Essay: Artificial Intelligence and Its Applications

by Maria Ingold

Introduction

Ethical Bank (EthiBank) is a London-based fintech startup founded in 2019 providing ethical online financial services including loans, investments, accounts and software to businesses, individuals, and sole-traders. Although EthiBank has experienced growth, funding over 40 startups and serving over 100,000 UK customers, it faces competition from AI-driven contenders and macroeconomic conditions (Dietz et al., 2022). To combat this and increase return on investment, Artificial Intelligence (AI) can enable churn reduction, automation, and revenue optimisation while still upholding EthiBank's commitment to transparency, ethics, and industry standards (Google Cloud, N.D.). This report identifies three AI applications that follow the Cross Industry Standard Process for Data Mining: predicting client churn, assessing startup potential, and forecasting revenue (Niakšu, 2015).

Predicting Client Churn

Problem

EthiBank currently suffers from high churn, resulting in revenue loss, with customer acquisition costing up to five times retention (de Lima Lemos et al., 2022). As an online-only company, EthiBank has rich usage, transaction, and engagement data, including account closures but lacks demographic data. Fortunately, research shows core banking data alone can effectively predict churn using machine learning (ML) (Rahman and Kumar, 2020; de Lima Lemos et al., 2022).

Solution

Supervised learning algorithms like random forest predict churn using historical labelled data (Bell, 2020; Chui et al., 2020; Russell & Norvig, 2021). In both examples, random forest outperformed other algorithms, including decision trees, k-nearest neighbours, and support vector machines (SVM), with higher accuracy (Rahman & Kumar, 2020; de Lima Lemos et al., 2022).

Because churn datasets are often imbalanced, affecting accuracy, pre-processing is required. A large Brazilian banking dataset of nearly 10 million customers created a balanced subset of 500,000 current accounts with 50% churners (de Lima Lemos et al., 2022). Out of 35 attributes, random forest predicted those with higher loans were most likely to stay, followed by more transactions and more products. In conclusion, a stronger customer relationship reduced churn.

By comparison, Rahman & Kumar (2020) analysed an imbalanced Kaggle dataset of 10,000 bank clients, with 20% negative class churners. Preprocessing reduced the 13-attribute dimensionality to 9, speeding up training and simplifying the model (Neal, 2019; Rahman & Kumar, 2020). Accuracy was evaluated using 10-fold cross-validation and confusion matrices which examine true and false positives and negatives. Oversampling—randomly duplicating the minority negative class to balance the majority—significantly increased random forest's accuracy to 95.74%—outperforming the other models. However, further feature reduction decreased accuracy, likely because the oversampling handled the remaining 9 features well. As loans were excluded, more products per customer was the top retention indicator.

At 10% churn, a balanced EthiBank training set would contain 10,000 past churners and 10,000 random non-churners from its 100,000 customers. At a minimum, data should cover loan value, number of transactions, number of products and churn status.

Considerations

The small balanced training dataset risks bias from underrepresented churners (de Lima Lemos et al., 2022). While accurate, random forest models are opaque, making bias hard to detect, and more interpretable models like decision trees underperformed. Post-hoc explainability methods like Local Interpretable Model-Agnostic Explanations (LIME) can clarify opaque models (Belle & Papantonis, 2021). However, customer feedback may provide more insight than explainability metrics alone. Rahman & Kumar (2020) and de Lima Lemos (2022) demonstrated that customers with higher loans and more products are less likely to churn, indicating relationship building is key for retention. Therefore, combining AI with customer interviews could enhance analysis and reduce churn.

Assessing Startup Potential

Problem

During the COVID-19 pandemic, startups helped reduce unemployment and poverty (Bangdiwala et al., 2022). In 2020, EthiBank launched EthiFund, seed funding 40 startups. With approximately 20% of UK startups failing in year one and 60% by year three, due diligence is required, but manual investigation limits scalability (Horne, 2022). All EthiFund startups must use EthiBank services—free of charge. However, 40 records are insufficient for prediction.

Solution

Corporate models cannot predict startup success. Therefore, Bangdiwala et al. (2022) preprocessed Crunchbase into 48,130 startups with 18 features like category, funding, region, funding, founding, and acquisition status. Total funding did not guarantee success, but high cash burn led to failure. Testing decision trees, random forest, gradient boosted trees, logistic regression, and neural networks all yielded over 91.75% accuracy. However, decision trees, gradient boosted trees, and neural networks achieved higher area under curve (AUC) scores at 0.88-0.89—where 1 perfectly predicts and 0.5 is random chance.

Żbikowski & Antosiuk (2021) trained models on 213,171 Crunchbase companies, reserving 10,000 for testing. By selecting only nine pre-operation features like country, region, industry, gender, and education, they avoided “look-ahead bias”. However, the 1995-2015 timeframe presented biases, as early data underrepresented failure and some companies are now inactive. Key predictors were geography and sector. London led European cities, and the top industries were manufacturing, healthcare, software, mobile and financial—albeit all under 20% success. With success defined as initial public offering (IPO), acquisition, or completing Series B while operating, the 12.2% positive (17% for UK) class imbalance challenged models. With cross-validation reducing overfitting and selection bias, XGBoost, which boosts decision trees, outperformed logistic regression and SVM with 85% accuracy. Yet low precision and recall indicated difficulty predicting less frequent success patterns.

With internal data limited to just 40 startups, EthiFund would need large external datasets like Crunchbase, with over 8,000 UK startups and thousands more globally, to sufficiently train models (Crunchbase, 2023). If EthiFund supplements Crunchbase data with its own, the data should align temporally to avoid bias.

Considerations

Bangdiwala et al. (2022) improved on corporate models by creating a startup model, with decision trees best balancing accuracy, AOC and explainability. Limiting data to pre-funding features avoids look-ahead bias, but risks regional, sector, gender, and recency biases (Żbikowski & Antosiuk, 2021). Though accurate, opaque models like XGBoost require post-hoc explainability (Belle & Papantonis, 2021). A two-model approach enables pre-funding evaluation, and opted-in incorporation of indicators like cash-burn into EthiBank's services can optimise success predictions as startups scale.

Forecasting Revenue

Problem

With declining year-over-year revenue growth, EthiBank want to evaluate AI to improve on manual, error-prone spreadsheet forecasting (Lei and Cailan, 2021). EthiBank has four years of accounting data. EthiFund startups have 1-5 years, while small and medium-sized enterprises (SME) and sole-traders may have more history.

Solution

Given the limited recent data, time series modelling may not capture seasonality, and revenue forecasting typically uses regression rather than time-series (Pavlyshenko, 2018; Lei & Cailan, 2021).

However, Lei & Cailan (2021) found limitations with regression for enterprise revenue forecasting, including explaining operations and potentially losing information during processing. Analysing top performer public data from 2015-2020, they trained models on 80 firms and tested on 20, examining 18 standard accounting features like revenue, growth, debt, assets, profit, cost, and cashflow. They standardised the data to prevent errors from inconsistent levels across features. Despite high feature correlation, models coped. With the small data size, deep learning was excluded. Surpassing random forest and gradient boosted regression tree (GBRT), support vector machines had the lowest mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE), making it potentially ideal for predicting limited data.

Kureljusic & Reisch (2022) found ML matched or exceeded European financial analyst revenue forecasting accuracy. Using 300 companies' public financial data from 2010-2019, augmented with macroeconomic indicators like gross domestic product (GDP), inflation, and unemployment, tests found an 80/20 train/test split ideal. However, while essential, preprocessing introduced risks: incorrect coding from data vectorisation to convert non-numbers into 0 or 1, misestimated missing values, misinterpreting scaling to standardise numbers, and losing information when extracting features to reduce dimensions and compute time. Measuring prediction accuracy across four quality metrics, random forest performed best out of eight models—on par for three and significantly better for mean absolute percentage error (MAPE) versus analysts.

Given data privacy needs and potentially small datasets, EthiBank and EthiFund could train and test on public data then beta test with opted-in users, potentially offering premium access in exchange. While Lei & Cailan (2021) shared 18 features without denoting importance, Kureljusic & Reisch (2022) catalogued feature relevance by model. The top indicator by far is revenue. For random forest, the next highest are retained earnings, accounts payable, and accounts receivable—all available in standard accounting data.

Considerations

Both analyses have limitations. Lei & Cailan (2021) only analysed enterprises, while Kureljusic & Reisch (2022) focused on European Union (EU) blue-chip indices, so findings may not generalise to SMEs, startups, and sole proprietors. The 2015-2020 and 2010-2019 timeframes respectively, mostly reflected periods of economic stability. Additionally, SVM and random forest are opaque models requiring post-hoc explainability (Belle & Papantonis, 2021). With further testing, the forecasting tools could extend to EthiBank's startups, SMEs, and sole-traders—potentially improving retention and return on investment (ROI).

Conclusion

In conclusion, thoughtfully implementing AI applications in client churn prediction, startup evaluation, and revenue forecasting could help EthiBank scale operations to compete with AI-driven fintech rivals while maintaining ethics, transparency, and adherence to global standards. However, small or imbalanced training datasets pose risks of bias. While potentially most accurate, opaque models like random forests, require explainability methods to ensure transparency, and models like decision trees may, in some cases, offer a better balance of accuracy and inherent interpretability. A phased approach is recommended, starting with small beta tests of AI tools verified by customer feedback. With iterative improvements guided by user insights and established AI practices like cross-validation and oversampling, EthiBank can selectively roll out ethical AI systems that augment human analysis while optimising processes for profitability.

References

Bangdiwala, M. et al. (2022) Predicting Success Rate of Startups using Machine Learning Algorithms, in *2022 2nd Asian Conference on Innovation in Technology (ASIACON)*. DOI: https://doi.org/10.1109/ASIANCON55314.2022.9908921.

Bell, J. (2020) *Machine Learning: Hands-On for Developers and Technical Professionals*. Wiley. DOI: https://doi.org/10.1002/9781119642183.

Belle, V. & Papantonis, I. (2021) Principles and Practice of Explainable Machine Learning, *Frontiers in Big Data*. Frontiers Media S.A. DOI: https://doi.org/10.3389/fdata.2021.688969.

Chui, M., Kamalnath, V. & McCarthy, B. (2020) *An executive's guide to AI*. Available from: https://www.mckinsey.com/capabilities/quantumblack/our-insights/an-executives-guide-to-ai [Accessed 14 August 2023].

Crunchbase (2023) *United Kingdom Startups*. Available from: https://www.crunchbase.com/hub/united-kingdom-startups [Accessed 8 October 2023].

Dietz, M. et al. (2022) *Banking on a sustainable path Global Banking Annual Review 2022*. Available from: https://www.mckinsey.com/~/media/mckinsey/industries/financial%20services/our%20insights/global%20banking%20annual%20review%202022%20banking%20on%20a%20sustainable%20path/global%20banking%20annual%20review%202022%20banking%20on%20a%20sustainable%20path.pdf?shouldIndex=false [Accessed 7 October 2023].

Google Cloud (N.D.) *What is artificial intelligence (AI) in finance?* Available from: https://cloud.google.com/discover/finance-ai [Accessed 7 October 2023].

Horne, B. (2022) *How Many Businesses Fail in the First Year in the UK?* Available from: https://www.nerdwallet.com/uk/business/start-up-failure-statistics/#:~:text=In%20the%20UK%2C%20according%20to,within%20the%20first%20three%20years. [Accessed 8 October 2023].

Kureljusic, M. & Reisch, L. (2022) Revenue forecasting for European capital market-oriented firms: A comparative prediction study between financial analysts and machine learning models, *Corporate Ownership and Control 19*(2): 159–178. DOI: https://doi.org/10.22495/cocv19i2art13.

Lei, H. & Cailan, H. (2021) Comparison of Multiple Machine Learning Models Based on Enterprise Revenue Forecasting, in *2021 Asia-Pacific Conference on*

Communications Technology and Computer Science (ACCTCS). DOI: https://doi.org/10.1109/ACCTCS52002.2021.00077.

de Lima Lemos, R.A., Silva, T.C. & Tabak, B.M. (2022) Propension to customer churn in a financial institution: a machine learning approach, *Neural Computing and Applications 34*: 11751–11768. DOI: https://doi.org/10.1007/s00521-022-07067-x.

Neal, B. (2019) *On the Bias-Variance Tradeoff : Textbooks Need an Update*. MSc thesis. Université de Montréal. DOI: https://doi.org/https://doi.org/10.48550/arXiv.1912.08286.

Niakšu, O. (2015) *CRISP Data Mining Methodology Extension for Medical Domain*, *Baltic J Modern Computing*.

Pavlyshenko, B.M. (2018) Machine-Learning Models for Sales Time Series Forecasting † 21–25. DOI: https://doi.org/10.3390/data4010015.

Rahman, M. & Kumar, V. (2020) Machine Learning Based Customer Churn Prediction In Banking, *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) 1196*–1201. DOI: https://doi.org/10.1109/ICECA49313.2020.9297529.

Russell, S. & Norvig, P. (2021) *Artificial Intelligence: A Modern Approach, Global Edition*. 4th ed. Pearson Education, Limited.

Żbikowski, K. & Antosiuk, P. (2021) A machine learning, bias-free approach for predicting business success using Crunchbase data, *Information Processing and Management 58*: 102555. DOI: https://doi.org/10.1016/j.ipm.2021.102555.