TEAM A3

EXECUTIVE SUMMARY

PREDICTING CUSTOMER CHURN FOR AN ONLINE STORE

STAT 642 DATA MINING FINAL PROJECT

PRESENTED BY - TEAM A3

UNDERSTANDING THE CHALLENGE

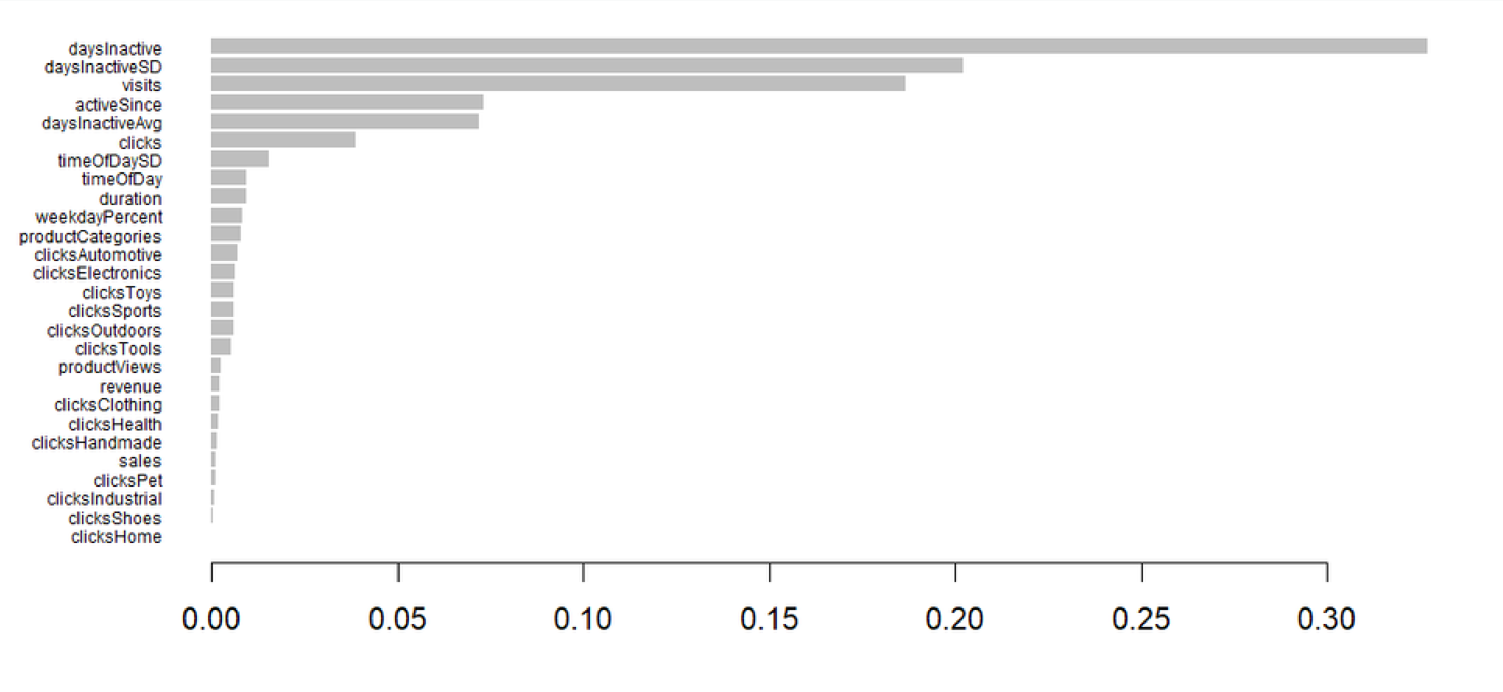
- **Problem at hand**: Significant churn rate of roughly 90%.

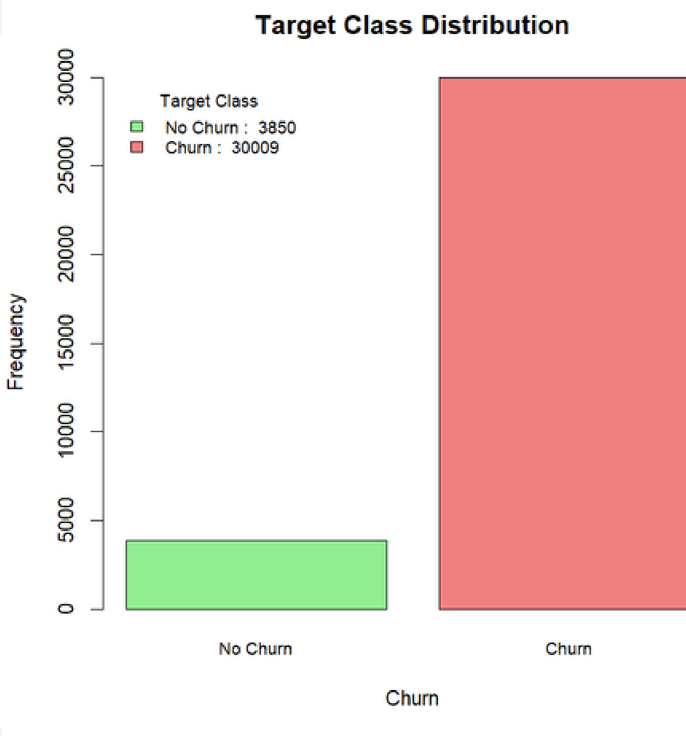
- **Impact**: Failing to retain customers can lead to decreased revenue, increased marketing costs and negative brand reputation

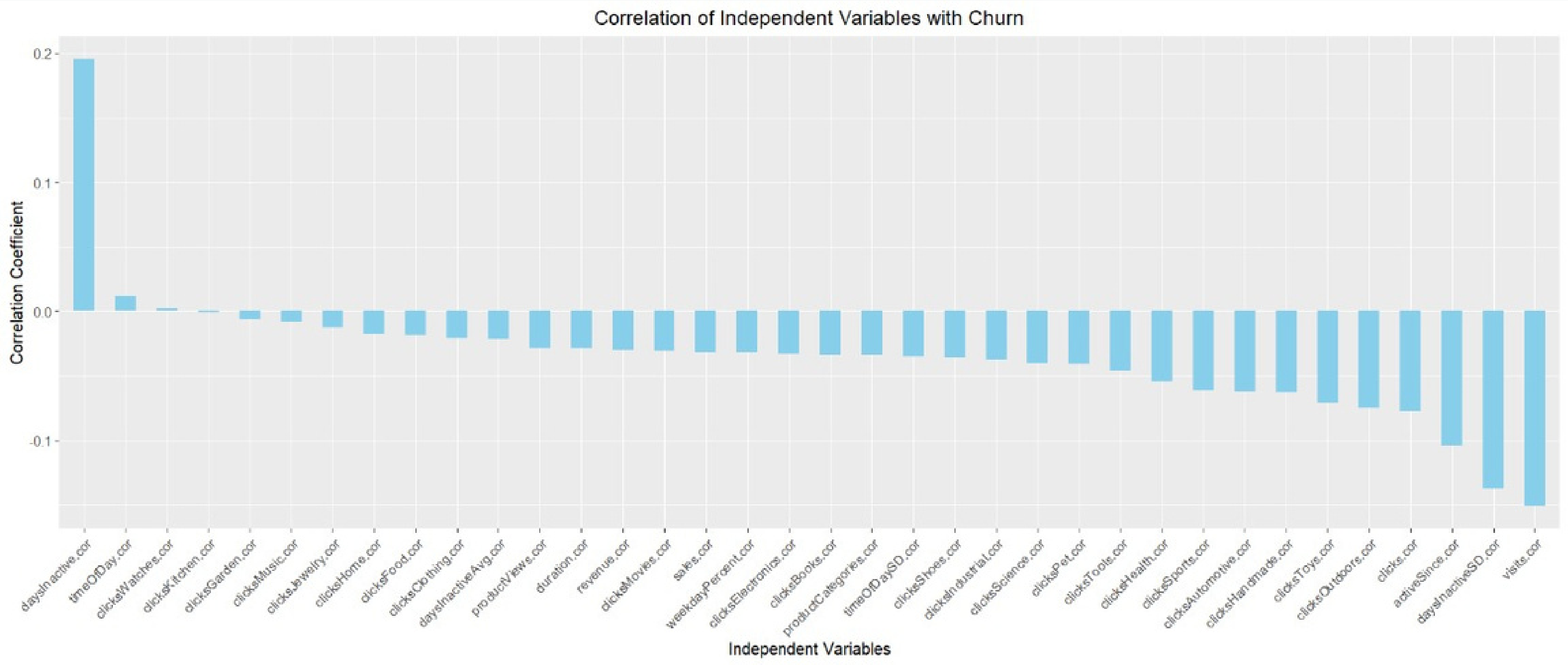
- **Solution**: To address the issue, we developed a robust machine learning model capable of predicting customers who are unlikely to return. The model serves as a tool for idnetifying at-risk customers and develop strategies to tackle the issue and spur sustainability and growth of the business

DATA UNDERSTANDING

- To analyze the significance of the variables in the dataset we plotted the correlation of each one to churn as well as the importance plot which is a feature of the XGBoost package that ranks features by their impact on model predictions
  - The top 5 variables by correlation are days inactive, visits, days inactive SD, active since and clicks.
  - the top 5 by importance are days inactive, visits, days inactive SD, active since and average days inactive
  - These findings highlight how instrumental active engagement is in preventing churn
- The target variable churn appears moderately imbalanced, with 3850 customers not churning over a total of 30009
  - steps were taken in the model development process to address the problem







DATA PREPARATION

- we looked for null values and found none, therefore no further processing was needed

- We opted for no under sampling/oversampling for handling target variable imbalance. However, we used stratified cross-validation in the hyperparameter tuning process, which ensures that each fold has the same proportion of the target class as the whole dataset.

- We tried to employ feature engineering, namely by creating new variables out of the already available ones and also by removing some.
  - For instance, we tried to remove the specific categories and only use the “clicks”, as it sums them up.
  - we decided not to pursue this approach as the accuracy of our model was negatively affected by it

- for reproducibility, we set a random seed.
  - This step is essential for reproducibility in processes that involve random number generation.
  - It ensures that everyone running the code will encounter the same random numbers sequence, which is essential for consistent outputs

```r
library(xgboost)

train_path <- "train.csv"
test_path <- "test.csv"
data_train <- read.csv(train_path, stringsAsFactors = FALSE)
data_test <- read.csv(test_path, stringsAsFactors = FALSE)

set.seed(42)
# Summary statistics
summary(data_train)
```

MODELING

XGBOOST FOR CUSTOMER CHURN PREDICTION

Model Used

- **XGBoost (Extreme Gradient Boosting)** Known for its efficiency and accuracy in classification problems.

Why XGBoost?

- Handles unbalanced data well, flexible with extensive customisation through hyperparameters, and excels in predicting categorical outcomes like customer churn.

Core Mechanism

- Utilizes a series of decision trees, each correcting its predecessor, to iteratively refine predictions. The gradient boosting technique minimizes prediction errors using a gradient descent algorithm, enhancing accuracy over iterations.

Algorithmic Efficiency

- XGBoost's speed and performance stem from its capability to parallelize the tree construction process and implement advanced regularization techniques, which prevent overfitting and improve model stability.

Interpretability

- Despite its complexity, XGBoost offers tools for understanding feature importance, allowing insights into which customer attributes significantly impact churn. This transparency is invaluable for crafting targeted retention strategies.

EVALUATION METHODOLOGY

```r
# Define XGBoost parameters
params <- list(
  objective = "binary:logistic",
  booster = "gbtree",
  eval_metric = "auc",
  eta = 0.01,
  max_depth = 4,
  subsample = 0.8,
  colsample_bytree = 0.8,
  min_child_weight = 5,
  lambda = 3,
  alpha = 2,
  colsample_bylevel = 0.8,
  colsample_bynode = 0.8
```

```r
# Cross-validation
cv <- xgb.cv(
  params = params,
  data = dtrain,
  nrounds = 100,
  nfold = 4,
  stratified = TRUE,
  print_every_n = 10,
  early_stopping_rounds = 10,
  maximize = TRUE
)
```

- **Hyperparameter Tuning**: A grid search approach to find the optimal model parameters, focusing on learning rate (eta) and tree depth (max_depth)

- **Cross-validation:** Employed to estimate the effectiveness of the model, ensuring robustness and avoiding overfitting.

```r
# Grid search for hyperparameter tuning
for (eta in c(0.05, 0.1)) {
  for (max_depth in c(4, 6)) {
    params$eta <- eta
    params$max_depth <- max_depth
```

```r
> # Report the best AUC and parameters
> print(paste("Best AUC:", best_auc))
[1] "Best AUC: 0.754089620276968"
```

Model Performance

- The model's performance was evaluated using the Area Under the ROC Curve (AUC), emphasizing its ability to differentiate between churned and retained customers.
- We employed a stratified k-fold cross-validation approach to enhance the reliability and generalizability of our findings.
- With the graphical representation of AUC progress we illustrate the models learning and influential predictors.
- The model with optimal parameters was then applied on full dataset and predicted churn probabilities for the test data.

MANAGERIAL IMPLICATIONS AND LIMITATIONS

Reasons for customer retention:

1. Engagement and experience - engagement metrics ssuch as days inactive, clicks, and visits are strong churn predictors
2. Product diversification - the diversity of product categories viewed and engaged with affect churn, showwing the important of offering a wide range of products
3. Timely and relevant communication - the variables related to frequency of customer interactions and timing provide insights into the best moments to engage with customers

Strategies for preventing churn

1. To increase engagement - personalize marketing communications and reccomendations based on customer data. Implement loyalty programs to incentivize repeat purchases
2. To enahnce user experience - optimize product offerings based on customer interest trends and online shopping experience by continuosly updating the platform to keep it user-friendly and functional
3. To enhance customer relationship - adopting a data driven approach to manage customer relationship. Utilize predictive analysis and ML to anticipate customer needs and engage with the right approach at the right time.

- For consistency - regularly monitor key performance metrics to track trends and intervene if necessary

Limitations:

1. Model interpretability - while our model shows strong performance in identifying churn risk, it is not as straightforward to pinpoint the specific factors causing customers to churn
2. Data preparation - we acknoledge that we could have employed better practices to handle outliers and target imbalance to provide a more robust and accurate model.
3. Overfitting - although the model does not overfit excessively, further reducing the probelm would result beneficial

THANK YOU