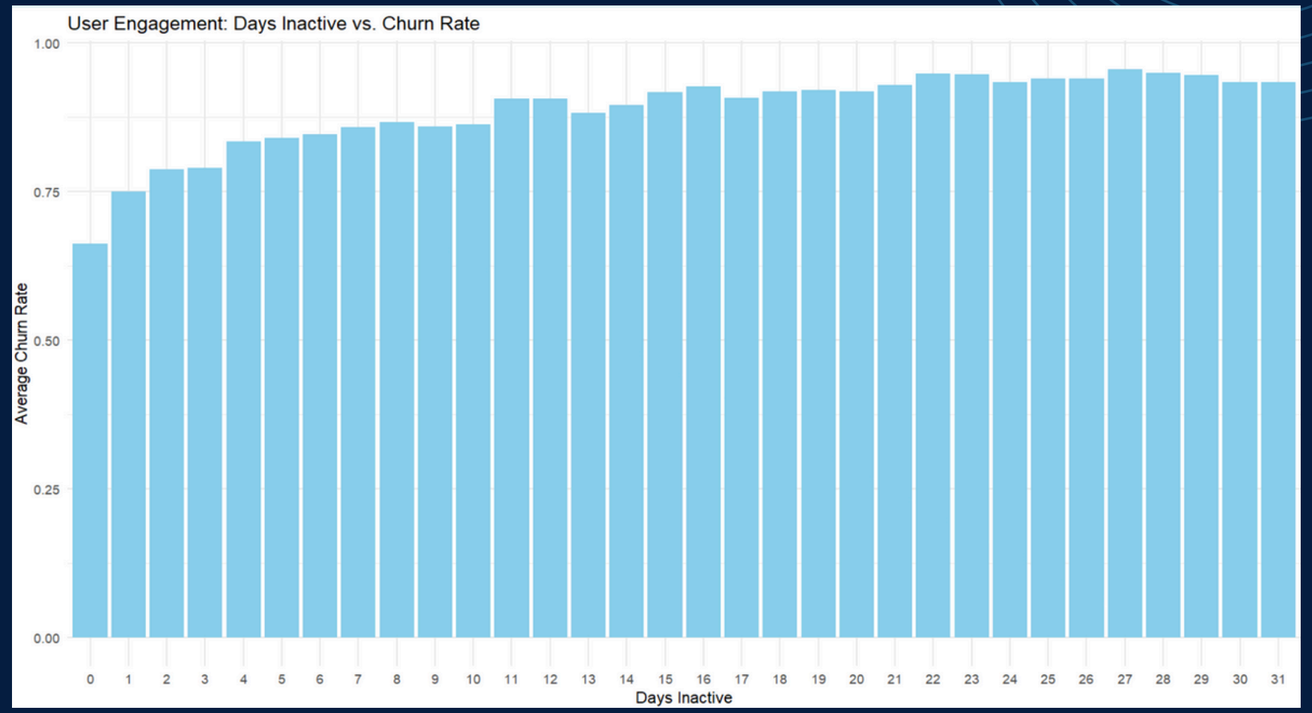
EXECUTIVE REPORT

Kaggle Competition
Churn prediction for online retail store

Submitted by Team Insight Seekers
Team members:-
Aayushi Marathe
Anvi Kothari
Muskan Bagde

Data Understanding

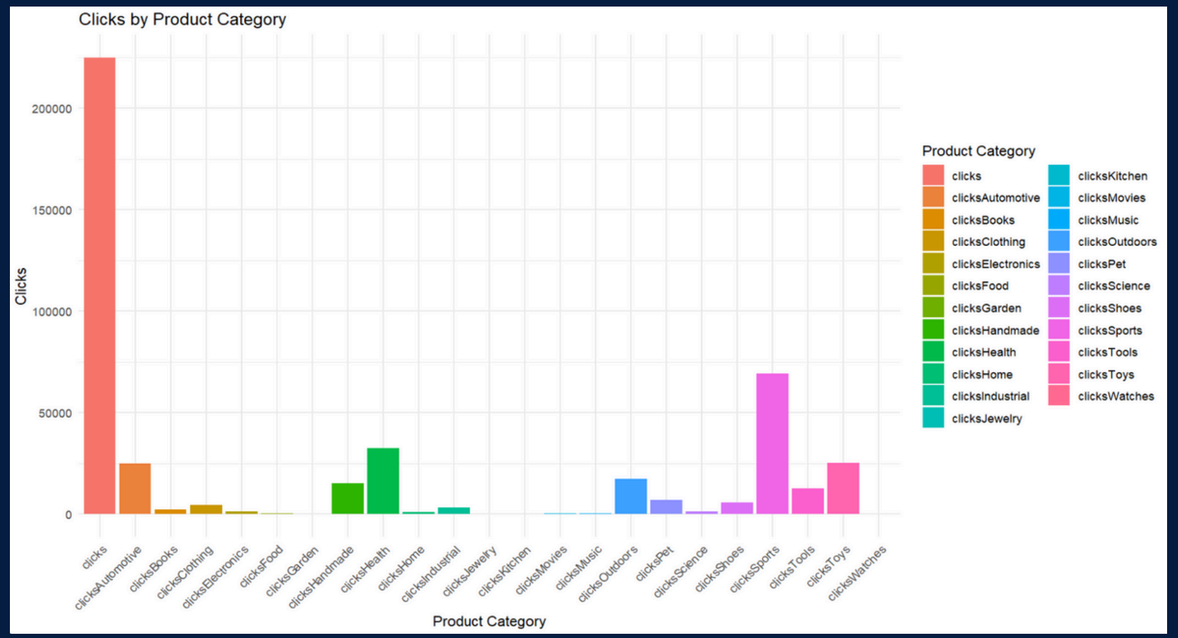


User Engagement:

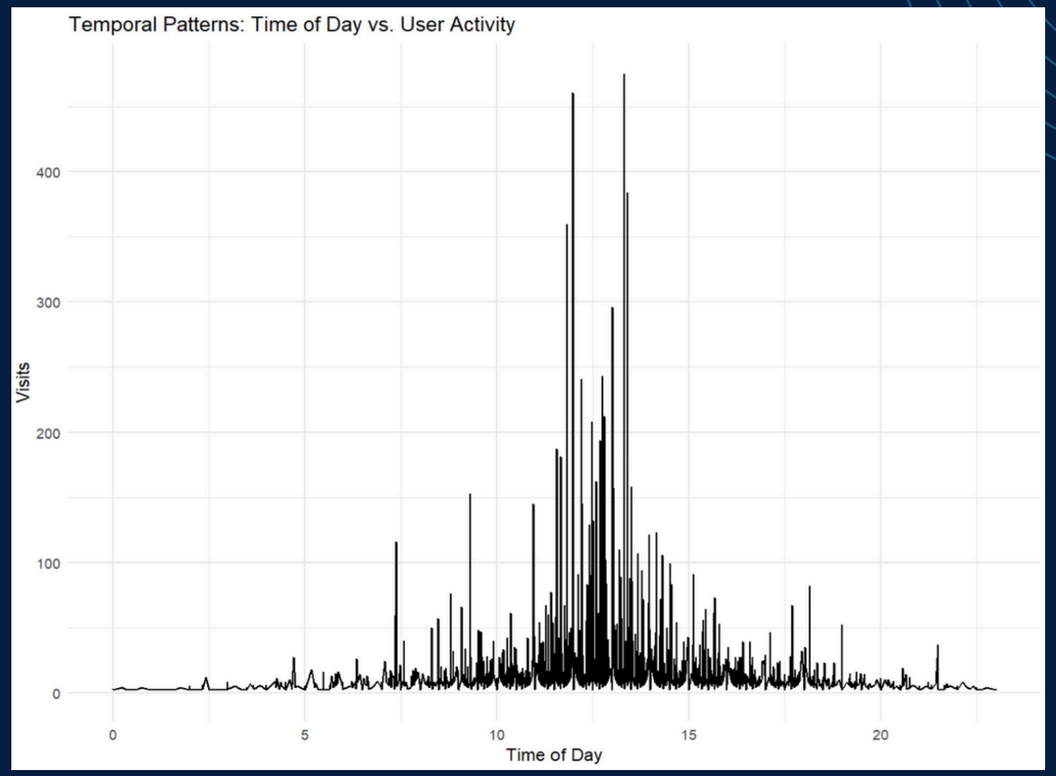
Variables like “daysInactive”, “visits”, and “clicks” are crucial in understanding user engagement. A preliminary examination suggests a potential correlation between daysInactive and churn, implying that more inactive days could lead to higher churn rates.

Product Interaction:

The dataset contains detailed interaction data across various product categories (“**clicksClothing**”,“**clicksShoes**”,“**clicksElectronics** etc.). These could be leveraged to identify popular categories and tailor recommendations or marketing strategies accordingly.



Data Understanding



Temporal Patterns:

With “timeOfDay”, ”weekdayPercent”, and similar variables, there's an opportunity to explore patterns in user activity—identifying peak times and days for engagement. This could optimize timing for promotional campaigns or product launches and help in reducing the overall churn.

Churn Rate:

The target variable churn has a high prevalence of 1 (churned users) at approximately 88.6% in the given train dataset. This indicates a significant churn problem, necessitating strategies to improve user retention.

Data Preparation

- **Initial Data Exploration:** Thoroughly examined dataset statistics and structure to understand its makeup and identify potential issues.
- **Data Cleaning:** Employed strategies to handle missing values and outliers, ensuring data integrity and quality.
- **Scaling and Standardization:** Utilized scaling techniques to bring data to a common scale and standardization for enhanced model interpretability.
- **Addressing Imbalanced Data:** Tackled class imbalance using oversampling methods, maintaining class proportions for unbiased model training and evaluation.

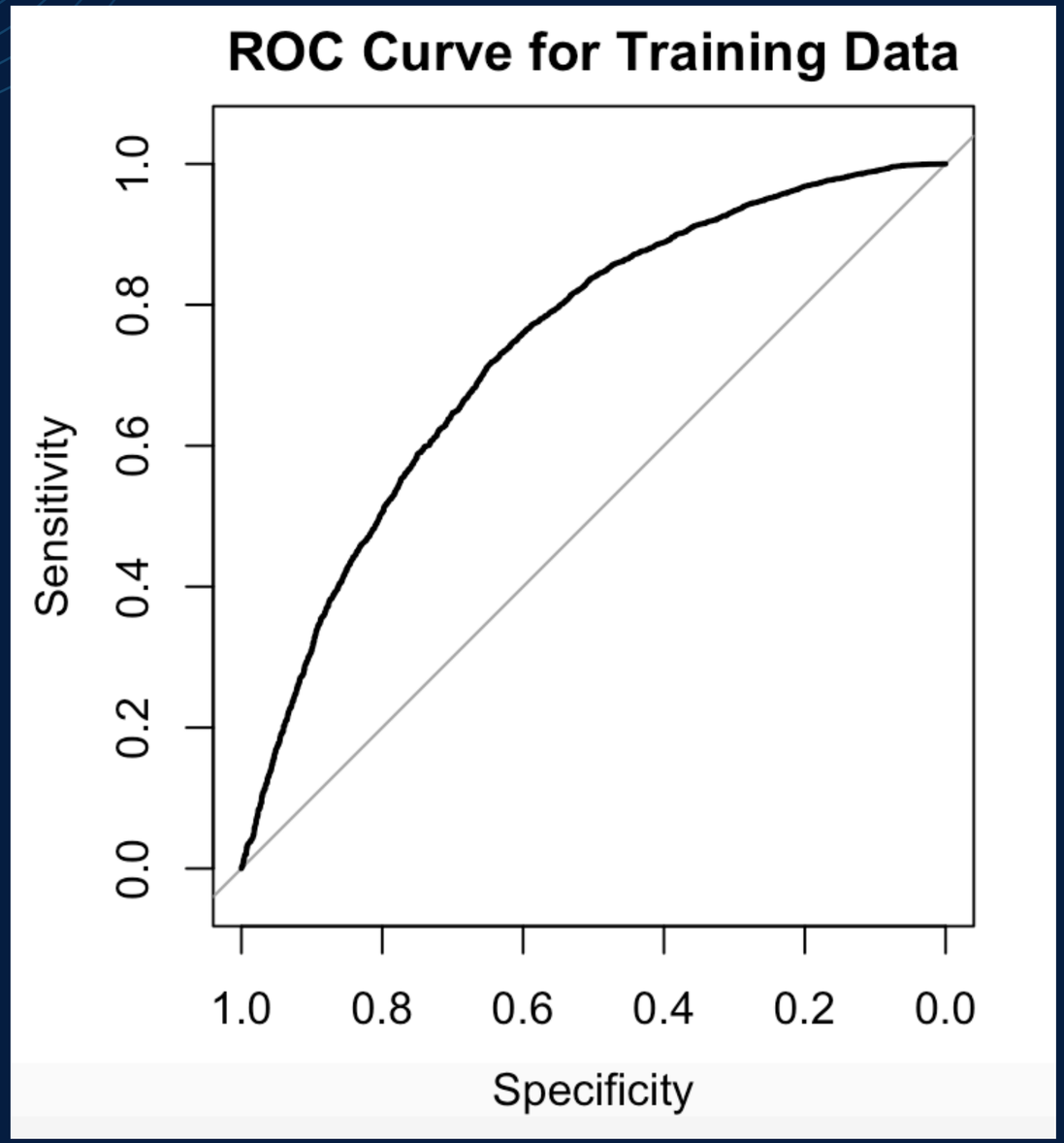
Data Modelling

- **Model Used:** XGBoost (eXtreme Gradient Boosting)
- **Algorithm Application:** Applied XGBoost for churn rate prediction in an online retail company.
- **Base Learners:** Utilized decision trees as base learners within the XGBoost framework.
- **Gradient Optimization:** Used gradient-based optimization to minimize the loss function, fine-tuning model parameters iteratively.
- **Regularization Techniques:** Incorporated L1 (Lasso) and L2 (Ridge) penalties, alongside the "gamma" parameter, to prevent overfitting and control model complexity.
- **Hyper parameter Tuning:** Tuned hyper parameters such as learning rate (eta), maximum tree depth (max_depth), subsampling ratio (subsample), and column subsampling ratio (colsample_bytree) to optimize model performance.
- **Scalability:** Designed the model to handle our large dataset and high-dimensional feature spaces efficiently, ensuring scalability and computational efficiency.
- **Outcome:** Successfully applied the model to predict churn probabilities, providing actionable insights for proactive customer retention strategies.

The chosen model delivered the most accurate predictions, supported by comprehensive exploratory data analysis and data preprocessing efforts. Logistic regression emerged as the second most effective model. Additionally, we explored various other models, including random forest, decision tree, and neural network, during our experimentation phase.

Evaluation Methodology

- **Selection of Key Metric**: Chose AUC-ROC as the main measure to assess model accuracy, particularly suited for imbalanced data like customer churn prediction.
- **Comparison with Various Benchmarks**: Compared our model to different methods, including basic rules and common algorithms, to gauge its superiority.
- **Ensuring Repeatable Results**: Documented all steps clearly to enable others to replicate our process accurately, promoting transparency and trustworthiness.



Managerial Implications

- **Targeted Audience for Marketing:** By understanding the characteristics of these segments of customers, targeted marketing strategies can be devised to retain them.By understanding the characteristics of these segments, targeted marketing strategies can be devised to retain them. For instance, customers who have been inactive for a certain period or exhibit erratic browsing behavior might require personalized offers or reminders to keep them engaged.

- **Optimizing Product Offerings:** Understanding customer behavior, like which categories they often check out or buy from, helps managers tweak their offerings. They can then customize promotions or even add new products that match what different groups of customers like.Managers can use this information to optimize product offerings, tailor promotions, or introduce new products that align with the interests of different customer segments. For instance, if many customers who might leave often look at electronics but don't buy, we could offer them special deals or package deals to encourage them to buy.

- **Enhancing Customer Experience:** Knowing when customers visit our website and how long they stay can help us make it better. For example, if lots of customers are online at specific times, we can make sure the website works well and have staff ready to help them. This could make the experience better and might stop some customers from leaving.

- **Predictive Customer Service:** Using predictive models in our customer service systems lets us spot customers who might leave before they do. This means our customer service team can reach out to them, solve any problems they have, and give them personalized help, which can make them more likely to stick with us.

- **Focus on Engagement and Inactivity Metrics:** Monitor key indicators such as frequency of visits and clicks, alongside periods of inactivity, to gauge customer satisfaction and predict churn.

Limitations

Limitations:

- **Data Quality and Completeness:** The accuracy of our predictions depends on how good our data is. If we're missing info or it's wrong, our predictions might be off. To make our predictions better, we could work on collecting more accurate data and adding in info from other places, like what customers say on social media.

- **External Factors:** Our dataset might miss external factors like economic changes, competitors' moves, or industry trends. To enhance predictions, we could incorporate outside data or adjust models to include these factors, improving accuracy.