Leveraging Statistics & Machine Learning to Accelerate 5G Adoption

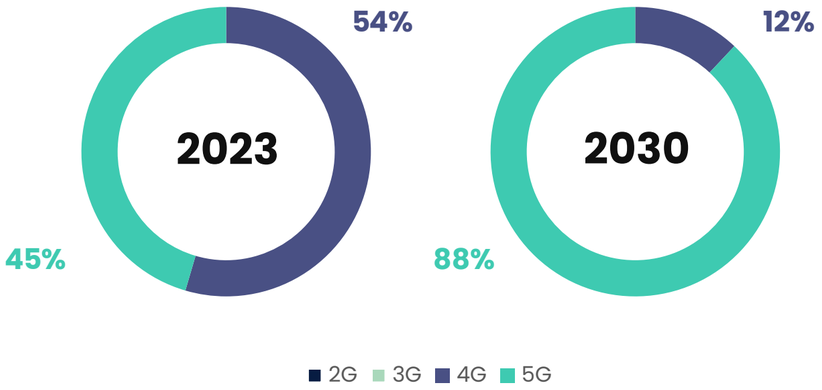
A Strategic Approach for Telecommunications Growth

AN6003 Group A – Team 9

Entering A New Era: 5G Adoption & Business Opportunity

5G Advantages

5G Adoption & Prediction in China
(% of total connections)



For successful 5G expansion, in 2023-2030, Chinese operators are projected to invest in mobile capex for **$320 billion** ~ 20% global mobile capex

Business Problem

Leveraging machine learning to accurately identify potential 5G users for targeted marketing campaigns and optimized resource allocation.

Limeylia Ng

Dataset Introduction

A real dataset from a telecommunications company

140K records

10 variable groups

43 feature variables

1 target variable

feature variables

- User Identification
- User Information
- Consumer Behavior Information
- Excess Information
- Broadband Information
- Contract Information
- Package Information
- Data Saturation Information
- Other Information

target variable

- Label

Whether this user is converted to a 5G user this month

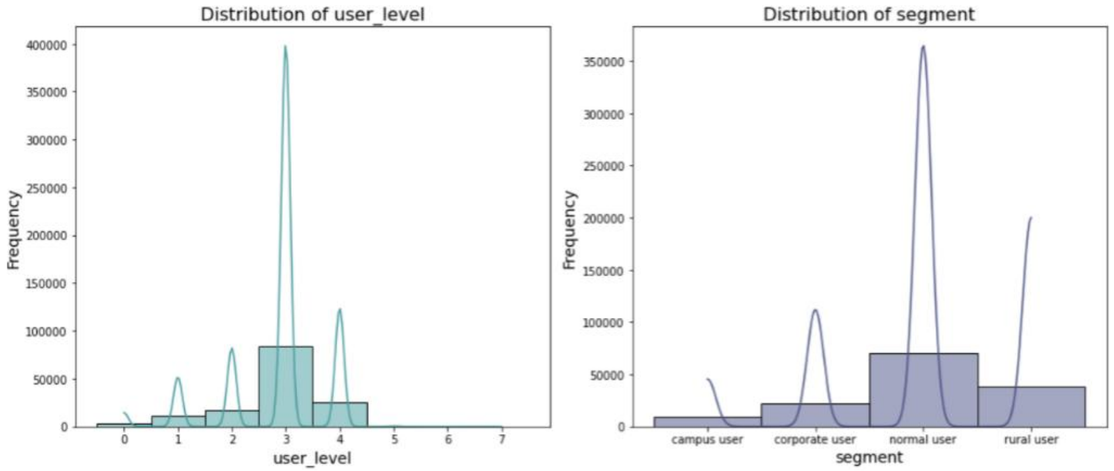
Data Preprocessing

Handling default values in variables

categorical variables

- **user_level**
- **segment**

Fill NAs with modes



numerical variables

Fill NAs with medians

multiple default values in feature variables within a single data record

Remove the records

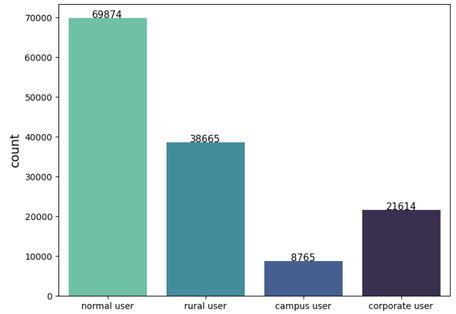
Exploratory Data Analysis

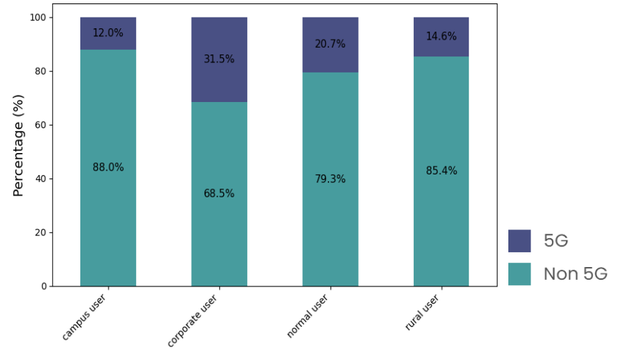
The conversion rate of 5G is quite low, so there is great potential for future 5G growth. Besides, Unique patterns can be identified among different variables.

non 5G users

The distribution of 5G and non 5G users is quite unbalanced, with a ratio of 4:1.

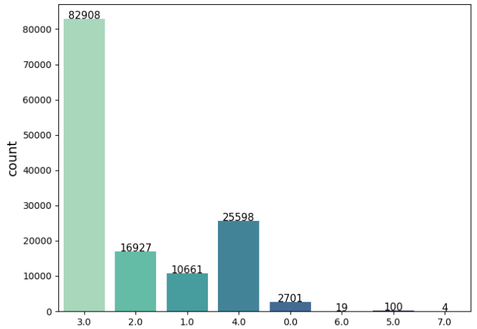
Segment

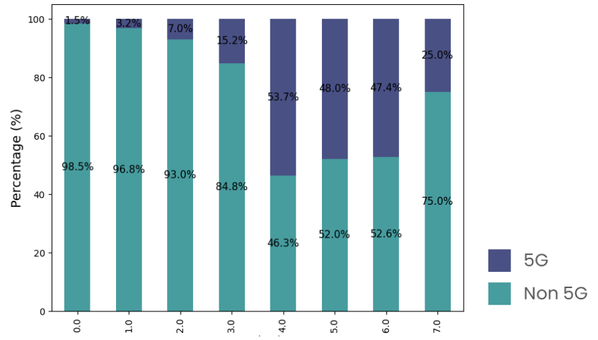




Corporate users have the highest 5G conversion rate, while campus users have the lowest 5G rate and user amount.

User Level





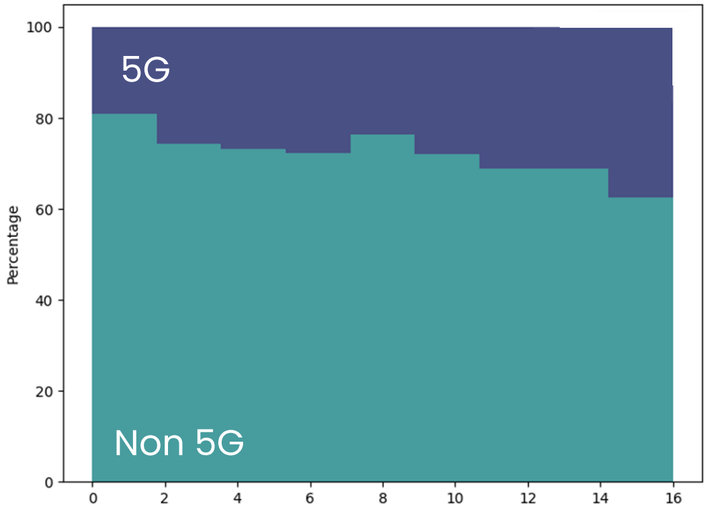
Starting from user level 4, 5G conversion rate is much higher than those of lower-level users.

Zhang Hang

Trend of 5G Conversion

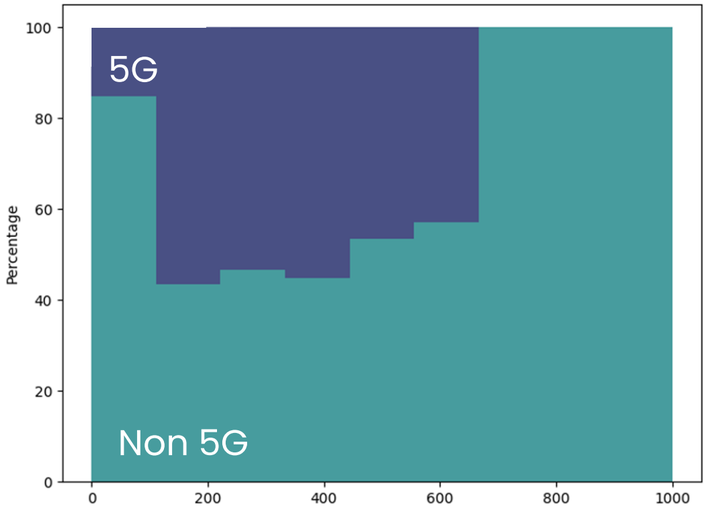
The proportion of 5G conversions varies with the growth of the numerical variable.

Tenure in Years



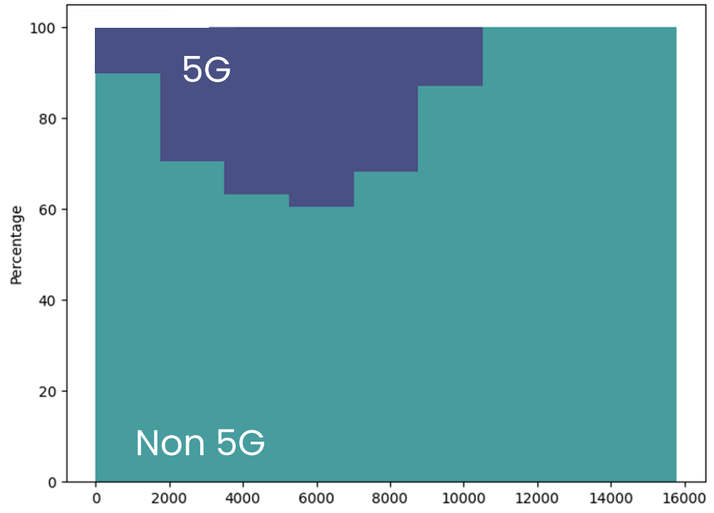
Users with longer tenure years switch to 5G more than those with shorter ones.

Avg ARPU in past 3 months



The proportion of 5G increases significantly at ARPU range of 100 – 600 RMB, compared to 0 -100 RMB.

Avg DOU in Past 3 Months

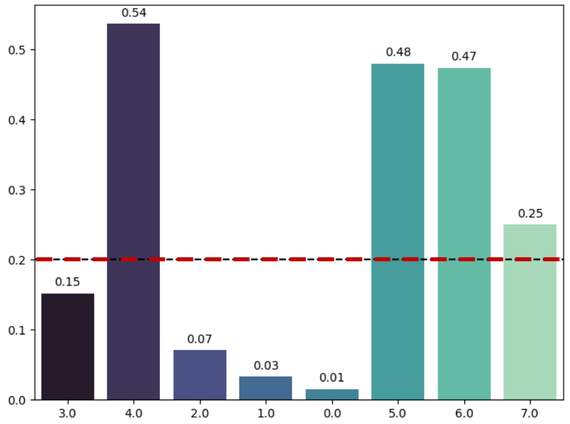


5G user proportion peaks at round 6000MB of DOU, which means users using more data tend to switch to 5G in certain range.

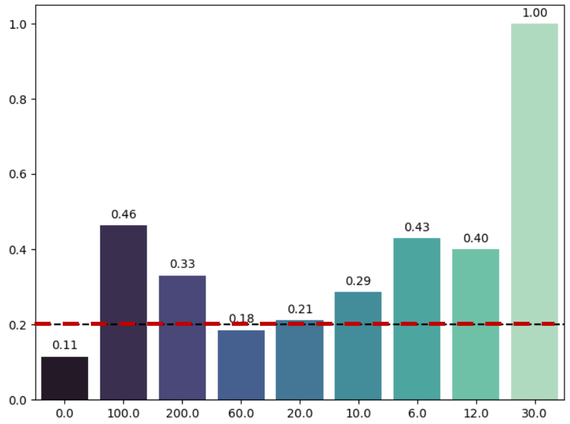
ARPU: Average Revenue Per User | DOU: Data Usage per User

Zhang Hang

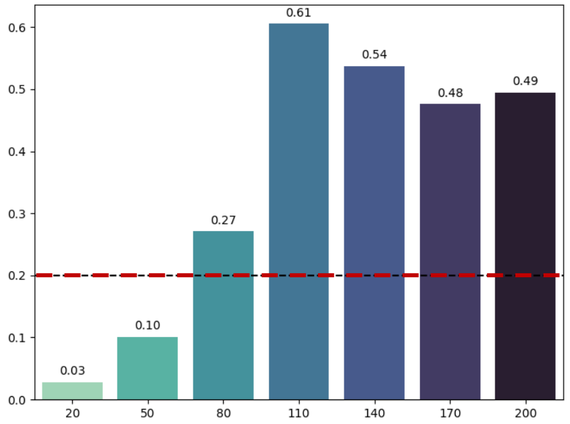
User Level



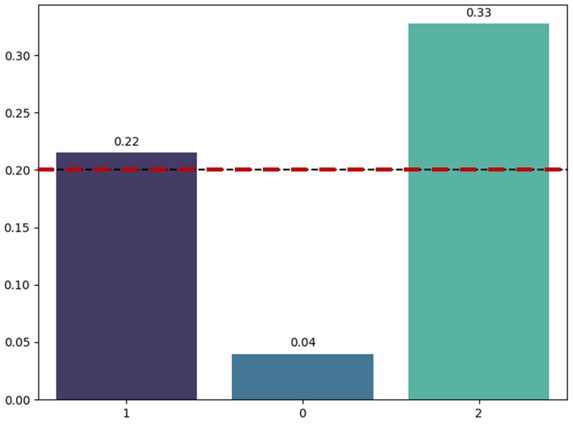
Broadband Width



Avg ARPU in Past 3 Months



Terminal Type



5G Rate

The charts on the left show the categorical attributes that have a significant impact on the 5G adoption rate, using 0.2 as the benchmark.

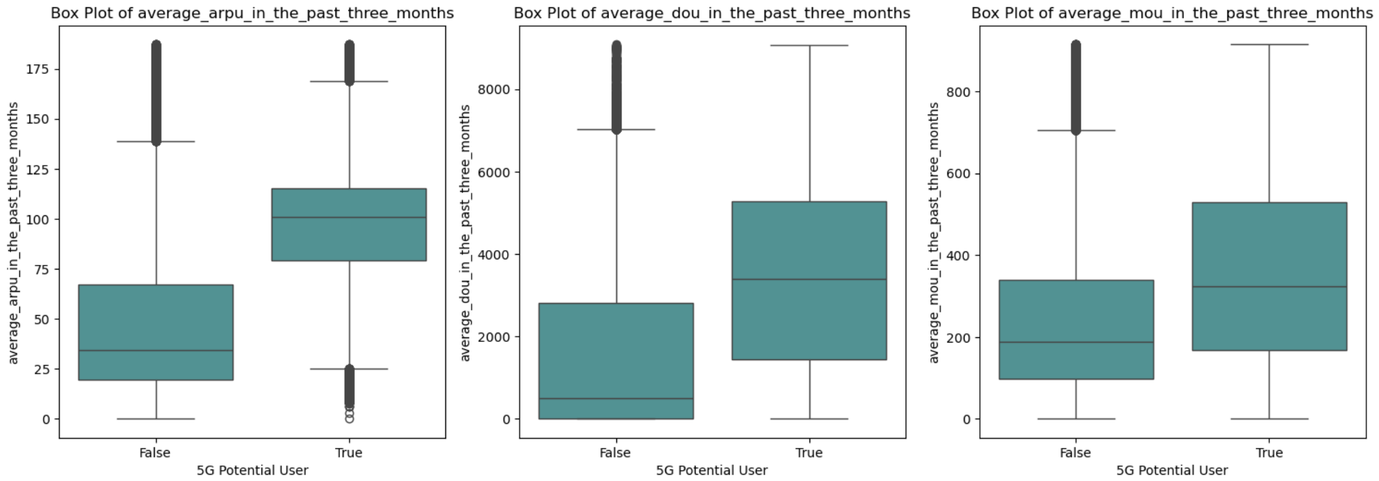
Cao Yawei

Key Variables of 5G Conversion

Analysis revealed key variables that are important and positively-correlated with customer adoption of 5G services.

Top 5 Important Features	Top 5 Positively-Correlated
Average ARPU	Average ARPU
Average DOU	User level 4
Average MOU	Broadband bandwidth
Age	Average DOU
Broadband bandwidth	Home user

Variables Relationship



Users with higher ARPU are more likely to convert to 5G.

Heavy data users are more inclined to upgrade to 5G.

Users who spend more time using services are more likely to adopt 5G.

Target Customers

Users with high ARPU, DOU, and MOU
Broadband users with higher bandwidth
Higher-tier customers (level 4)
Users with home network

ARPU: Average Revenue Per User | DOU: Data Usage per User | MOU: Minutes of Usage

Limeylia Ng

Machine Learning - Features Engineering

Feature Generation

- Features in current dataset reflect only static behavior.
- Developed features based on past three months data to track average & trend.

Average Features

- Voice Overage Fees
- Data Overage Fees
- Data Saturation Level

3 Months Average

Trend Features

- ARPU
- DOU
- MOU
- Voice Overage Fees
- Data Overage Fees
- Data Saturation Level

Monthly Growth Rate

Feature Selection

- Exist high degree of correlation among features in the dataset, may lead to several issues during model training.

Correlation Analysis → **Threshold : 0.9**

Features to Drop

- 3 Months Average ARPU
- 3 Months Average MOU
- Broadband Bandwidth
- Is Boardband Activated
- Is Boardband Contract Bundled
- Main Package Fee
- 3 Months Avg Data Saturation Level

Standardization & Encoding

- Numerical features in the dataset have inconsistency.
- Features with larger values may influence model training.

Numerical Features → **StandardScaler** → **Mean = 0 Std. = 1**

- Dataset includes categorical features with multiple categories like user level and segment.

Categorical Features → **Target Encoding** → **Mean of "Label"**

Segment	Campus User	Corporate User	Normal User	Rural User
Encoded Segment	0.120251	0.314981	0.206844	0.146179

STEP 1

Generate 15 new features

138918 Records | 59 Columns

STEP 2

Drop 7 highly correlated features

138918 Records | 52 Columns

STEP 3

Adjust Numerical & Categorical Features

138918 Records | 52 Columns | All between 0 - 1

Note: ARPU: Average Revenue Per User | DOU: Data Usage per User | MOU: Minutes of Usage

Feng Wenzhuo

Machine Learning - Model Preparation

Evaluation Metrics

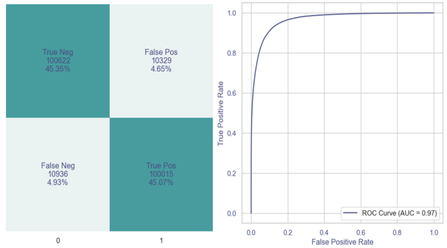
Primary Model Evaluation Strategy : **5 Fold Cross-Validation**

Model Performance Metrics

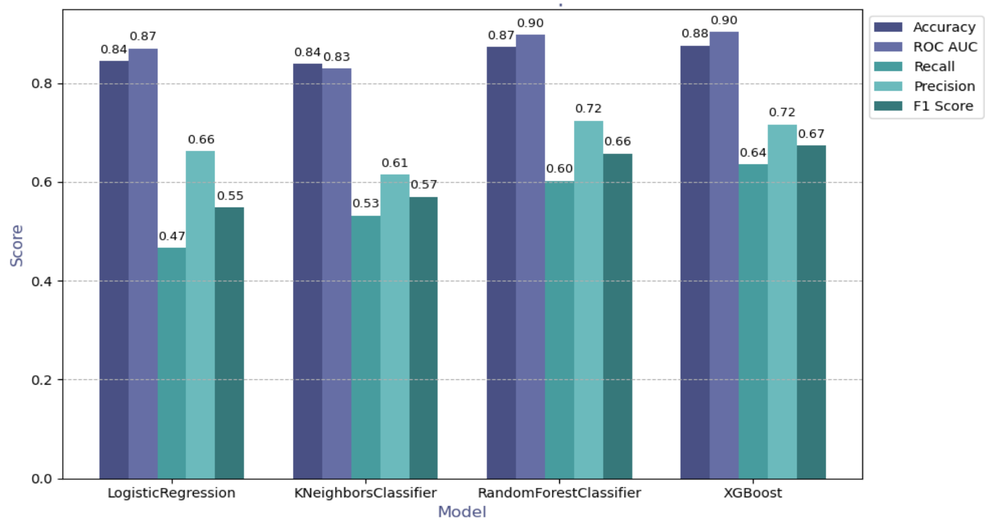
Accuracy **Precision** **Recall**

ROC AUC **F1 Score**

Confusion Matrix & ROC Curve



Model Performance with Unbalanced Data



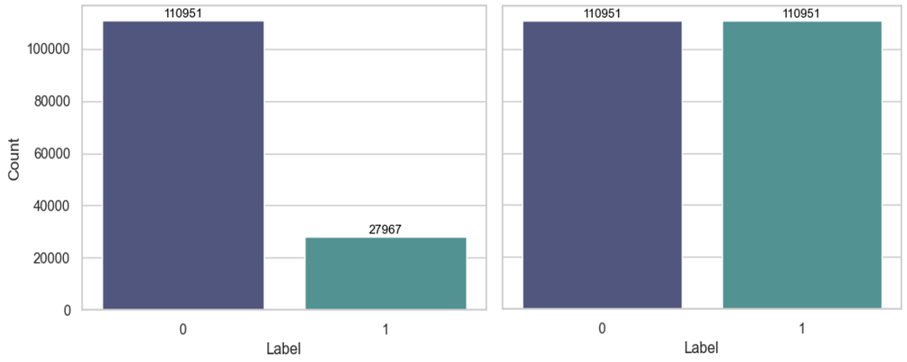
Imbalanced Data Resampling

- Initial machine learning models with **high accuracy** (0.84+) but **low recall** (0.64-).
- Distribution of 0 and 1 samples is uneven, approximately **4:1**, make the model to favor the majority class.
- Use **SMOTE** to adjusted the ratio to **1:1** and get a **balanced dataset**.

SMOTE

Synthetic Minority Over-sampling Technique, generates synthetic minority samples to balance the ratio between positive and negative samples.

Label Distribution Before and After SMOTE



Feng Wenzhuo

Machine Learning – Logistic Regression Model

Model Overview

Logistic Regression

Logistic Regression estimates class probabilities for binary classification by applying the logistic function to a linear combination of input features.

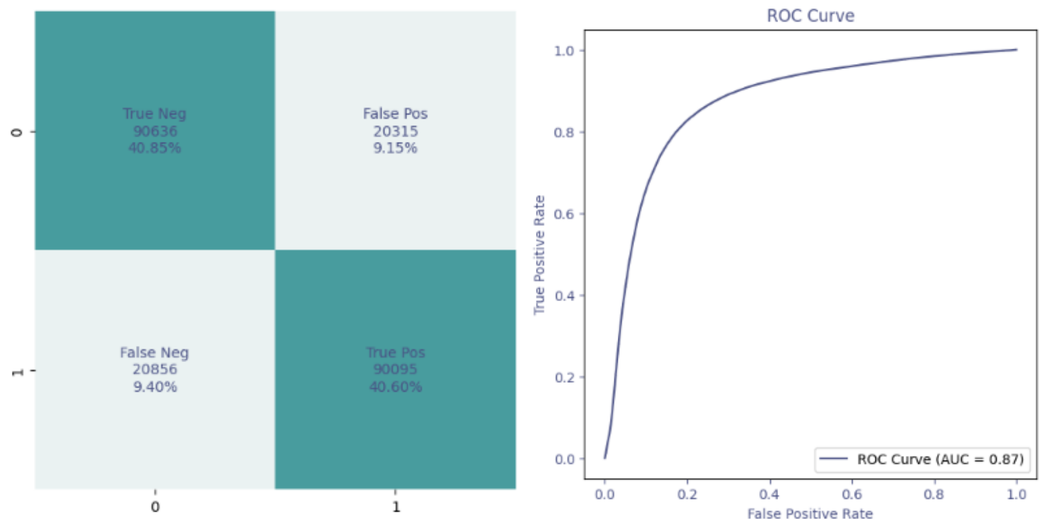
Results and Analysis

Logistic Regression Model Performance Metrics

Model Name	Accuracy	AUC	Recall	Precision	F1 Score
Logistic Regression	0.8143	0.8724	**0.8118**	0.8157	0.8134

- Logistic Regression model achieved an **accuracy** of **81.43%** and a **high recall** of **81.18%**.
- Accuracy means, out of 100 predictions whether a user will switch to 5G or not, about 81 were correct.
- Recall means, if there are 100 actual 5G users, the model correctly identified about 81 of them.

Logistic Regression Confusion Matrix and ROC Curve



Marcus Vos

Machine Learning – Logistic Regression Variants

Regression Variants

The **Lasso variant (L1)** improves prediction accuracy and interpretability by selecting and regularizing variables. The **Ridge variant (L2)** prevents overfitting by penalizing large coefficients and addressing multicollinearity. The **Elastic Net variant (L1 and L2)** combines and balances the benefits of both.

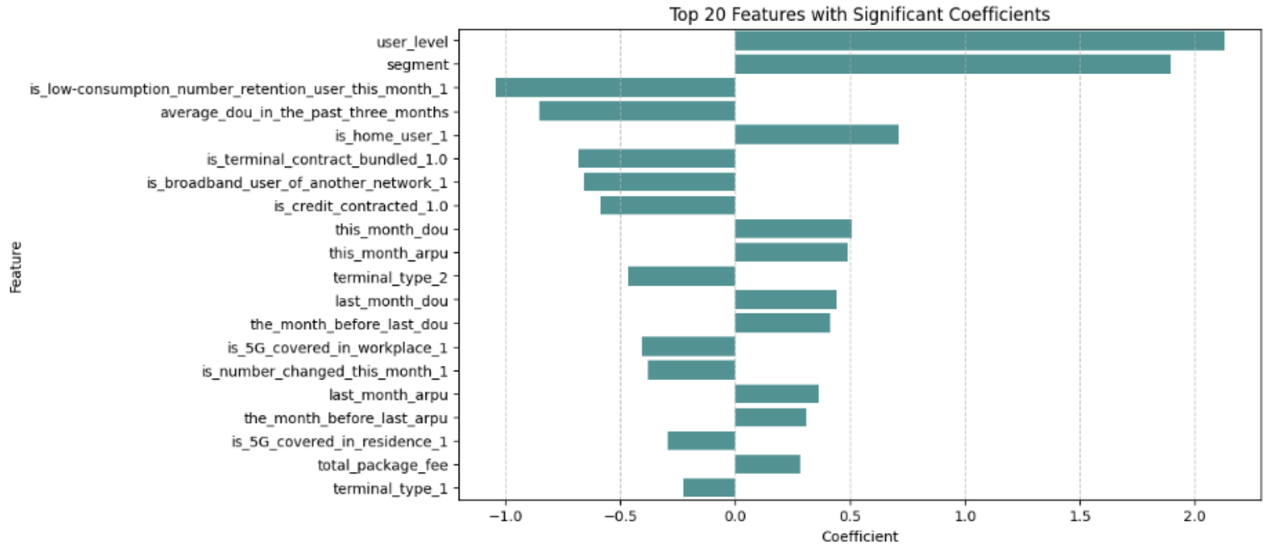
Logistic Regression Model Variants Performance Metrics

Model Name	Accuracy	AUC	Recall	Precision	F1 Score
Binary Logistic Regression	0.8143	0.8724	**0.8118**	0.8157	0.8134
Lasso Logistic Regression	0.8144	0.8724	**0.8120**	0.8158	0.8135
Ridge Logistic Regression	0.8144	0.8724	**0.8119**	0.8158	0.8135
Elastic Net Logistic Regression	0.8144	0.8724	**0.8120**	0.8158	0.8135

- All variants of logistic regression provide similar performance across all metrics.
- Consistent performance across variants could be due to the dataset having linear relationship between features and the target, with minimal noise.
- Different regularization techniques did not significantly impact performance.

Marcus Vos

Machine Learning – Lasso Coefficients Analysis



Coefficient Analysis

- Based on the **coefficients from the lasso regression model, we identified the most significant factors** influencing the likelihood of users switching from 4G to 5G.
- **Positive coefficients**, such as User Level, User Segment, and Home User Status, indicate an increased likelihood of conversion to 5G.
- **Negative coefficients** decrease the likelihood of conversion:
  - Low Consumption Retention User:
  - Low Average Data Usage (DOU)
  - Terminal Contract Bundled
  - Is Broadband User of Another Network
  - Is Credit Contracted
- **6 out of the 50 features** in the lasso regression model have **coefficients reduced to zero**, indicating that **these features provide no benefit** to the model's predictions.

Machine Learning – Top 10 Features Logistic Regression

Model Overview

Logistic Regression

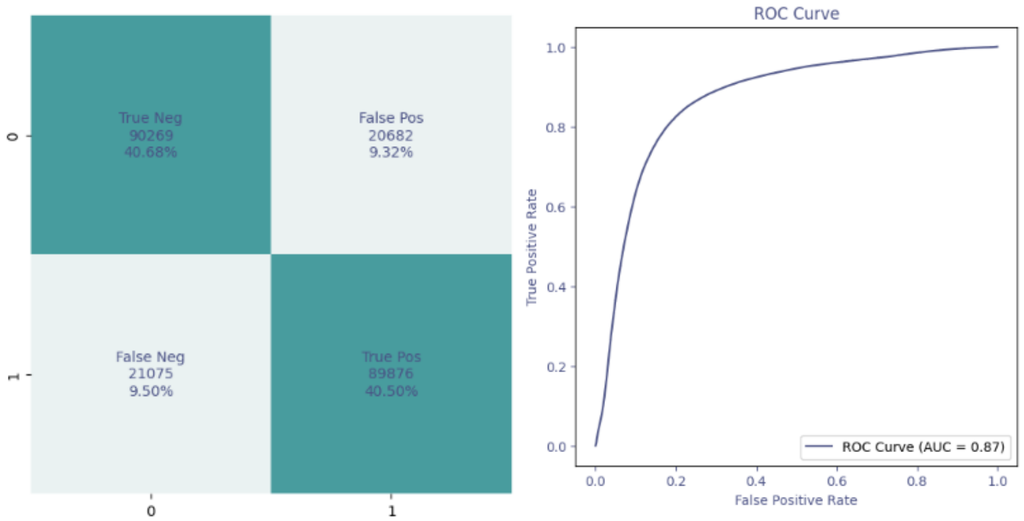
Our testing revealed that using only the top 10 features determined by the coefficients from the lasso regression model is sufficient to build a satisfactory regression model.

Results and Analysis

Top 10 Features Logistic Regression Model Performance Metrics

Model Name	Accuracy	AUC	Recall	Precision	F1 Score
Top 10 Logistic Regression	0.8118	0.8681	**0.8100**	0.8129	0.8113
Logistic Regression	0.8143	0.8724	**0.8118**	0.8157	0.8134

Top 10 Features Logistic Regression Confusion Matrix and ROC Curve



- Top 10 Logistic Regression model achieved similar performance metrics as the other logistic regression models.
- By using the top 10 features , we can get similar model performance as compared to the original model which had 51 features.
- Reducing the number of features to the top 9 causes the accuracy to drop to 0.72 .
- This model with reduced feature requirements can be advantageous if data collection is costly.

Marcus Vos

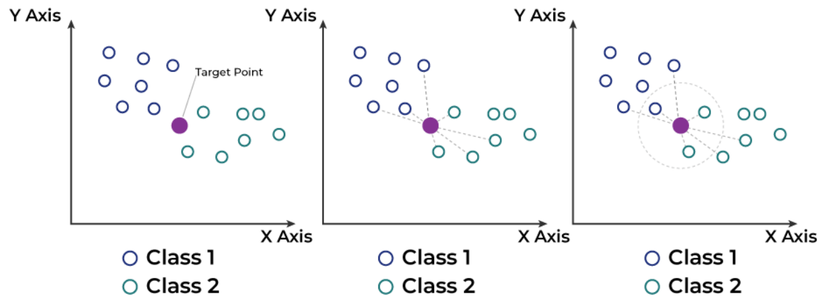
Machine Learning - KNN Model

Model Overview

KNN

K-Nearest Neighbors (KNN) algorithm is a simple, instance-based learning method. It classifies a sample based on the majority class of its K-nearest neighbors, determined by a distance metric (e.g., Euclidean distance).

KNN Algorithm

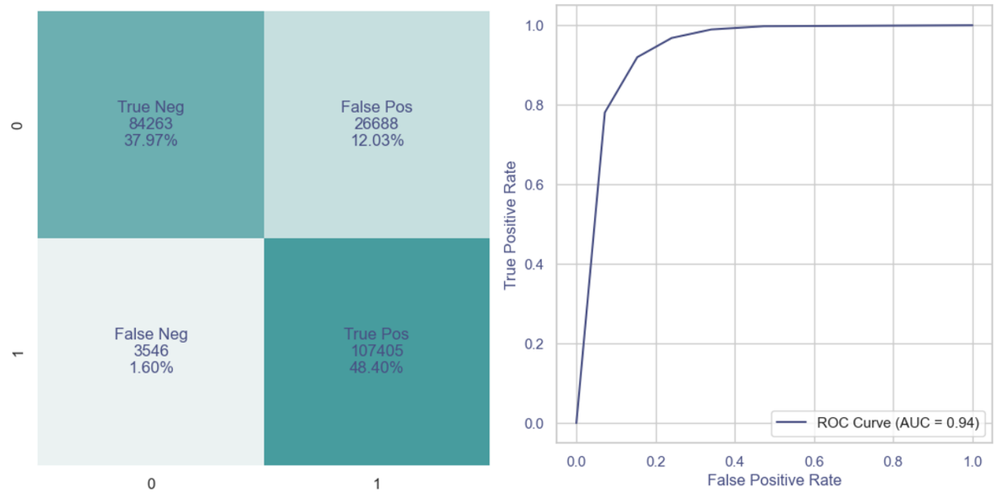


Results and Analysis

KNN Model Performance Metrics

Model Name	Accuracy	AUC	Recall	Precision	F1 Score
KNN Model	0.8638	0.9355	**0.9680**	0.8010	0.8766

KNN Model Confusion Matrix and ROC Curve



- KNN model achieved an **accuracy** of **86.38%** and a **high recall** of **96.80%**.
- Highly effective at identifying positive class users, which aligns with our business goal of accurately recognizing positive users (e.g., 5G users).
- Precision is 80.10%, showing that the model misclassifies more negative class users (non-5G users) as positive, incorrectly flagging them as 5G users.

Feng Wenzhuo

Machine Learning - Random Forest Model

Model Overview

Random Forest

Random Forest algorithm builds multiple decision trees using different subsets of the data and combines their predictions. One of its key strengths is its ability to reduce overfitting, meaning it performs well on both the training and testing data.

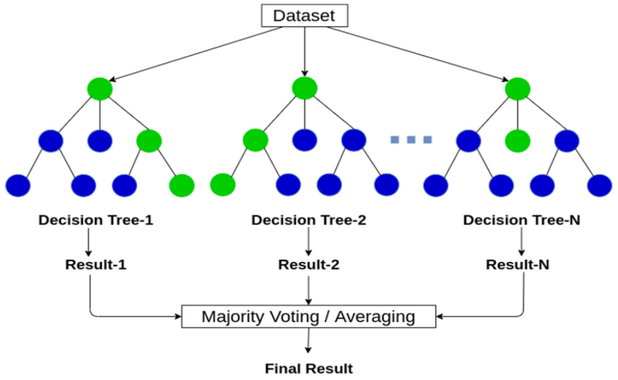
Results and Analysis

Random Forest Model Performance Metrics

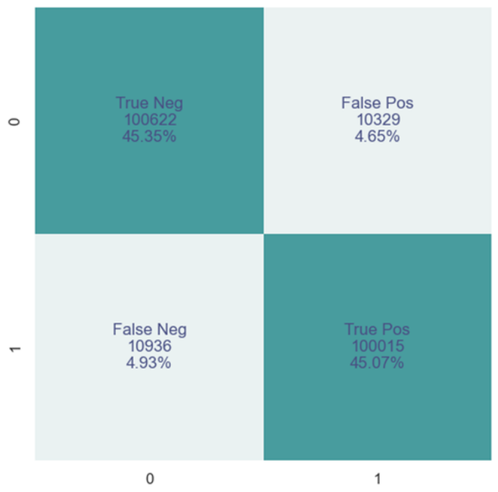
Model Name	Accuracy	AUC	Recall	Precision	F1 Score
Random Forest	0.9042	0.9705	**0.9014**	0.9070	0.8998

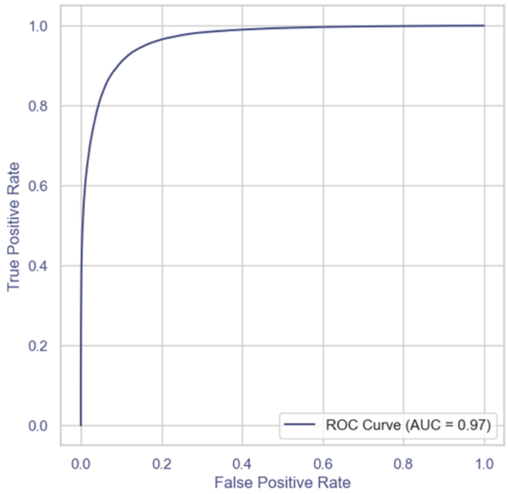
- Random Forest achieved an **accuracy** of **90.42%** and a **recall** of **90.14%.**
- Results indicate that RF model performs very well in identifying both positive and negative class samples, with a strong balance between sensitivity (Recall) and specificity.
- The high Precision of 90.70% suggests that most positive samples predicted by the model are indeed true positives, reducing the number of false positives.

Random Forest Algorithm



Random Forest Model Confusion Matrix and ROC Curve





Feng Wenzhuo

Machine Learning - XGBoost Model

Model Overview

XGBoost

XGBoost is a gradient boosting framework. It builds multiple weak learners (typically decision trees) sequentially, where each new tree aims to correct the errors made by the previous ones. XGBoost is highly efficient for large datasets and excels in handling imbalanced data.

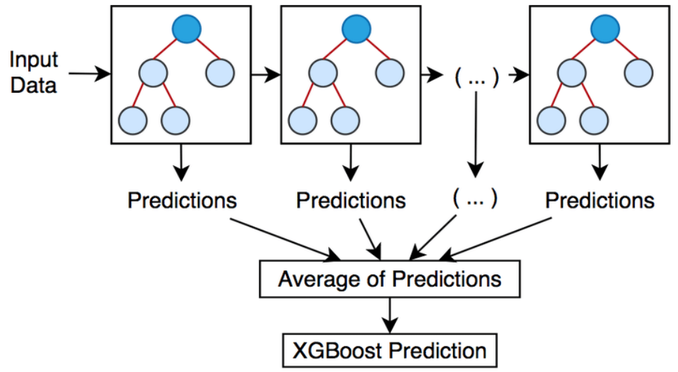
Results and Analysis

Random Forest Model Performance Metrics

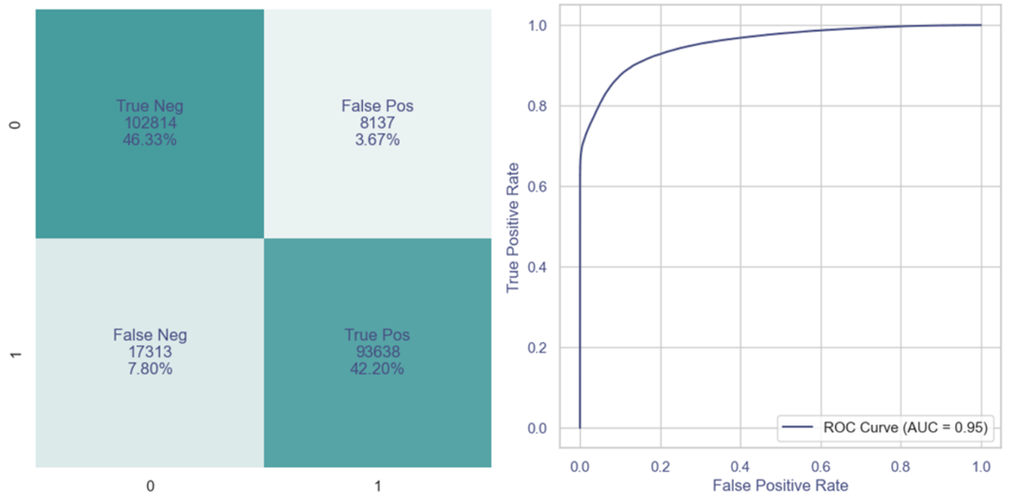
Model Name	Accuracy	AUC	Recall	Precision	F1 Score
XGBoost	0.8853	0.9685	**0.8440**	0.9209	0.8596

- XGBoost model achieved an **accuracy** of **88.53%** and a **recall** of **84.40%**.
- While the recall is lower compared to KNN and Random Forest, the Precision is notably higher at 92.09%, meaning XGBoost is more effective at reducing false positives.

XGBoost Algorithm

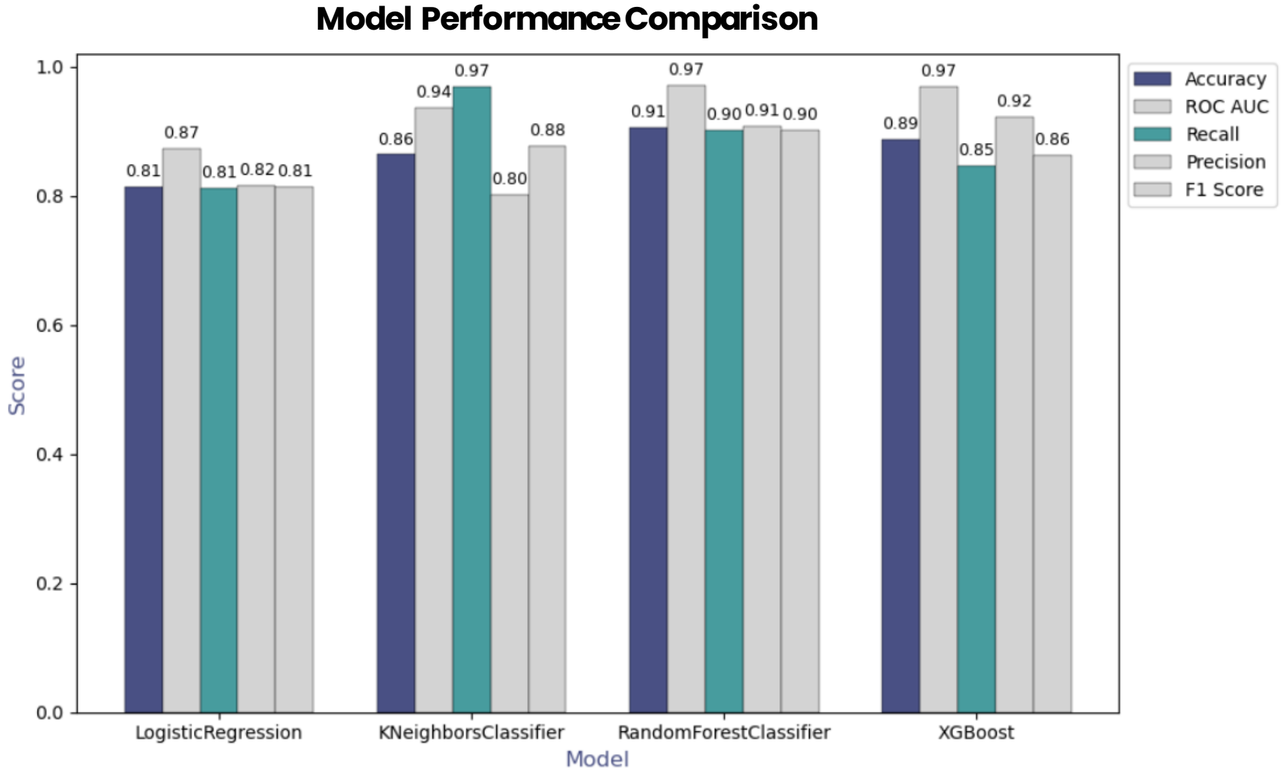


Random Forest Model Confusion Matrix and ROC Curve



Feng Wenzhuo

Machine Learning – Model Comparison



Results and Analysis

- **Logistic Regressions** show balanced performance, with accuracy, AUC, recall, precision remaining stable at consistent levels.
- **KNN Model** stands out with the highest recall (96.80%) but low precision (80.10%), meaning the model can identify most of potential 5G users but have a higher number of false positives.
- **Random Forest Model** shows median recall (90.14%), along with high precision (90.70%), not only effective at identifying potential 5G users but also manages false positives well.
- **XGBoost Model** achieves the highest precision (92.09%), excelling at minimizing false positives. However, its recall is comparatively lower (84.40%).

Business Requirement

In the context of 5G business applications, expanding the marketing scope to include as many potential 5G users as possible, within acceptable cost limits, is the priority. Thus, the goal is to identify potential 5G users for targeted marketing, making recall the most critical metric.

Best Model

👑 **KNN**

Feng Wenzhuo

Business Strategies

Target on ARPU, DOU & MOU

From the exploratory data analysis, we can conclude that the key factors influencing the switch to 5G are consumption-related variables, dataflow usage variables and voice usage variables.

High-spending Users

Focus on promoting premium 5G plans with enhanced features:

- **Faster speeds**
- **Exclusive content**
- **Additional services**
- ……

Call-focused Users

Promote the superior call quality and network stability of 5G:

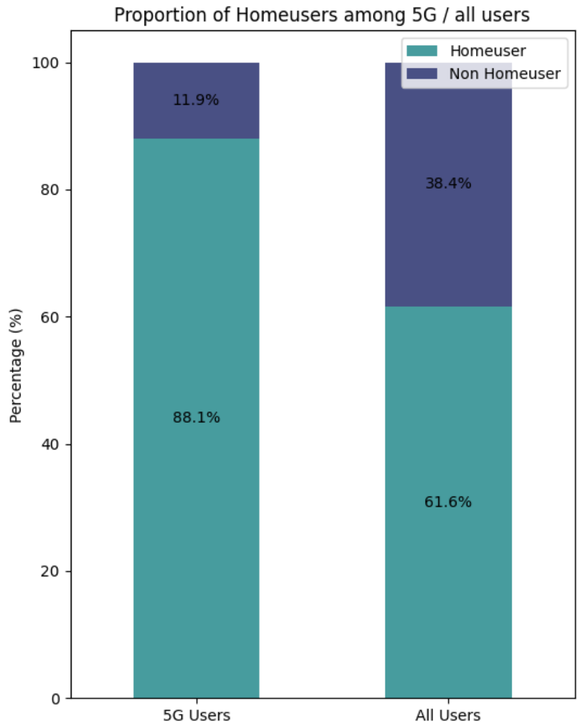
- **HD voice calls**
- **Video calls**
- **Rich text messages**
- ……

Data-hungry Users

Encourage upgrade for a better experience of activities like:

- **Streaming**
- **Gaming**
- **Cloud Services**
- ……

Zhang Hang



5G users are more likely to be home users compared to the overall user base. 88.1% of 5G users are home users, compared to 61.6% of all users.

Business Strategies

Target on Home Users

5G Household Bundles

Introduce family or household bundles that combine 5G mobile services with home broadband, TV, or home devices, offering discounted prices.

Family Shared Data Plans

Offer family shared data plans that allow multiple members to share 5G data within the same plan, making it easier to manage costs.

5G-Enabled Smart Home Solutions

Highlight the advantages of 5G for smart home devices (e.g., smart cameras), demonstrating how 5G can enhance home life.

Zhang Hang

Business Strategies

Target on Broadband Users

Network Bundling Discounts

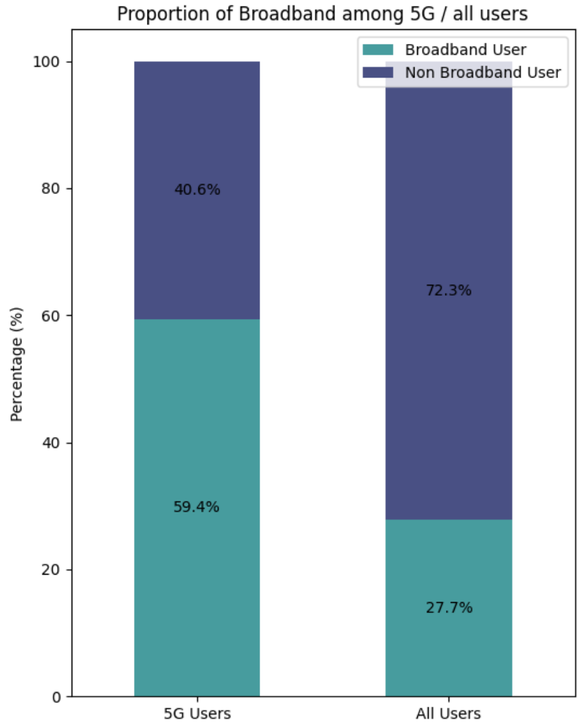
Provide existing in-network broadband users with exclusive discounts on bundled 5G mobile and broadband services.

Seamless Upgrade and Experience

Highlight the seamless transition between 5G and home broadband networks. Emphasize the superior, uninterrupted network experience.

Data Prioritization

Offer prioritized data speeds for broadband users on their 5G mobile plans, ensuring faster and more reliable connectivity when both networks are in use.



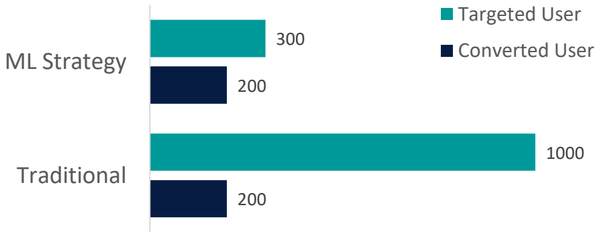
5G users are more likely to be broadband users compared to the overall user base. 59.% of 5G users are broadband users, compared to 27.7% of all users.

Zhang Hang

Model Application in Business

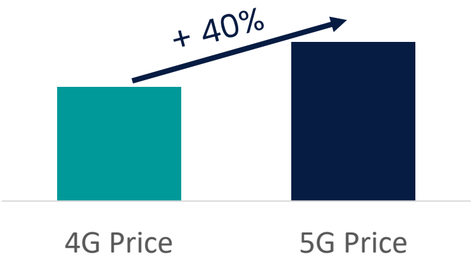
Targeted Outreach

- **Traditional 5G promotion** relies on a broad-based approach, using large-scale marketing campaigns to reach as many potential 5G users as possible.
- With the **machine learning model**, telecom operators can leverage data-driven insights to precisely **target users** who are **most likely to transition** from 4G to 5G.



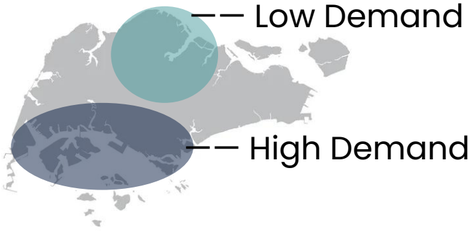
Precision Marketing

- Tailor specific 5G plans and service offerings based on users' behavior.
- Once high-potential customers are identified, operators can incentivize these users to upgrade by offering small discounts or promotions.
- Operators can provide **proactive technical support** and upgrade assistance for potential users to accelerate their conversion to 5G.



Network Resource Optimization

- Identify user groups most likely to upgrade to 5G and pinpoint their **geographic distribution or user segment**.
- Prioritize the deployment of 5G base stations; Expand network capacity in these areas, ensuring that users receive high-quality service when transitioning to 5G.



Feng Wenzhuo

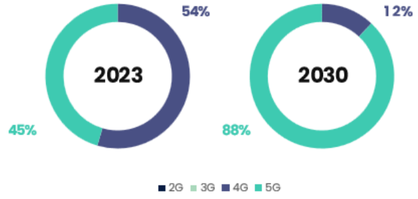
Conclusion

This data-driven strategy empowers the company to maximize customer adoption and resource optimization, securing its position in the 5G market.

Business Problem

Accurate identification of potential 5G users for targeted marketing campaigns and optimized resource allocation

5G Mobile Adoption in China
(% of total connections)



Exploratory Data Analysis

Unique patterns can be identified among different variables.

Top 5 Important Features	Top 5 Positively-Correlated
Average ARPU	Average ARPU
Average DOU	User level 4
Average MOU	Broadband bandwidth
Age	Average DOU
Broadband bandwidth	Home user

Machine Learning

Feature Engineering

Model Preparation

- Logistic Regression Model
- Random Forest Model
- XGBoost Model

Best Model 👑 **KNN**

accuracy 86.38%
recall 96.80%

Business Solutions

Segmentation

- ARPU, DOU & MOU
- Home Users & Broadband Users

Model Application in Business

- Targeted Outreach
- Precision Marketing
- Network Resource Optimization

Cao Yawei

Thank you for listening!

Leveraging Statistics & Machine Learning to Accelerate 5G Adoption

A Strategic Approach for Telecommunications Growth

AN6003 Group A – Team 9