EXECUTIVE SUMMARY

Presented by:
Hitankshi Jain
Ishita Poorey
Vaishali Patidar

DATA UNDERSTANDING

- **UNDERSTANDING THE DATA**

The data tracks user activity and shopping habits to predict churn, aiding in the development of customer retention strategies.

- **TARGET CLASS**

The target class distribution provides insights into customer churn rates. Understanding this distribution is vital for businesses to strategize customer retention.

- **BUSINESS CONTEXT**

This information can then be used to develop strategies to improve customer retention, such as personalized marketing or improved customer service. Understanding these patterns and acting upon them can lead to increased customer loyalty, improved customer satisfaction, and ultimately, increased revenue for the business.

DATA PREPARATION

1. Data Loading:

- The data was loaded from CSV files into R data frames using the read.csv function.

2. Preprocessing Steps:

- These steps are essential to ensure the quality and usability of the data.
- They include:

1. Handling Missing Values:
   - Machine learning models cannot handle missing values.
   - This could involve removing rows or columns with missing values, or imputing missing values based on other observations.
2. Encoding Categorical Variables:
   - Many machine learning models require categorical variables to be converted into a numerical format.
   - This could be achieved through various encoding techniques like one-hot encoding, ordinal encoding, etc.
3. Scaling Numerical Variables:
   - Some models are sensitive to the scale of the input features.
   - Therefore, numerical variables might need to be scaled (e.g., normalization or standardization) to ensure that they contribute equally to the model's performance.

These preprocessing steps ensure that the data is in a suitable format for the machine learning models and enhance the reproducibility of the analysis. It's important to note that the same preprocessing steps applied to the training data must also be applied to any new data (like the test set in this case) before making predictions. This ensures consistency and allows the models to make meaningful predictions on new data.

MODELING

- Two machine learning models were used: *Support Vector Machine (SVM)* and *Gradient Boosting Machine (GBM)*.
- These models were chosen for their ability to handle complex, non-linear relationships in the data and their robustness to overfitting.

1. Support Vector Machine (SVM):

- SVM is a powerful and flexible classification algorithm that can handle both linear and non-linear data. It works by finding the hyperplane that best separates the classes in the data.
- The SVM model was trained using a radial basis function (RBF) kernel, which allows the SVM to handle non-linear data by mapping it to a higher-dimensional space.

2. Gradient Boosting Machine (GBM):

- GBM is a type of ensemble learning method that builds multiple weak learners in a sequential manner.
- The GBM model was trained using a Bernoulli distribution, suitable for binary classification problems like customer churn.
- The number of trees (n.trees) was set to 1000, indicating the number of sequential trees to be modeled.

- After training, predictions were made on the test set using both the SVM and GBM models.
- These predictions were then combined by averaging to create a final prediction, a simple form of model blending that can often improve prediction accuracy. in this case, blending the SVM and GBM predictions resulted in an AUC of 0.74, which is a moderate performance.

EVALUATION METHOD

In this stage, the performance of the models was evaluated using a separate test set. This is a common practice in machine learning to ensure that the model's performance is assessed on unseen data, providing a realistic measure of its predictive power.

1. Predictions: After training the SVM and GBM models on the training data, they were used to make predictions on the test set. These predictions represent the models' classification of whether a customer will churn or not.

2. Model Blending: The predictions from the SVM and GBM models were then combined by averaging. This technique, known as model blending, is based on the idea that a combination of predictions from different models can provide a more robust and accurate prediction than any single model. This is because different models may capture different patterns in the data, and combining their predictions can help to balance out their individual weaknesses.

3. Assumptions: This approach assumes that the blend of predictions from the SVM and GBM models will provide a more accurate prediction of customer churn. However, it's important to note that this may not always be the case. The effectiveness of model blending can depend on the correlation between the models' errors. If the models make very similar errors, blending may not improve the predictions significantly.

MANAGERIAL IMPLICATIONS & LIMITATIONS

1. Managerial Implications:

- The SVM and GBM models can help the company identify customers who are likely to churn.
- This information allows the company to take proactive measures to retain these customers.
- The company could target identified customers with special offers or personalized services to increase their satisfaction and loyalty.

2. Data Limitations:

- The models' predictions are only as good as the data they are trained on.
- If the training data is not representative of the company's customer base, the predictions may not be accurate.
- This could lead to ineffective or even counterproductive retention strategies.
- It's crucial to ensure that the data used to train the models is comprehensive, up-to-date, and accurately reflects the characteristics of the customer base.

3. Future Work:

- There are several ways to potentially improve the performance of the models.
- This could involve tuning the models' parameters, incorporating more features that could be predictive of churn, or trying different modeling techniques.
- These efforts could lead to more accurate predictions and more effective customer retention strategies.

- In conclusion, while the models provide a valuable tool for predicting customer churn, it's important to be aware of their limitations and continuously strive for improvement. This will help ensure that the company's retention strategies are as effective as possible.