

Infinite Investment Systems Ltd.

**Exploring Applications of A.I in Wealth Management: Customer Churn**

*Submitted By: Nathan Chu, Cherry Yang,*

*Matthew Yang, Cloris Zhang*

1. **Introduction**

Our team was tasked with addressing the challenge outlined in the case study: to predict and analyze customer churn for a nationally renowned wealth management platform using machine learning. We recognize the importance of customer retention in sustaining and enhancing Infinite Investment Systems’ continued growth and success. Customer churn prediction means knowing which customers are likely to leave or unsubscribe from Infinite Investment Systems’ services. Thus, the primary objective is to devise algorithms that can help us predict and analyze which customers are likely to churn so that we can enable proactive marketing strategies to retain valuable clientele.

By leveraging the provided historical data from the company's platform, supplemented by insights gained from external research from the wealth management industry and machine learning techniques, our team was able to build multiple classification models, which were then evaluated against each other for the best possible model to predict customer churn. Our initial research uncovered possible data manipulations we could perform such as normalizing numerical data, encoding categorical data and effective models to use for classification such as XGBoost and Random Forest. Our final model drew upon a comprehensive understanding of customer behaviours, preferences, and industry trends to accurately forecast which current customers are most likely to churn.

Our investigation included the following main stages:

1. Research
2. Data Cleaning
3. Exploratory Data Analysis (EDA)
4. Data Preprocessing
5. Model Training & Refinement
6. Results Analysis & Marketing

Our team outlined a strategic plan beforehand, assigning responsibilities to team members and establishing clear timelines for each stage of the project. This comprehensive approach encompassed tasks ranging from data cleaning and exploratory data analysis to model development and refinement, ensuring a structured path towards completion.

1. **Methodology**

**Data Cleaning**

To ensure the effectiveness of our machine learning model in predicting customer churn for the wealth management platform, we performed a comprehensive data cleaning process. This critical phase involved several key steps aimed at preparing the dataset for analysis and modeling.

Firstly, we identified and dropped unuseful columns from the dataset, such as columns with the same values for all entries or columns with too many empty values. This step was crucial to prevent the model from being trained on irrelevant or redundant data. Each team member contributed to this effort by selecting and dropping columns within their assigned range, resulting in a streamlined dataset ready for further processing.

Next, we addressed boolean columns present in the dataset. These columns represented binary values, such as True/False, which needed to be standardized to numeric values for modeling purposes. Through a mapping function, we converted boolean values to 0s and 1s, ensuring consistency and compatibility with the model.

Next, we parsed datetime columns to standardize their values to a numeric range between 0 and 1. This step involved converting datetime values to integers, normalizing them within the range of the minimum and maximum values, and replacing missing values with the minimum date to maintain consistency.

Finally, we numerized type columns within the dataset to represent categorical variables in a numeric format suitable for modeling. This involved parsing and encoding categorical values within specific columns, such as risk tolerance and investment objective, into numeric representations based on predefined criteria.

Throughout the data cleaning process, we meticulously handled missing values by replacing them with appropriate placeholders or filling them with default values like -999 as necessary. This ensured that the dataset remained robust and complete, ready for subsequent analysis and modeling tasks.

By systematically addressing data cleaning tasks and collaborating effectively as a team, we were able to transform the raw dataset into a clean, standardized format optimized for machine learning analysis. This laid a solid foundation for our subsequent efforts in model development and predictive analytics aimed at accurately predicting customer churn for the wealth management platform.

**Exploratory Data Analysis (EDA)**

During the exploratory data analysis (EDA) phase, our team meticulously examined the dataset performing statistical feature selection to determine which columns would be included in the final dataframe for model training and testing. This process involved several steps to understand the data's composition and identify relevant variables for predicting customer churn.

Initially, we partitioned the dataset into subsets, ranging from columns A to AZ and BA to CZ, for a more focused analysis. For each subset, we performed preliminary data exploration, generating summary statistics and visualizations to gain insights into the distribution and characteristics of the data.

One crucial aspect of the EDA was identifying categorical variables within the dataset. These variables represent qualitative data and play a significant role in predicting churn. We evaluated the unique values and frequencies of each categorical variable, considering their relevance to the prediction task. Variables with a small number of unique values and meaningful categories were retained for further analysis, while others were discarded if they were deemed irrelevant or redundant.

Similarly, we identified numerical variables within the dataset, which provide quantitative information that can be leveraged for predictive modeling. For these variables, we examined their descriptive statistics and distributions between churn and non-churn instances. Variables with a wide range of values and clear distinctions between churn and non-churn instances were considered valuable for the prediction task and retained for further analysis.

Furthermore, we conducted a correlation analysis to identify relationships between variables and the target variable (churn). Heatmaps were generated to visualize the correlation matrix, highlighting variables with the strongest correlations to churn. This analysis helped prioritize variables with the highest predictive power for inclusion in the final dataframe.

Based on the insights gained from the EDA, we curated the final dataframe for model training and testing. This dataframe included a subset of both numerical and categorical columns deemed most relevant for predicting customer churn. Columns were selected based on their potential predictive power, relevance to the prediction task, and absence of redundant or irrelevant information.

By systematically conducting an exploratory data analysis and making informed decisions about variable selection through statistical feature selection, we ensured that the training dataset provided a solid foundation for building and evaluating predictive models to address the challenge of customer churn prediction in the wealth management platform.

**Model Development**

In our process of implementing and choosing tools for model development and tuning, we employed various techniques and tools to build and evaluate predictive models for predicting customer churn in the wealth management platform.

Firstly, we initialized the features (X) and target variable (y) from the final dataframe, setting up the dataset for model training. We then defined a collection of six classification models to evaluate their performance and compare them against each other: K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Gradient Boosting, LightGBM, and XtremeGBRFM classifiers. Additionally, we utilized techniques like threading to expedite the cross-validation process.

Using stratified k-fold cross-validation with 10 folds to account for imbalance in the target variable, we evaluated each model's performance based on accuracy and F1 score metrics. We decided to use 10 folds for the k-fold cross-validation as it was noted to be industry standard and common practice for datasets of this relative size from many different sources. This allowed us to assess the models' ability to generalize to unseen data and identify the most effective model for predicting churn. We stored the results of each model's performance, including average accuracy, standard deviation, and average F1 score, for comparison.

| **Model** | **Mean Accuracy** | **Std. dev of Accuracy** | **F1 Score** |
| --- | --- | --- | --- |
| KNN | 0.93642 | 0.000889 | 0.960121 |
| CART | 0.952403 | 0.00071 | 0.970096 |
| RF | 0.962952 | 0.000747 | 0.976498 |
| GB | 0.967741 | 0.000721 | 0.979416 |
| LightGBM | 0.968266 | 0.000798 | 0.979758 |
| XtremeGBRFM | 0.967698 | 0.000768 | 0.979388 |

Upon evaluating the models, we identified the model with the highest F1 score as the most suitable for predicting churn and selected it as the final model for deployment. This is because we felt that the F1 score was one of the more important metrics for a binary classification task especially when our churn and no churn counts were quite heavily imbalanced. Using the F1 score provided better insight into the precision and recall results of the models and helped with our decision on which model to choose. Relying on accuracy alone would have been a poor indicator of the model performance as most predictions would have landed in the majority class which was churn. We then trained the chosen model, Light Gradient Boosting Random Forest Classification, on the entire dataset to leverage all available data for optimal performance.

Following model training, we prepared the given test dataset for prediction by applying the same preprocessing steps used for the training data, including handling missing values, encoding categorical variables, parsing datetime columns, and numerical type conversion. Once the test dataset was preprocessed, we used the trained model to make predictions on the test data.

Finally, we saved the predictions to a CSV file for submission and further analysis. Throughout the process, we maintained a systematic approach, leveraging a combination of tools and techniques to develop and tune the predictive model effectively, ultimately enabling accurate predictions of customer churn in the wealth management platform.

1. **Analysis & Marketing**

*Please refer to our notebook for the results of our analysis (4.0 Marketing)*

After looking through our model’s predictions, one thing we noticed was that within the type\_code column, the TFSA and RRSP type codes corresponded more towards no churn making up a combined 61% of all type\_codes which don’t churn. On the other hand, cash, cash sweep, and margin account type codes were much more likely to churn with a combined 62% distribution over all type\_codes which do churn. From these observations, we came up with the idea of launching educational campaigns on the benefits and advantages of opening up and retaining TFSA and RRSP accounts. These campaigns would specifically highlight the long-term positive value that these accounts can provide in terms of financial well-being, tax advantages, and retirement planning benefits. With a greater number of clients having opened a TFSA or RRSP account, it is more likely that they will not churn based on predictions made by our model. To add to the campaigns we could also offer promotions to incentivize opening up these types of accounts or transferring from cash, cash sweep, and margin accounts to a TFSA or RRSP account. These promotions could include better interest rates or a cash bonus for creating the account. On the other hand, we could also look at offering more tailored guidance on how to optimize cash, cash sweep and margin accounts, so that people are more comfortable with maintaining their accounts and using hedging strategies to meet their financial objectives leading to less churn.

Another common theme we saw among people who were churning was the fact that they had blank plan\_effective\_dates and plan\_end dates. This made us realize that we could target people who aren’t currently on any plans and focus our marketing towards them. Similar to how we would draw more people to TFSA and RRSP accounts, we could use special offers and promotions to draw people into using plans which should lead to fewer churns among existing clients. The appeal of plans can be enhanced by increasing advertising and emphasizing the positive impacts that clients could benefit from with the use of the plans.

One final observation we noticed across all clients which was that clients who had more empty entries were more likely to churn. We thought that this could mean that not enough information was gathered about the client and well-informed decisions could not be made for the client to help improve their experience with the company. To combat this problem, we came up with the idea of sending out surveys and developing an appointment booking system to allow clients to interact and engage with the company more collaboratively. This would allow for more data and important information to be gathered while informing the client of ways to enhance their experience with the company. All of this would happen with their consent because of the medium in which the information is collected ultimately resulting in better long-term business-to-customer relationships and a lower churn rate according to our predictions.

1. **Reflection**

Our team executed a well-structured and collaborative approach to address the challenge of predicting customer churn for the wealth management platform. One of the key strengths of our methodology was the planning and division of responsibilities, ensuring that each team member contributed effectively to the project's success. Additionally, our approach to data cleaning was systematic and thorough, resulting in a clean and standardized dataset optimized for machine learning analysis. We successfully handled missing values, standardized data types, and removed irrelevant columns. This attention to data quality enhanced the reliability and accuracy of our predictive models, contributing to the robustness of our findings. Furthermore, our utilization of exploratory data analysis (EDA) provided valuable insights into the dataset's composition and characteristics. By examining correlations, distributions, and trends within the data, we gained a deeper understanding of potential predictors of customer churn.

There remain aspects where our approach could be further improved. Firstly, while conducting research within the wealth management industry, investing more time in researching industry trends could have provided additional insights and enhanced the relevance of our predictive models. Bringing in external data sources to supplement the existing dataset could also have improved model accuracy and predictive performance by capturing broader consumer market trends and customer behaviors. Moreover, performing exploratory data analysis (EDA) before data cleaning could have provided a better understanding of what the dataset contained and informed more targeted data cleaning decisions. This sequential approach would have allowed us to identify important variables and relationships earlier in the process, potentially improving the effectiveness of our predictive models. Lastly, if we had more time, conducting further investigation into the impact of including or excluding certain columns on model accuracy and F1 scores would have been beneficial. Testing and comparing different model configurations could have provided insights into the relative importance of specific features and informed more informed decision-making regarding variable selection.

Although initially, over half our team members were entirely new to machine learning, we were able to learn along the way and produce a model capable of predicting customer churn. Our approach demonstrated several strengths in planning, data cleaning, exploratory analysis and there are opportunities for improvement in research, data exploration, and model evaluation. By incorporating these into future projects, we hope to enhance the effectiveness and accuracy of our predictive models.

1. **Resources**

Brownlee, Jason. "Random Forest Ensembles with XGBoost." Machine Learning Mastery, 10 Aug. 2021, [https://machinelearningmastery.com/random-forest-ensembles-with-xgboost](https://machinelearningmastery.com/random-forest-ensembles-with-xgboost/).

Brownlee, Jason. "K-Fold Cross-Validation." Machine Learning Mastery, 19 Oct. 2021, <https://machinelearningmastery.com/k-fold-cross-validation/>.

Ereken, Simge. "Churn Prediction using Machine Learning." Kaggle, 16 Nov. 2021, <https://www.kaggle.com/code/simgeerek/churn-prediction-using-machine-learning#2--Data-Preprocessing>.

"Cross Validation in Machine Learning: What Is It & How Does It Work?" G2, <https://learn.g2.com/cross-validation>.

Data Science Stack Exchange. "Effects that Empty Cells have in Unsupervised Machine Learning." Stack Exchange, 2021, <https://datascience.stackexchange.com/questions/67639/effects-that-empty-cells-have-in-unsupervised-machine-learning>.