Presented by Martina Navarria

Churn Prediction

Using Logistic regression and Decision Tree Classifier

Business context

This project focuses on a subscription-based online service (e.g., streaming, software as a service). Churn prediction is critical for the business to retain customers, reduce acquisition costs, and maintain revenue. Identifying customers likely to churn allows the company to take proactive measures (e.g., discounts, personalized outreach) to prevent cancellations.

Topic analyzed

The churn prediction model seeks to predict which users are likely to discontinue their subscriptions. By analyzing factors such as subscription type, contract length, age, sex, usage frequency, and payment delays, the model provides insights into customer behavior.

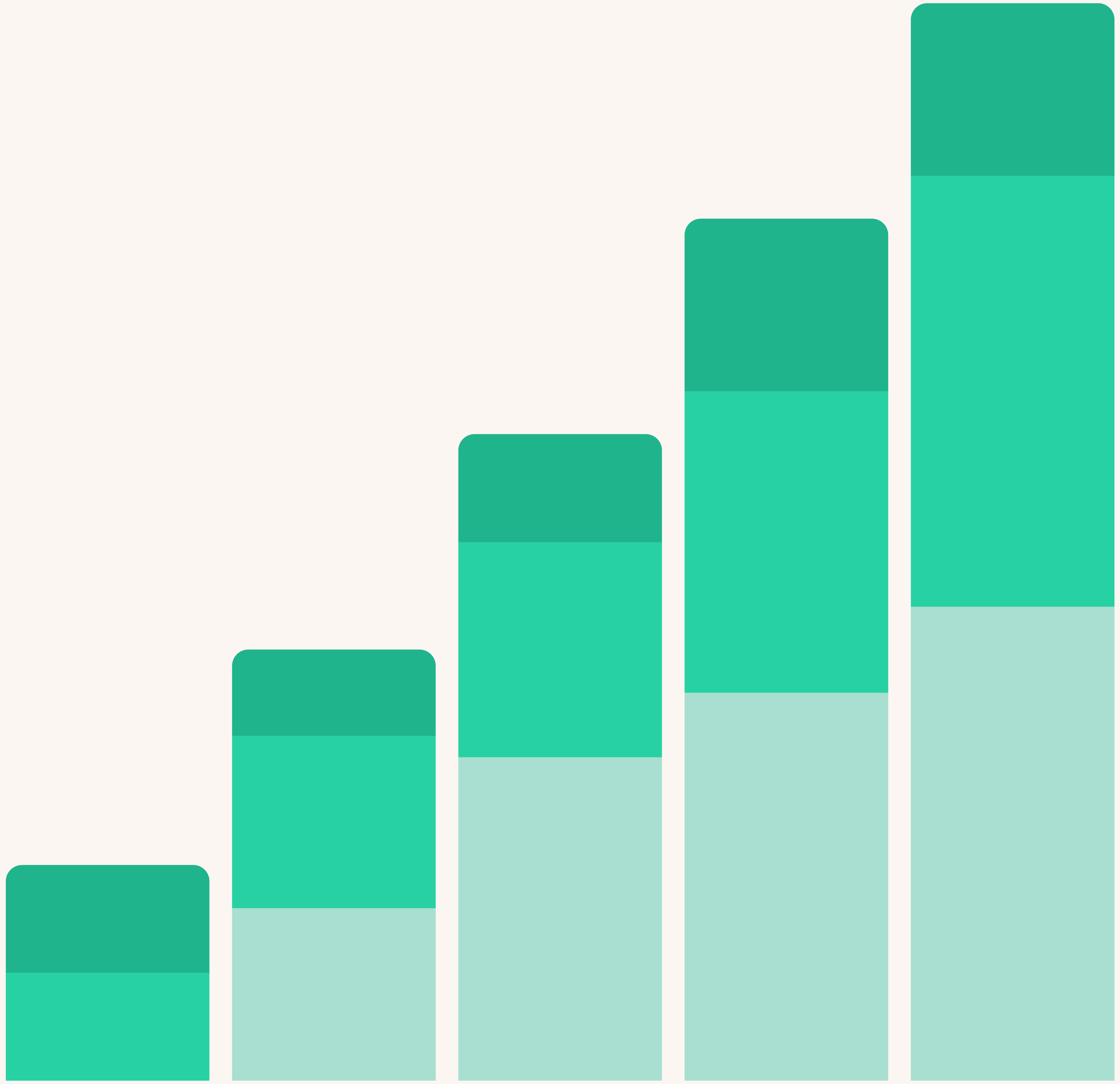
AI Model Support: The model enables the company to efficiently allocate resources to retention strategies by focusing on high-risk customers, leading to potential increases in customer lifetime value and reducing churn rates.

Dataset overview

The dataset contains customer-level data for a subscription-based service.

Main Variables:

- Subscription Type: Identifies the level or type of subscription (e.g., basic, premium).
- Contract Length: Duration of the subscription (e.g., month-to-month, annual).
- Age: Age of the customer.
- Sex: Gender of the customer.
- Usage Frequency: How often the customer uses the service.
- Payment Delay: Number of instances where the customer has delayed payments.



Methodology: Linear Regression

Linear regression was employed initially to explore the relationships between customer features and the likelihood of churn.

Model Development:

- Preprocessing: Missing value imputation, feature scaling, and encoding of categorical variables.
- Split data into training and testing sets for model validation.
- Performance Evaluation: Model accuracy, precision, recall, F1 score, ROC-AUC, and log-loss metrics were calculated.

Key Findings:

- Logistic Regression achieved good accuracy on both training (84.8%) and testing (84.9%) sets.
- Precision: 87.3% (test set), indicating that most predicted churners indeed churned.
- AUC: 0.910, reflecting good discriminative power.

Methodology: Decision Tree Classifier

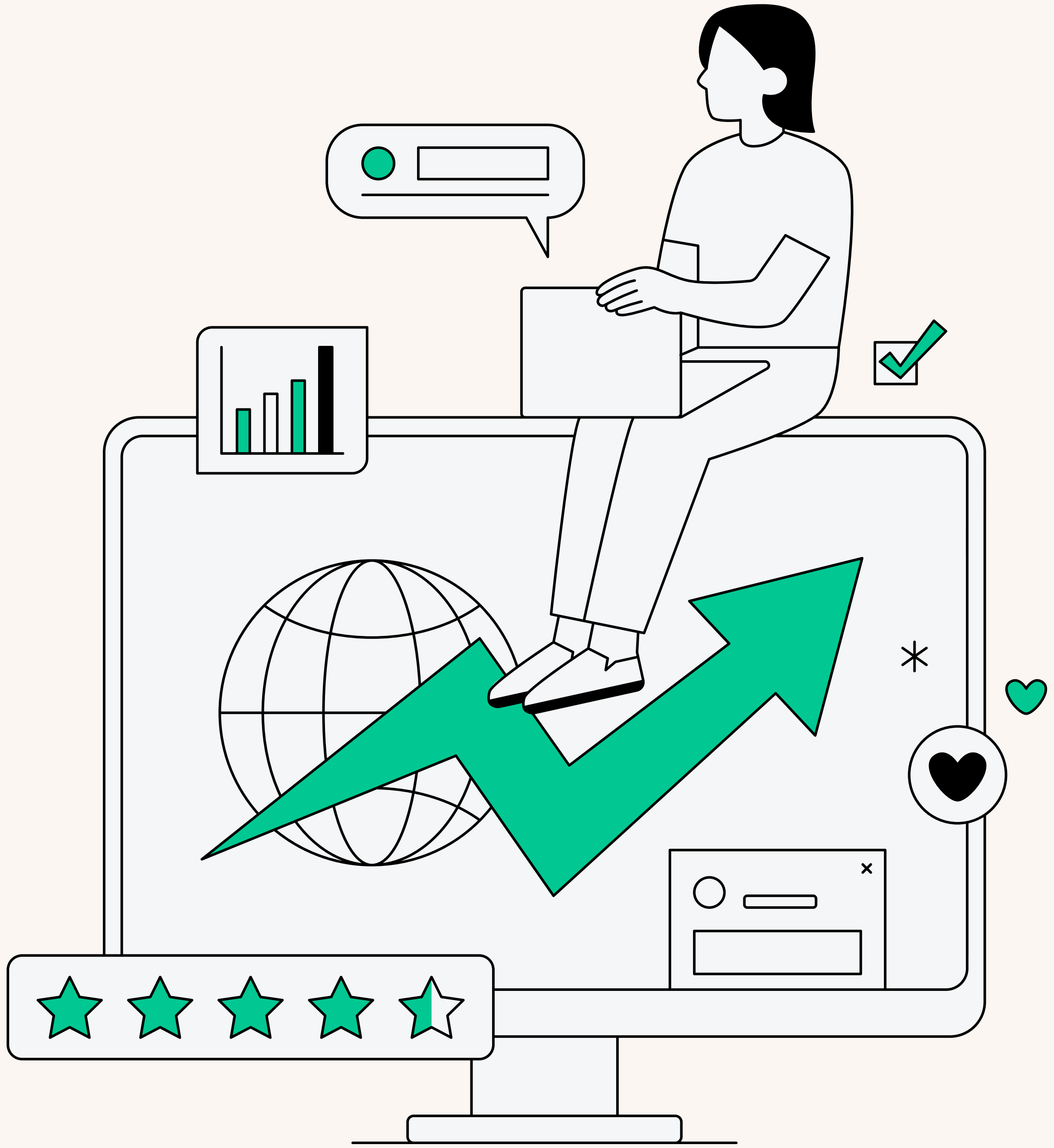
The Decision Tree Classifier was used for a more interpretable model to understand customer churn behavior.

Model Development:

- Hyperparameters like tree depth were optimized to prevent overfitting.
- Decision trees can handle non-linear relationships and complex interactions between variables.

Key Findings:

- High Accuracy: Training (93.5%) and testing (93.5%) accuracy indicates strong model performance.
- High Recall: 99.7% (test set), meaning the model effectively identifies almost all churners.
- ROC-AUC of 0.953 shows excellent classification performance.



Business Results & Implications

Business Insights:

Both models performed well, with the Decision Tree offering slightly better recall and interpretability, while the logistic regression model was more balanced in terms of overall accuracy and AUC.

Implications:

The high recall from the Decision Tree Classifier suggests that it is particularly suitable for situations where catching as many churners as possible is critical (e.g., reducing the churn rate aggressively).

Logistic Regression is more robust for precision, making it better for targeting interventions where resources are limited, and precision is necessary (e.g., personalized campaigns).

Next Steps:

The company can use these insights to target customers with specific retention strategies, such as offering discounts to those with longer payment delays or personalized promotions to lower usage frequency users.

Further model tuning or combining both approaches (ensemble methods) could enhance results.