Executive Summary

Problem Statement

In our quest to optimize user engagement on NetEase Cloud Music's (NCM) platform, we delved into understanding the core characteristics and actions of active users. Our objective was to decipher early user interactions-clicks, likes, shares-that could forecast their continued activity. The ultimate goal was to refine the existing recommendation system within the Cloud Village tab, thereby fostering an increase in active user participation.

Data Exploration

Our extensive analysis covered over 57 million impressions from a 2 million user sample. The demographic insights revealed a predominance of users aged between 15-25, with a surge of new user registrations noted from 2017 to 2019. However, a disconnect was observed between user preferences and demand, and the recommended content in the Discovery Subtab portrayed in low engagement with the impressions, suggesting room for optimization.

Models Summary

User engagement was defined to be User interaction with a card from the Discovery subtab, specifically through a click on an impression (isClick variable = 1). Various modelling techniques were explored to determine the best fit for predicting user engagement, including Neural Networks, Regularized Logistic Regression, Decision Trees and Ensembled Decision Trees (Random Forest). A consistent methodology was employed across the four models, where the train data was used with a K-Fold cross validation approach and a search methods to fine-tune the model's hyperparameters. The optimal model of each was then thoroughly analysed with classification performance metrics, such as accuracy, AUC and recall. With that, Neural Networks were excluded from consideration due to their inability to predict the positive class.

The three remaining models underwent further assessment using extended metrics, such as recall at K and precision at K. After comprehensive testing, we selected a decision trees model, striking the best balance between precision at K and recall at K, surpassing the baseline recommendation system in these metrics. Decision Tree Recall at K = 0.546, compared to a baseline of 0.407, and precision at K = 0.0358, compared to a baseline of 0.0326. In contrast, Random Forests and Logistic Regression fell short in terms of recall. This decision tree model forms the cornerstone of the proposed recommendation system, prioritizing content aligned with individual user behaviour and preferences, thereby enhancing the relevance of content in the Discovery Subtab.

Furthermore, as interpretable models, decision trees and logistic regression were employed for feature analysis. Common insights emerged from both models, such as the significance of publish time, where more recent cards were more likely to be clicked. Additionally, users' historical preferences for specific artists played a significant role, laying the groundwork for developing a content-based filtering system in the future. The analysis also revealed that longer user view durations on previous impressions increased the likelihood of clicks, emphasizing the importance of the recommendation system in engaging these users with relevant content. Other findings included the higher likelihood of clicks from new users, underscoring the importance of early user engagement. Conversely, longer-tenured users tended to transition from the Discovery Tab to the Follow Tab after identifying their interests. Lastly, users who were also creators exhibited a higher likelihood of clicks, suggesting they may seek inspiration or explore the platform.

Business Actions and Recommendations

The analysis underlines the necessity for NCM to recalibrate its recommendation system. A personalized content delivery strategy, attuned to the tastes and early interactions of users, is recommended over a generalized approach that favours content from prominent creators. A proposed recommender system functions as a preference-based matching network, where user-card rankings are determined by predicted click probabilities from the best model, the decision tree. These rankings dictate the positions of card recommendations on the Discovery Subtab within the Cloud Village Tab.

Limitations and Future Improvements

Moving forward, we propose prioritizing the enhancement of our modelling techniques. A pivotal aspect will be the adoption of content-based filtering models that leverage user-item interactions, thereby reducing dependency on complex feature engineering. This approach is expected to refine the recommendation system's accuracy, ensuring a more personalized and responsive user experience. reduce the need for extensive user scrolling, similar to vertical scrolling interfaces like Spotify. This approach could enhance user experience and increase engagement with a wider variety of content. Additionally, addressing computational limitations and expanding our dataset will enable the exploration of sophisticated models such as SVMs and allow for a more granular hyperparameter optimization. Tackling data quality issues, particularly the variables with high unknown values, will improve model fidelity. Lastly, by overcoming the inherent bias introduced by the current recommendation system, our future strategy will seek to craft a recommendation system that not only reflects but also stimulates diverse user interests, fostering a robust and dynamic user base on NetEase Cloud Music's platform.

Index of Submitted Files

File	Description
00 Executive Summary and Index of Files.pdf	This document encapsulates the executive summary of the analysis, modelling, and recommendations. It serves as a guide with an index, providing easy access to key sections within the project files.
01 Final Project EDA.html	This Jupyter Notebook printout encompasses a detailed exploratory data analysis (EDA) and data preparation phase conducted before entering the modelling stage. It offers insights into data patterns and pre-processing steps, setting the foundation for subsequent modelling.
02 Modelling - Hyperparameter Tuning Notebooks	This folder contains a collection of Jupyter Notebook printouts, each dedicated to hyperparameter tuning for a specific model. These files detail the meticulous process of optimizing model performance through hyperparameter searches.
02 Final Project Modelling.html	A Jupyter Notebook printout consolidating the pre-modelling setup, optimal models resulting from hyperparameter tuning scripts, performance indicators, and an analysis of feature outcomes. This file provides a unified view of the modelling phase, facilitating a comprehensive understanding of model selection and evaluation.
03 Final Project Recommender System Assessment.html	A Jupyter Notebook printout specifically focused on assessing potential ranking recommendation systems relying on the predictive models that were built, utilizing precision at k and recall at k metrics. This document sheds light on the effectiveness of recommendation systems developed, enabling stakeholders to gauge system performance.
04 Final Project Report.pdf	A comprehensive business report presenting an in-depth overview of all analysis and findings. This PDF serves as a central document encapsulating the entire project journey, from data exploration and modelling to conclusive recommendations.