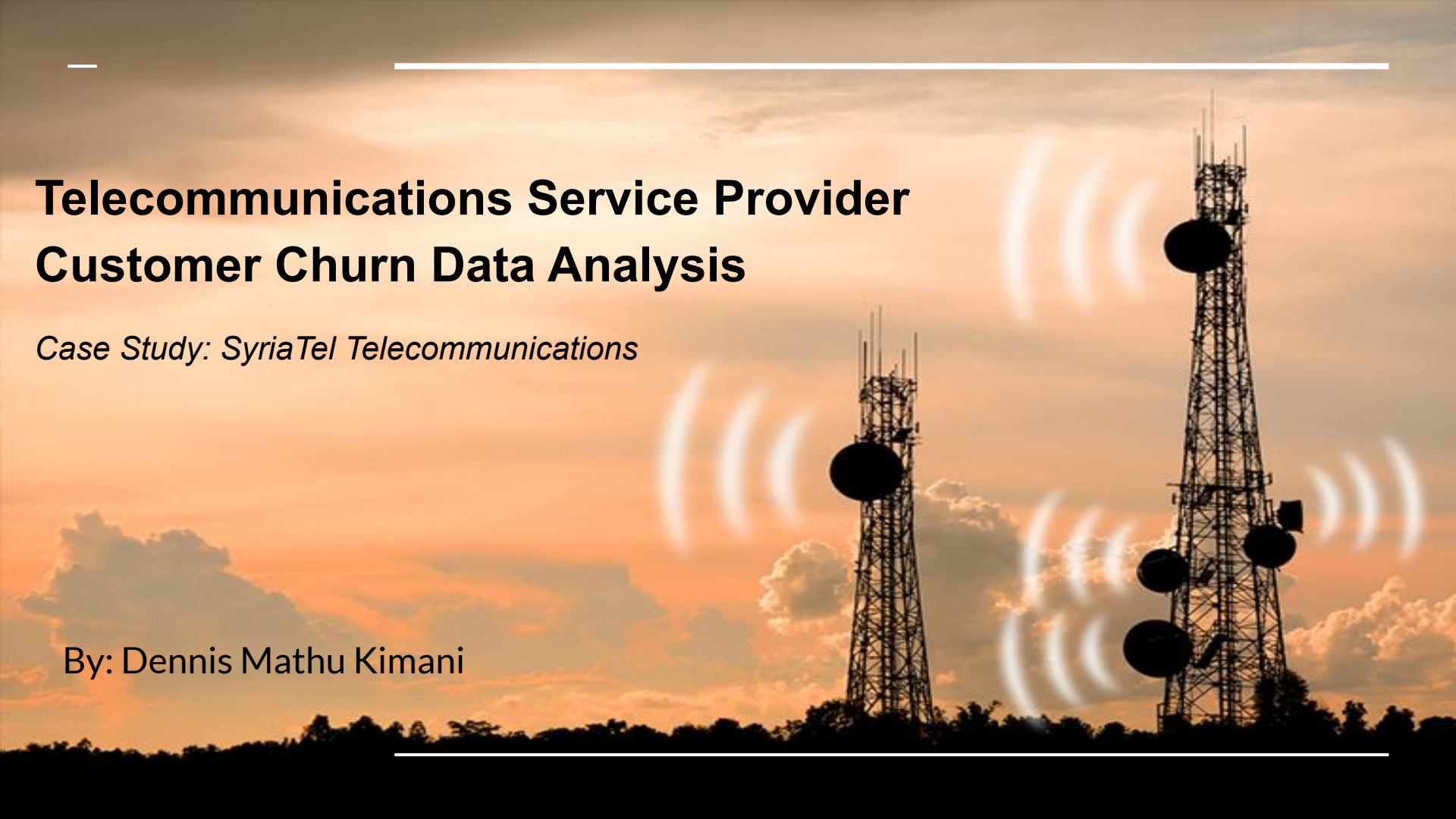
Telecommunications Service Provider Customer Churn Data Analysis

Case Study: SyriaTel Telecommunications



By: Dennis Mathu Kimani

Presentation Outline

- Overview
- Business Understanding
- Problem Statement & Objective
- Data Understanding
- Modeling
- Evaluation
- EDA
- Recommendations
- Next Steps

Overview

Customer churn definition: *rate of customer terminating a subscription or involvement to a business*

- A data analysis of customer churn in the telecommunications industry
- Goal is to come up with a model with high prediction of customers at the risk of churning
- Use model and data insights to make recommendations and next steps

Business Understanding

- SyriaTel, a mobile telecommunications offering calling both locally and internationally

Domain knowledge

- Average churn rate in telco industry is 20%
- 1% reduction in churn can lead to a 5% profit increase in the telecom industry

Source: subscriptionflow.com

Problem Statement and Objective

Problem Statement

The telecom industry is experiencing challenges in retaining customers, with an increasing churn rate affecting overall business sustainability.

Objective

To create a model that identifies customers likely to churn, providing insights to reduce churn

Data Understanding

Dataset has features related to SyriaTel Telecom customers

Key identifier: Phone Number

Target: Churn

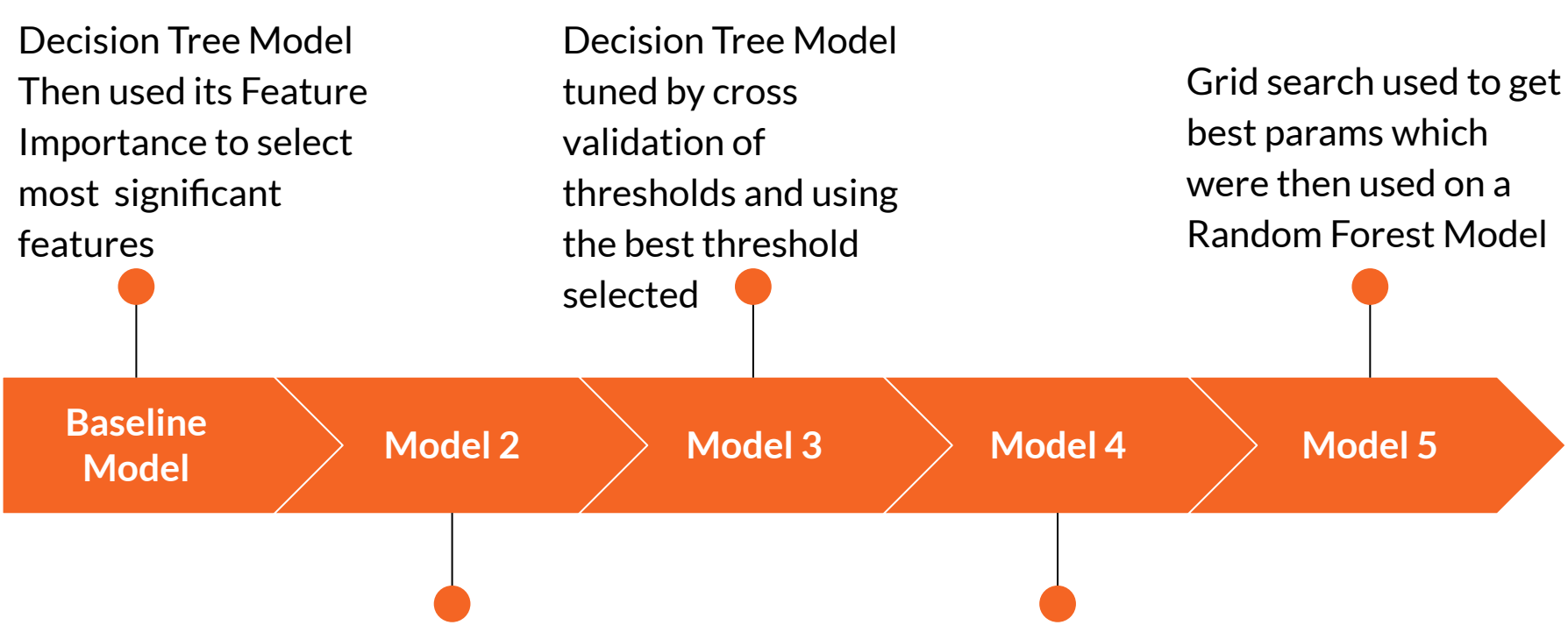
Feature engineering: Cumulative daily charges

Other features include:

Customer's region (area code and state), the length their account has been active, whether they are on international and voicemail plans and calls and charges at different periods of the day either locally or internationally.

Models and Evaluation

Our primary evaluation metric was recall: it shows the actual positive instances (churned customers) that the model correctly identifies. .


Decision Tree Model Then used its Feature Importance to select most significant features
Decision Tree Model tuned by cross validation of thresholds and using the best threshold selected
Grid search used to get best params which were then used on a Random Forest Model
Baseline Model
Model 2
Model 3
Model 4
Model 5
Decision Tree Model with the top most important features selected and 0.1 threshold
Random Forest Model: With the most important features selected after Feature

Model 1 and its Evaluation

Baseline Model

- Decision Tree Model

Feature importance results

- Cumulative daily charges,
- Customer service calls,
- Number of voicemail messages,

.

Evaluation

- High Accuracy of 0.98
- High Recall of 0.87

Model 2 and its Evaluation

Model

- Decision Tree Model with the top most important features selected
- 0.1 threshold

Evaluation

- Lower accuracy of 0.95
- A reduced recall of 0.68

Model 3 and its Evaluation

Model

- Decision Tree Model tuned by cross validation of thresholds between 0 and 0.2
- Best threshold selected 0.011
- Used the best threshold selected to create model

Evaluation

- Recall rose to 0.871
- Accuracy went up marginally 0.951 .

Model 4 and its Evaluation

Model

- Random forest model with the most important features from the results of Feature Importance

Evaluation

- Recall rose marginally to 0.8713
- Accuracy went up to a very remarkable 0.9805.
- F1 score of 0.9312
- The cross-validation mean recall 0.8559 and F1 score 0.9220

Model 5 and its Evaluation

Model

- Grid search-tuned random forest model
- Model using best parameters indicated using Grid Search
- Best parameters were: 'max_depth': None, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100,

Evaluation

- accuracy of 0.9805
- recall of 0.8713.
- The F1 score was 0.9312,
- Cross-validation mean recall and F1 score at 0.8571 and 0.9230,

Evaluation Interpretation

The accuracy, confusion matrix, and F1 score remained almost identical before and after grid search.

Interpretation

Hyperparameter tuning did not result in significant changes in the model's performance.

This could mean that the model is at its optimal performance

Evaluation Conclusion

Given the marginal improvement and resource intensity of the grid search tuned random forest model, the **original Random Forest is chosen as the final model.**

confusion matrix shows a high number of true positives (88) no false positives (FP = 0).

This indicates that the model is not incorrectly classifying non-churn instances as churn.

EDA

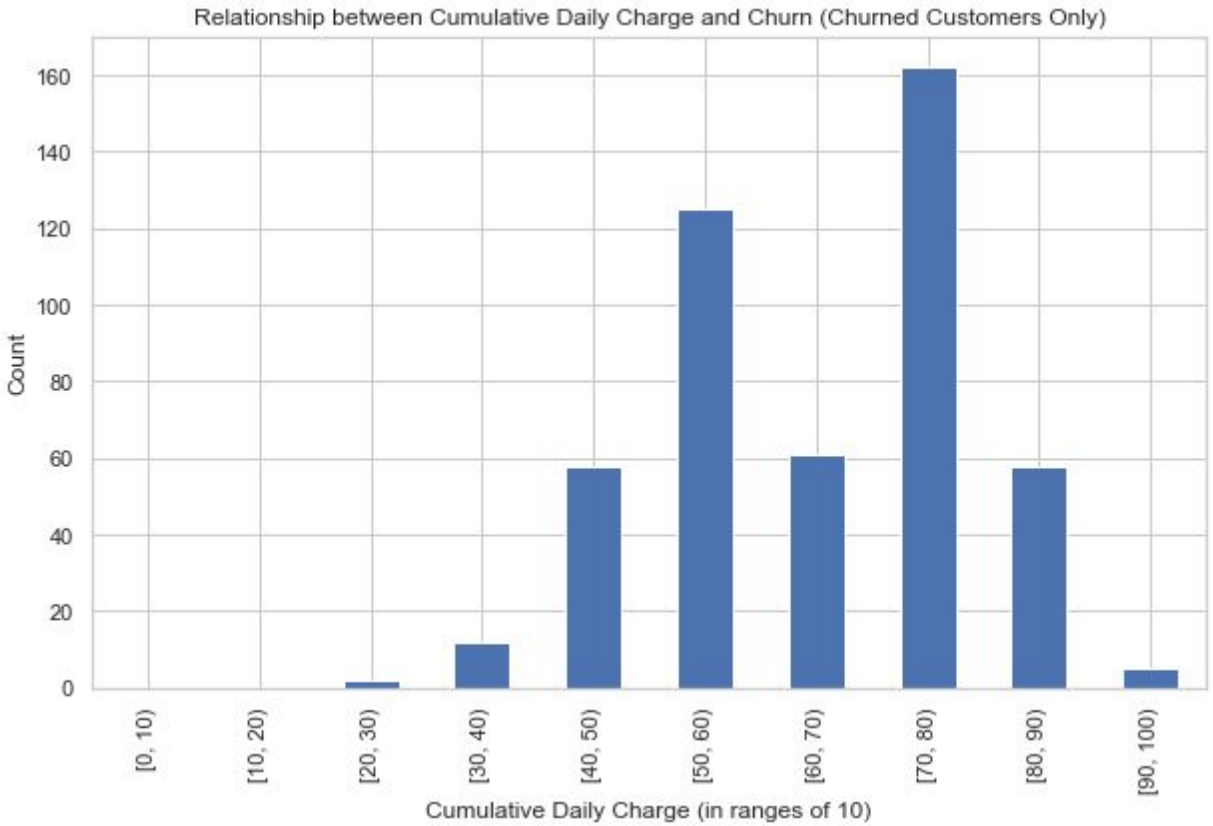
Churn rate : 14%
From initial EDA

From the features biggest most importance after doing Feature Selection.

Most impotant features in ascending order

1. Cummulative daily charge
2. Customer service calls
3. Number of voicemail messages

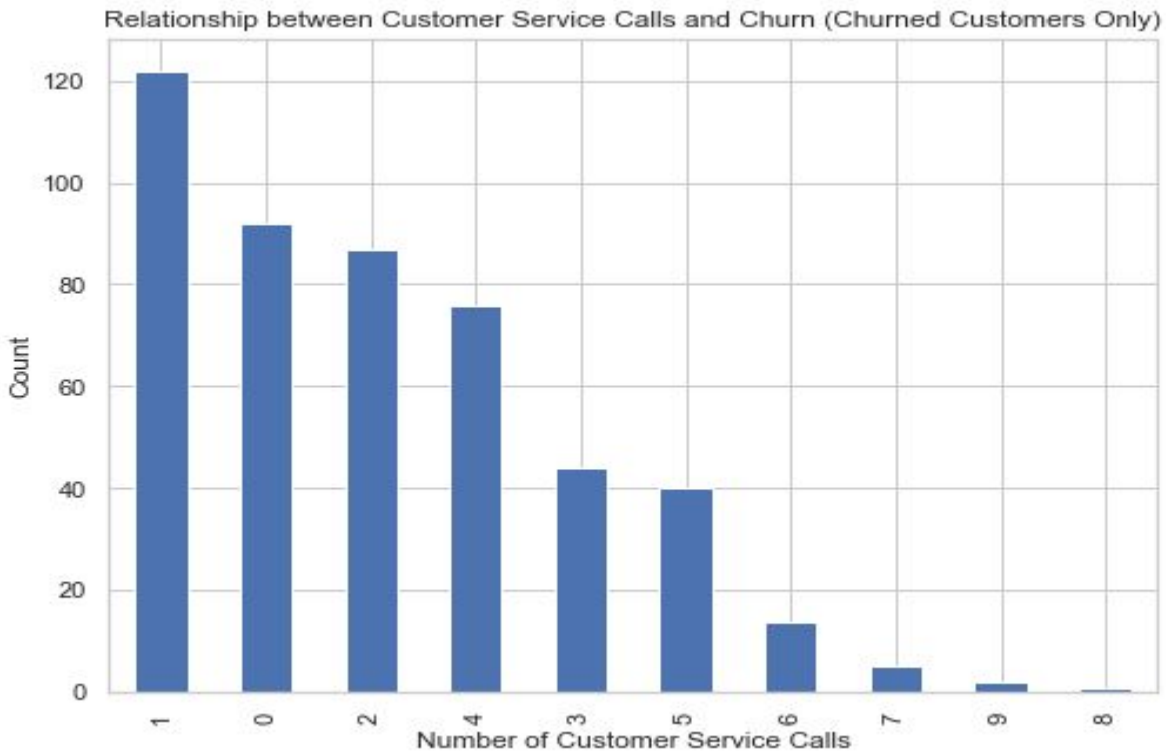
Cummulative Daily Call Charges vs Churn



Most customers who churned had high cummulative charges per day.

High charges above the competitors rate could lead to high churn rate

Cummulative Daily Call Charges vs Churn

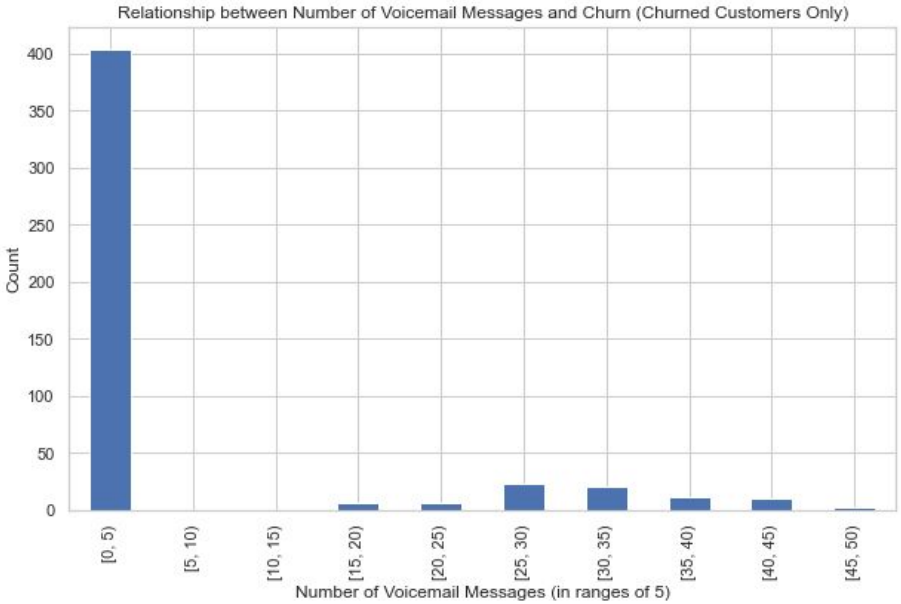


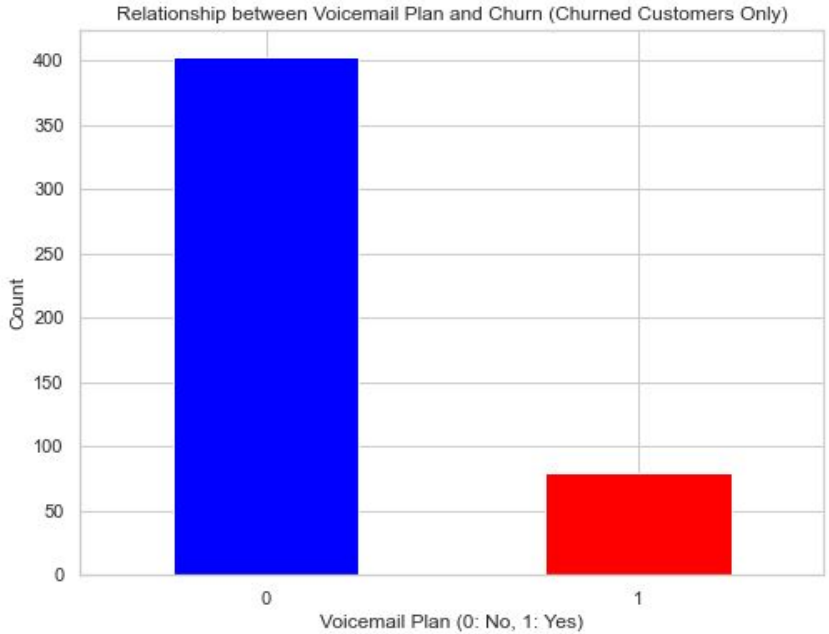
More than half the churned customers made less than 4 calls.

Could be an indication of

- Low customer service agent availability
- Poor assistance from customer service agents

Relationship between churned customers and voicemail





Most customers who churned made less than 5 voicemail messages and didn't have a voicemail plan

Recommendations

SyriaTel is recommended to:

1. I**mprove customer service availability and support** know how to avoid churn from frustrated customers
2. Make their **call charges competitive** to avoid church from customers with high cumulated charges.
3. M**ake voicemail plan part of normal call charges (industry standard)** and not an additional fee to reduce churn from customers without the plan.

Next Steps

For deeper insight and model improvement, SyriaTel could provide data about based on other telecom services they offer such as

- Internet services
- Money transfer and payment services
- Customer age groups

Q&A

Thank you for your time.

Floor is open for questions