1. Pendahuluan

KKBOX, didirikan di Taiwan pada tahun 2005, telah menjadi *leading streaming music platform* di Asia, menampilkan lebih dari 40 juta *track* dan mencapai lebih dari 10 juta *customers* di Taiwan, Hong Kong, Jepang, Singapura, dan Malaysia. Meskipun sukses, KKBOX menghadapi tantangan signifikan terkait dengan *"churn" customers*, di mana *users* menghentikan *subscription* mereka, yang berdampak langsung pada *revenue* dan *growth*. Memahami dan memprediksi *churn customers* penting bagi layanan berbasis *subscription* seperti KKBOX. *Paper* ini mengeksplorasi pengembangan *model churn prediction* menggunakan *machine learning*.

Churn, merupakan fenomena penting dalam perusahaan-perusahaan yang bergantung pada customer sebagai *main asset* mereka. Ketika *customers* merasa *dissatisfied* dengan *product* atau *services* yang ditawarkan, mereka cenderung untuk berhenti menggunakan *product* tersebut. *High Churn rate* dapat mengakibatkan penurunan *growth* perusahaan secara keseluruhan. Oleh karena itu, penting bagi perusahaan untuk dapat memprediksi *customer churn behavior* dengan *accurate* agar dapat mengambil *strategic decision* yang sesuai untuk mempertahankan *customer*.

Churn prediction melibatkan identifikasi pengguna yang berpotensi untuk membatalkan *subscription* mereka, memungkinkan perusahaan untuk menerapkan *rentention strategy*. *Churn prediction* yang efektif dapat secara signifikan meningkatkan kemampuan perusahaan untuk mempertahankan pelanggan [1]. Penelitian ini menggunakan dataset KKBOX yang juga digunakan untuk competition "WSDM - KKBox's Churn Prediction Challenge". Pendekatan ini didukung oleh studi Huang dkk. yang menekankan pentingnya data komprehensif dalam meningkatkan *accuracy prediction* [2]. *Churn prediction* merupakan masalah *binary classification problem* yang bertujuan untuk menentukan apakah seorang *customers* akan *churn* atau tidak. Dalam konteks ini, *churn* dapat didefinisikan sebagai *customers* yang tidak memperbaharui *subscriptions* mereka dalam periode tertentu setelah masa *subscriptions* mereka berakhir.

Metodologi yang digunakan dalam penelitian ini melibatkan *data exploration* dan *data preparation*, diikuti dengan penerapan berbagai model machine learning, termasuk logistic regression, SVM, dan metode ensemble seperti XGBoost dan LightLGM. Mengkomparasikan model yang berbeda dapat memberikan gambaran masing-masing kelemahan dan keunggulan, yang mana menjadi salah satu tujuan riset ini [3]. Selain itu, perlu juga untuk memperhatikan *feature engineering* dan bagaimana mengatasi *data imbalances* untuk meningkatkan *model's accuracy* [4]. Dengan menerapkan teknik-teknik canggih ini, penelitian ini bertujuan untuk membangun *model churn prediction* yang dapat diterapkan secara *practical* dalam *business operation streaming music platform*. Nilai *machine learning* dalam sektor *telecomunication*, yang memiliki karakteristik serupa dengan industri *streaming music* [5].

Secara keseluruhan, paper ini memberikan kontribusi bagi bidang analitik pelanggan dengan menyediakan kerangka kerja untuk *churn prediction* yang dapat membantu layanan berbasis langganan seperti KKBOX meningkatkan *user retentions* dan mendorong *growth*. Temuan ini diharapkan relevan tidak hanya pada case seperti dataset KKBOX tetapi juga bagi sektor lain yang menyediakan *digital content*. Vafeiadis dkk. dan Tsai & Lu lebih lanjut menekankan *potential impact* dari model *prediction* semacam ini di berbagai industri [6,7].

1.1 Latar belakang

Masalah *churn* bukan hanya masalah *finances* saja, tetapi juga memiliki *impact* besar pada *growth* jangka panjang sebuah perusahaan. Penelitian dari Harvard Business School menunjukkan bahwa menjaga hanya 5% *customers retentions* yang ada dapat menghasilkan peningkatan *profitability* yang signifikan, hingga 25% hingga 95% [8]. Dengan demikian, mengurangi tingkat *churn* dapat meningkatkan *profitability* yang substansial bagi perusahaan. Pemilihan topik ini sangat relevan dalam konteks saat ini karena semakin banyak perusahaan yang berusaha memanfaatkan *data* untuk meningkatkan *strategy customer retentions* mereka. Namun, masih terdapat kesenjangan antara pemahaman konseptual tentang *churn* dan implementasi praktisnya. Oleh karena itu, *experiment* ini akan mengisi celah ini dengan mengembangkan model *churn prediction* yang dapat membantu perusahaan dalam memutuskan langkah-langkah strategis berdasarkan data pelanggan mereka.

Studi-studi terdahulu telah menunjukkan berbagai pendekatan untuk memodelkan dan memprediksi *churn.* Sebagai contoh, pada studi oleh Liao & Chen, mereka menggunakan pendekatan *machine learning* untuk mengidentifikasi faktor-faktor yang mempengaruhi *churn customers* [9]. Begitu juga, penelitian oleh Verbeke menunjukkan bahwa *predictive model* dapat secara signifikan meningkatkan akurasi dalam mengidentifikasi pelanggan yang berpotensi melakukan *churn* [10].

Dengan menggunakan algoritma *machine learning* dan *data analytics* pada *customers historical data*, ekperimen ini akan mengembangkan model *churn prediction* yang dapat membantu perusahaan untuk lebih efektif dalam menjaga dan meningkatkan basis *customers*. Pemilihan metode ini didasarkan pada kemampuan algoritma *machine learning* untuk menangani *volume* data yang besar dan kompleksitas dari masalah *classification churn*. Contohnya metede seperti *Ensemble methods*, seperti *bagging* dan *boosting*, dapat mengurangi bias dan variance, dan mengasilkan performances yang lebih baik pada *cases imbalanced data*. *Bagging* misalnya, dapat mengurangi *variance* dengan melakukan *averaging* dari beberapa *models*, yang dapat mencegah *overfitting* [11]. Boosting mengurangi *bias* dengan berfokus kepada data yang sulit di klasifikasikan, yang mana sering melibatkan *minority class samples* [12]. Model dari ensemble learning dipilih untuk jadi acuan

pada *experiment* ini.

1.2 Topik dan Batasannya

Competition yang semakin ketat dalam *industry streaming music* pada kasus ini KKBOX, telah menuntut platform untuk secara efektif mempertahankan customer mereka. Pada Data train dan test yang ada, terdapat class imbalances dimana *class churn* sangat rendah. Untuk mengatasi masalah ini, sangat penting untuk mengembangkan model churn prediction yang secara akurat melakukan prediction terhadap customers behavior –khususnya pada dataset yang mengalami *class imbalances* pada class target– agar dapat mengidentifikasi customers yang kemungkinan akan akan melakukan churn.

Paper ini bertujuan untuk membangun *model churn prediction* menggunakan pendekatan ensemble learning seperti XGBoost,LightBGM dan metode *traditional machine learning* lainnya pada *data customers* dari KKBOX. Dataset yang digunakan dalam penelitian ini mencakup table members, transaction, dan user logs. Input utama untuk model ini mencakup *feature* yang behubungan dengan *customers* seperti registration_date, membership_expired_date, music_activity_per_users, dan auto_renew. Output dari sistem ini akan berupa *prediction probability of churn* (berupa angka antara 0 dan 1). *Performance model* akan dievaluasi menggunakan log loss function, untuk memastikan keandalan dan efektivitasnya.

Beberapa *constraints* harus dipertimbangkan dalam penelitian ini. Dataset yang digunakan pada data train dan test yang tersedia di WSDM *competition* hanya terbatas pada periode April dan Mei 2017. Selain itu, karena keterbatasan *hardware*, hanya 1-5 juta *record* dari tabel user_logs yang digunakan dari total 300 juta *record* yang ada. Untuk memastikan perbandingan performa yang lebih adil dan objektif, 3 nilai hyperparameter digunakan per model.

1.3 Tujuan

Dalam penelitian ini, penulis membangun model churn prediction menggunakan *ensemble learning* dan metode lainnya untuk case *streaming music platform* KKBOX. Model ini bertujuan untuk memprediksi *customer behaviors*, apakah mereka akan *churn* atau tidak. Model XGBoost dan LightGBM diimplementasikan sebagai metode utama karena kemampuannya dalam menangani dataset yang besar dan kompleks serta memberikan hasil yang baik. Selain itu, kami juga akan mengevaluasi beberapa metode lain seperti Logistic Regression, SVM, dan LightLBM sebagai pembanding. *Performance* model-model ini akan dievaluasi menggunakan metric log loss, yang mana baik untuk *case imbalance data.*

Table 1: Keterkaitan tujuan riset paper, metrics, hipotesis

No	Tujuan	Metrics	Hipotesis
1	Membangun dan mengimplementasikan model *churn prediction* menggunakan Ensemble Learning dan metode lainnya untuk KKBOX.	Log Loss: Mengukur seberapa dekat prediciton dengan actual value, (dengan *higher penalties* untuk prediction yang salah) metode lainnya.	Model Ensemble Learning diharapkan menunjukkan kinerja terbaik dalam memprediksi churn dengan akurasi tertinggi dan log loss terendah dibandingkan dengan metode lain.
2	Mengidentifikasi fitur-fitur paling signifikan yang mempengaruhi churn pada model Ensemble Learning.	Importance scores dari features didalam model Ensemble Learning	Fitur yang digunakan dalam prediksi model diharapkan memiliki perilaku tertentu sesuai dengan metode (XGBoost dan LightGBM).

1.4 Organisasi Tulisan

Bagian awal dari penelitian ini akan membahas terkait Preprocessing dataset KKBOX, dan framework model. Pada bagian kedua, hasil eksperimen dari berbagai model yang digunakan akan dianalisis dan dibandingkan. Kesimpulan dari penelitian ini akan merangkum temuan model terbaik untuk kasus ini dan *future works* dalam konteks *streaming music platforms churn prediction*.