CUSTOMER CHURN PREDICTION

Prachi Mehta (202318008)
Dhruvi Mehta (202318003)
Simran Dalvi (202318042)

PROBLEM STATEMENT

- Acquiring new customer is more costly than retaining current customer
- Goal: Identify current customers who are likely to churn/cancel subscription

Remove null values from various columns

Extract features from user log data

User grid search cross validation for hyper-parameter tuning

Test the model using 20% of the dataset

DATA

- Sparkify is an imaginary digital music service similar to Spotify.
- The dataset contain 12GB of user interactions with this service.

```
root
 |-- artist: string (nullable = true)
 |-- auth: string (nullable = true)
 |-- firstName: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- itemInSession: long (nullable = true)
 |-- lastName: string (nullable = true)
 |-- length: double (nullable = true)
 |-- level: string (nullable = true)
 |-- location: string (nullable = true)
 |-- method: string (nullable = true)
 |-- page: string (nullable = true)
 |-- registration: long (nullable = true)
 |-- sessionId: long (nullable = true)
 |-- song: string (nullable = true)
 |-- status: long (nullable = true)
 |-- ts: long (nullable = true)
 |-- userAgent: string (nullable = true)
 |-- userId: string (nullable = true)
```

DATA PREPROCESSING

Data selection

Columns that were not significant to the modelling process will be dropped

- Firstname
- Lastname
- Id_copy

userID was retained as it was used for feature engineering step.

Unit conversion

Registration and TS were given in milliseconds.
These fields were converted to seconds by dividing the values by 1000.

Create churn label

Dataset only contains user log data used page column to identify churners:

- Visiting cancellation confirmation page indicated a churned user
- Creating a label column where 1 indicates a churned user and 0 indicated otherwise

FEATURE ENGINEERING

- Meaningful data has to be created from the user log data that could be used by the prediction models.
- The following features were used
  - Time since registration
  - Number of friends referred
  - Total songs listened to
  - Total songs liked
  - Total songs disliked
  - Number of songs in user playlist
  - Average songs played
  - Number of artists listened to
  - Number of user sessions logged
- More features were used initially but discarded after observing less than 1% feature importance during training of models.

MODELLING

- Dataset will be split into 80-20 train test split
- Grid search cross validation with three folds was used to built the following models
  - Gradient boosting trees
  - Random forests
  - Logistic regression
  - Support vector machine
  - Hybrid model
- Goal was to maximize F-1 score since the dataset is highly imbalanced

REFERENCES

- (PDF) Computational Efficiency Analysis of Customer Churn Prediction Using Spark and Caret Random Forest Classifier (researchgate.net)
- Customer churn prediction system: a machine learning approach | Computing (springer.com)

THANK YOU