics often involves making assumptions and simplifications about the underlying relationships and dynamics within the data. These assumptions may not always hold true in real-world scenarios, leading to potential inaccuracies or limitations in the predictive models.

4) Model Complexity: Decision Tree classification and K Means clustering are relatively simple and interpretable models. However, their simplicity may also limit their ability to capture complex patterns and relationships within the data. More sophisticated models or ensemble techniques may be required for capturing nuanced customer behaviour and improving prediction accuracy.

5) External Factors: The predictive models developed in this study may not account for external factors or variables that could influence customer churn, such as changes in market conditions, competitor strategies, or macroeconomic trends. Ignoring these external factors may limit the effectiveness of the predictive models in real-world applications.

International Research Journal of Management and Commerce
ISSN: (2348-9766)
Impact Factor 7.098 Volume 12, Issue 02, Feb 2025
©Association of Academic Researchers and Faculties (AARF)
www.aarf.asia, Email : editoraarf@gmail.com

6) Implementation Challenges: Implementing the recommendations derived from the study, such as enhancing customer experience or implementing targeted retention strategies, may pose practical challenges and require significant resources, coordination, and organizational buy-in. Overcoming these implementation challenges is essential for realizing the full benefits of the predictive analytics insights.

By acknowledging these limitations, the study aims to provide a realistic and transparent assessment of the developed classification models for predicting NPS status in the context of the specified scope.

CONCLUSION:

We have measured the performance of both Classification & Clustering algorithms respectively in our analysis. Based on the model performance analysis implemented in the steps above we can say for:

Decision Tree Classifier' model:

It has a relatively high accuracy, precision, recall and F1 score, indicating that it performs well in identifying churned customers. However, the AUC-ROC score suggests that there is room for model improvement. The model can be fine-tuned, or different features could be added to improve its discrimination capabilities.

K-Means Cluster' model:

Silhouette Score suggests that there is some level of separation between clusters, but it doesn't guarantee that the clustering is meaningful. The ARI and NMI scores are relatively low, indicating limited agreement with any ground truth labels if available. This suggests that the clustering may not align well with the actual classes. The FMI score is relatively high, indicating reasonable similarity between the true and predicted clusters.

Our dataset has a class label defined (target_class) and we are trying to predict if customers/visitors will visit the website during the subsequent month or not. This means that the customers who will not visit the website would churn out. The older accounts with short time interval between sessions tend to be retained.

Based on the performance metrics of both the ML models we can say that 'Decision Tree Classifier' model would work best to predict churn

REFERENCES:

Research Papers:

- https://www.researchgate.net/publication/342474107_Predicting_Customer_Churn_in_E-commerce_Industry_Using_Machine_Learning_Techniques
- https://www.researchgate.net/publication/240476385_Decision_Trees_An_Overview_and_Their_Use_in_Medicine
- https://www.researchgate.net/publication/334531488_A_Survey_on_K-Means_Algorithm_for_Big_Data_Analytics
- https://ieeexplore.ieee.org/document/8126741
- https://ieeexplore.ieee.org/document/8126741

Websites:

- https://www.optimove.com/resources/learning-center/understanding-customer-churn
- https://www.gooddata.com/blog/customer-churn-analysis
- https://towardsdatascience.com/understanding-gini-impurity-and-information-gain-in-decision-trees-ab4720518e1b
- https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Data Source:

- https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fwww.kaggle.com%2Fdatasets%2Ffridrichmrtn%2Fuser-churn-dataset%22