Brokerage Longevity Analytics in the Finance Market

Final Project

Talha Ahmed	13903620
Joel Larsen	14326856
Chi Vi Nguyen	13592629
Zhuoqian Pan	13647045
Alexander Zervos	13934841
Xingxing Zhang	13653530

41004 AI/Analytics Capstone Project

Table of Contents

1 | Executive Summary

Business Problem:

Maintaining client engagement and longevity is a significant challenge for brokers in financial markets. SIGMA Trading Management seeks to assist clients in maximising their performance and potential but faces hurdles in building enduring relationships due to impersonal transactions and limited engagement opportunities. Sevenfold Consultants, as consultants to SIGMA, identified two main challenges: understanding and optimising customer engagement and retention strategies, and utilising data-driven insights effectively to achieve these objectives. The multifaceted nature of obstacles faced by brokers underscores the need for a deep understanding of customer behaviour and factors influencing their decisions to enhance services and maximise client retention. In particular, as a means of determining where to prioritise their resources, SIGMA is primarily interested in understanding what kinds of traders tend to use the service for long periods of time, the patterns they exhibit, and how they can identify clients by their likely longevity.

Findings:

Analysis conducted by Sevenfold Consultants revealed several key insights into factors affecting trader longevity. Notably, higher trading frequency was inversely related to longevity, with traders engaging in fewer daily trades exhibiting longer-term engagement. Profitability analysis indicated that moderate, consistent performance is more sustainable than extreme profitability. Stability, prudent risk management, and strategic trading practices were found to correlate with longer trading lifespans. Additionally, the provision of high-quality resources and support, lower commissions, and effective risk management practices were associated with increased longevity. Finally, while credit is, unexpectedly, not a good indicator, it may still be useful as an incentive to traders, and a means of converting lower longevity clients.

Modelling:

Distinct differences in model performance were observed between the bottom 90% and top 10% of client accounts, suggesting that trading behaviours significantly influence client retention. While the XGBoost model showed superior performance for the bottom 90% of clients, predicting longevity for high-value clients proved more challenging, indicating the influence of complex factors not fully captured by available data or current modelling techniques.

Recommendations:

Based on findings, recommendations include:

- Focusing marketing efforts on key categorical variables such as algorithmic or mobile traders and targeting markets with higher average longevity tendencies.
- Monitoring client activity based on key variables identified and promoting conservative trading practices using marketing, recommendations, and incentives.
- Providing credit as an incentive for low longevity clients to adopt conservative positions and promoting effective risk management practices like setting take profit and stop loss limits.
- Crafting tailored approaches for managing extremely high longevity traders, recognizing their differing behaviours and priorities.

Implementing these recommendations can help SIGMA optimise marketing strategies, risk management, and client services to build enduring relationships with traders, thus addressing the critical business problem of client engagement and longevity in financial markets.

2 | Problem

Maintaining client engagement and longevity poses a significant challenge for brokers across financial markets. With differing priorities, impersonal transactions, and limited opportunities for engagement, building meaningful and enduring relationships with clients becomes arduous. SIGMA Trading Management recognise this challenge is pivotal to their mission of assisting clients in maximising their performance and potential in the financial markets.

As consultants to SIGMA, we at Sevenfold Consultants view the business problem as twofold: Firstly, brokers need to comprehend and optimise customer engagement and retention strategies. Secondly, there's an imperative to utilise data-driven insights effectively to achieve this objective.

The complexity and significance of this challenge is underscored by the multifaceted nature of hurdles faced by brokers in retaining clients. These obstacles encompass diverse aspects such as marketing campaigns, risk management strategies, budget allocation, and daily operational decisions. Without a deep understanding of customer behaviour and the underlying factors influencing their decisions, brokers risk missing out on crucial opportunities to enhance their services and maximise client retention.

In particular, as a means of determining where to prioritise their resources, SIGMA is primarily interested in understanding what kinds of traders tend to use the service for long periods of time, the patterns they exhibit, and how they can identify clients by their likely longevity.

At Sevenfold Consultants, we recognized the elucidatory nature of data in navigating this landscape and are committed to providing actionable insights to address this critical business problem. Leveraging our expertise, we navigated the highly quantitative financial brokerage landscape to determine factors affecting client longevity. Through meticulous data cleaning, processing, and transformation, we prepared the data to facilitate the analysis of client longevity. This involved categorising clients by their trading style, approach, and habits, yielding meaningful insights into their propensity to trade and the duration of their trading relationship. The analysis revealed potential patterns between longevity and key features, driving our feature engineering and modelling efforts.

3 | Data Exploration

The dataset provided by SIGMA consists primarily of five CSV files, including login.csv, trades.csv, symbol.csv, reason.csv and daily_report.csv. Moreover, the client also provided the daily_chart folder to convert other currencies to USD for consistency. Login.csv describes the general information about the trading accounts, while trades.csv describes their detailed trading history. The two columns, symbol and reason, in trades.csv are encoded, and their details are described in symbol.csv and reason.csv. Finally, the daily_report.csv provides a daily report of each trading account regarding their balance and profit details.

3.1 | Initial Data Attributes

Variable	Type	Level	Non-null Entries	Description
login	Integer	Nominal	40512	Account number, unique identifier of clients.
country	String	Nominal	40505	Country where the client is located, country name or Alpha-2 code.
account_currency	String	Nominal	40512	Currency of the account.
reg_date	Integer	Interval	40512	Accounts registration date in unix epoch time.

Table 1. login.csv: Information about trading accounts

Variable	Type	Level	Non-null Entries	Description
ticket	Integer	Nominal	4521777	Order number, unique identifier of trades.
login	Integer	Nominal	4521777	Trading account.
symbol	String	Nominal	4521777	Security/Product traded.
cmd	Integer	Nominal	4521777	Direction of the trade, 0=buy, 1=sell.
volume	Float	Ratio	4521777	Volume in lot.
open_time	Integer	Interval	4521777	Trade open time in unix epoch time.
open_price	Float	Ratio	4521777	Security price at which the trade is opened.
close_time	Integer	Interval	4521777	Trade close time in unix epoch time.
close_price	Float	Ratio	4521777	Security price at which the trade is closed.
tp	Float	Ratio	4521777	Take-profit, 0 if nil.
sl	Float	Ratio	4521777	Stop-loss, 0 if nil.
reason	Integer	Nominal	4521777	Platform/Method used to place the order.
commission	Float	Ratio	4521777	Commission charged by the broker, quoted

				in account currency.
swaps	Float	Ratio	4521777	Total overnight interest, quoted in account currency.
profit	Float	Ratio	4521777	Total profit/loss of the trade, quoted in account currency.
volume_usd	Float	Ratio	4521777	Total notional volume traded, quoted in USD.

Table 2. trades.csv: Trading history of the accounts

Variable	Type	Level	Non-null Entries	Description
symbol	String	Nominal	113	Symbol name.
description	String	Nominal	113	Description of the symbol.
type	String	Nominal	113	Symbol type, eg. Forex, Metal, Index etc.

Table 3. symbol.csv: Information about symbols traded

Variable	Type	Level	Non-null Entries	Description
code	Integer	Nominal	8	Code of reason as in trades.csv.
reason	String	Nominal	8	Platform/Method by which an order can be placed.

Table 4. reason.csv: Mapping of reason code

Variable	Type	Level	Non-null Entries	Description
login	Integer	Nominal	8664161	Trading account.
record_time	Object	Interval	8664161	Date for which the report is generated.
net_deposit	Float	Ratio	8664161	Total net deposit of the account on the day.
balance	Float	Ratio	8664161	Account balance at end of day, excluding credit and floating profit/loss.
equity	Float	Ratio	8664161	Account balance at end of day, including credit and floating profit/loss.
credit	Float	Ratio	8664161	Temporary funds given to the client as an incentive.
profit_closed	Float	Ratio	8664161	Total profit/loss of trades closed on the day..
profit_floating	Float	Ratio	8664161	Total profit/loss of trades that are still open at end of day.

margin	Float	Ratio	8664161	Funds that are withheld for open trades.

Table 5. daily_report.csv: End of day report of a trading account

3.2 | Preliminary Data Pre-processing

3.2.1 | CSV Files Considered For Longevity Analysis

1. **login.csv**: This table (see Figure 1) contains unique account ids, countries of origin, the currency that the accounts trades in and the registration date of the accounts. This csv acts as the master in which all other relevant files will eventually merge into to create a comprehensive dataset for longevity analysis and predictive modelling.

	login	country	account_currency	reg_date
0	457547	Romania	EUR	1614212132
1	474589	CA	CAD	1609987442
2	504321	CA	CAD	1602642710
3	504322	CA	USD	1602736545
4	504326	CA	USD	1603093152

Figure 1. Head of login.csv

2. **trades.csv:** This table (see Figure 2) contains largely numerical data regarding all trades made on the brokerage platform. This csv will be merged into login, where insightful trading metrics will be derived for each unique account.

	ticket	login	symbol	cmd	volume	open_time	open_price	close_time	close_price	tp	sl	reason	commission	swaps	profit	volume_usd
0	68880703	7062462	XAUUSD	0	0.01	1707843941	1991.35000	1708013114	2003.01000	0.00000	0.00000	1	0.00	-1.12	10.83	3994.36
1	68880910	7062462	XAUUSD	0	0.02	1707844129	1990.30000	1708013110	2002.76000	0.00000	0.00000	1	0.00	-2.25	23.16	7986.12
2	68120690	813125	US2000	0	1.00	1706083005	1983.22000	1706114503	1996.68000	2012.34000	1980.03000	1	0.00	0.00	105.42	39799.00
3	68169249	813125	US2000	0	2.00	1706139371	1960.62000	1706200143	1991.51000	2331.85000	1958.94000	1	0.00	-7.21	485.26	79042.60
4	68186877	88945036	NZDUSD	0	0.13	1706169600	0.61062	1706197106	0.61317	0.61312	0.55059	1	-0.91	0.00	33.15	15909.27

Figure 2. Head of trade.csv

3. **daily_chart directory:** This is a directory of csv files for every currency that is traded on the brokerage platform, with data on the exchange rate to USD from 2021-2024. This directory will be used to convert all currencies in trades.csv to USD to ensure consistency and validity for analysis.
4. **reason.csv:** This table (see Figure 3) translates the encoded ‘reason’ variable in trades.csv to its respective trading method. This is useful for analysis and visualisation.

	code	reason
0	0	Client
1	1	Expert
2	2	Dealer
3	3	Signal
4	4	Gateway
5	5	Mobile
6	6	Web
7	7	API

Figure 3. reason.csv

5. **daily_report.csv:** This table (see Figure 4) provides a daily report for each account in 2023, including information about deposits, equity, daily profits, and more. Some of this data is already captured in trades.csv. This dataset will also be merged into the master for the generation of extra features.

	login	record_time	net_deposit	balance	equity	credit	profit_closed	profit_floating	margin
0	457547	2023-01-01	0.0	0.00	0.00	0.0	0.0	0.0	0.0
1	474589	2023-01-01	0.0	0.56	0.56	0.0	0.0	0.0	0.0
2	504321	2023-01-01	0.0	2.03	2.03	0.0	0.0	0.0	0.0
3	504322	2023-01-01	0.0	0.51	0.51	0.0	0.0	0.0	0.0
4	504326	2023-01-01	0.0	0.01	0.01	0.0	0.0	0.0	0.0

Figure 4. Head of daily_report.csv

3.2.2 | Variable Level Pre-processing

1. **Country Mapping:** The country variable in login.csv describes the country of origin for a given account in either name or Alpha-2 code. After visualisation of the unique entries for this variable, it was noted that some countries had representations in both formats, which is an issue for both analysis and modelling.

```
login_dataframe['country'].unique()
✓ 0.0s
array(['Romania', 'CA', 'CI', 'PK', 'AF', 'KE', 'GB', 'Canada', 'MQ',
       'Slovakia', 'Czech Republic', 'Pakistan', 'Germany', 'France',
```

Figure 5. Top of unique entries for 'country' in login.csv

Figure 5 shows labels 'CA' and 'Canada' which both represent Canada, as well as 'PK' and 'Pakistan' which represent Pakistan. To remove this issue, Alpha-2 code labels were mapped to their proper country names in a separate JSON file (see Figure 6) resulting in a decrease of unique entries in 'country' from 384 to 238.

```
{} country_mapping.json > ...
{
    "CA": "Canada",
    "AU": "Australia",
    "CI": "Cote D'ivoire",
    "PK": "Pakistan",
    "AF": "Afghanistan",
    "KE": "Kenya",
    "MQ": "Martinique",
    "GB": "United Kingdom",
    "SK": "Slovakia",
    "CZ": "Czech Republic",
    "DE": "Germany",
    "FR": "France",
```

Figure 6. Top of mapping for 'country' in login.csv

2. **Preliminary Account Filtering:** Despite login.csv being the master dataset, trades.csv holds most of the vital data for account longevity, therefore data considerations must be made for trades.csv.

Login.csv has a minimum account registration date of 1/09/2016 and maximum of 5/03/2024, a range of 7.5 years, whilst trades.csv has a minimum trading date of 3/01/2023 and maximum of 29/02/2024, a range of only 1.2 years. This means that if accounts registered before 3/01/2023 were included in the final dataset, there would be missing trading data. These accounts are filtered out to avoid misrepresentation of data which could negatively influence a predictive model.

This filtering strategy reduces the number of unique accounts in the master dataset from 11976 to 8167.

3. **Currency Conversions:** Figure 7 shows conversion rates per day for every currency in the dataset from the daily chart directory, that have been merged into a single dataframe for calculation purposes.

date	AUD	EUR	GBP	NZD	CAD	CHF	CNH	HKD	HUF	JPY	MXN	NOK	PLN	SEK
2023-01-03	0.67233	1.05472	1.19663	0.62510	0.732032	1.068810	0.144454	0.128003	0.002637	0.007633	0.051531	0.099209	0.225695	0.094656
2023-01-04	0.68296	1.06030	1.20527	0.62923	0.741988	1.076403	0.144975	0.127946	0.002682	0.007541	0.051634	0.099242	0.227394	0.095211
2023-01-05	0.67513	1.05207	1.18981	0.62292	0.737039	1.068079	0.145178	0.127996	0.002658	0.007496	0.051754	0.097625	0.224621	0.093806
2023-01-06	0.68674	1.06437	1.20924	0.63485	0.743859	1.079319	0.146431	0.128114	0.002702	0.007573	0.052284	0.100017	0.227017	0.095164
2023-01-09	0.69091	1.07288	1.21725	0.63708	0.747010	1.085882	0.147518	0.128157	0.002709	0.007583	0.052259	0.100845	0.228689	0.096313
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2024-02-23	0.65621	1.08190	1.26710	0.61912	0.740960	1.134984	0.138903	0.127915	0.002827	0.006644	0.058499	0.094905	NaN	0.096900
2024-02-26	0.65397	1.08503	1.26843	0.61720	0.740505	1.136454	0.138720	0.127817	0.002789	0.006636	0.058492	0.095128	NaN	0.097266
2024-02-27	0.65426	1.08435	1.26837	0.61698	0.739262	1.138265	0.138654	0.127813	0.002775	0.006644	0.058609	0.094991	NaN	0.097055
2024-02-28	0.64935	1.08361	1.26595	0.60958	0.736469	1.137864	0.138644	0.127753	0.002756	0.006636	0.058509	0.094397	NaN	0.096729
2024-02-29	0.64952	1.08046	1.26240	0.60853	0.736480	1.130736	0.138812	0.127735	0.002754	0.006668	0.058641	0.094156	NaN	0.096415

Figure 7. Concatenated dataframe of conversion rates from daily chart directory

Three of the variables in trades.csv need to be converted from their respective currencies to USD to ensure consistent analysis, these are ‘commission’, ‘swaps’, and ‘profit’. Conversions are made for non-USD trades using the above dataframe.

	account_currency	commission	swaps	profit
0	EUR	0.00	-1.12	10.83
1	EUR	0.00	-2.25	23.16
2	GBP	0.00	0.00	105.42
3	GBP	0.00	-7.21	485.26
4	USD	-0.91	0.00	33.15

	account_currency	commission	swaps	profit
0	EUR	0.00	-1.199330	11.597089
1	EUR	0.00	-2.409368	24.800423
2	GBP	0.00	0.000000	134.137462
3	GBP	0.00	-9.174076	617.449677
4	USD	-0.91	0.000000	33.150000

Figure 8. Head of merged df before conversion (left)
Figure 9. Head of merged df after conversion (right)

As can be seen Figures 8 and 9, commission, swaps, and profit have been appropriately converted to USD dependent on the original account currency, whilst the USD account retains the same values.

3.3 | Feature Engineering

3.3.1 | Feature generation from trades.csv:

Trades.csv is merged into the master dataset login.csv on the unique account id in 'login'. In this process, sixteen new trading features are generated through various calculations, which will help generate trading insights for client accounts in analysis as well as inputs for training a predictive model (see Figure 10).

New Feature	Description	Formula	Justification
Total_Trades	The total number of trades an account has made.	Σ(count of 'ticket' per 'login')	This feature indicates the level of trading activity. Whilst total trades won't appear in the final dataset as it is a cumulative feature and not a representative average, it will be used for calculating other features.
Buy_Percentage	The ratio of trades that are buys vs sells for an account.	$\text{Buy\_Trades} = \sum_{\text{cmd}=0} (\text{count of 'ticket' per 'login'})$ $\left(\frac{\text{Buy\_Trades}}{\text{Total\_Trades}}\right) \times 100$	Reflects the trading preference of the client. A balanced buy-sell ratio might suggest a more stable trading strategy, potentially influencing longevity.
Average_Volume	The average volume of shares per trade for an account.	$\frac{\sum \text{volume}}{\text{Total\_Trades}}$	Helps to understand the scale of transactions. Larger volumes may indicate more committed or professional traders, potentially linked to longer client relationships.
Average_Volume_USD	The average price of shares traded per trade for an account.	$\frac{\sum \text{volume\_usd}}{\text{Total\_Trades}}$	Indicates the financial depth of trades. Higher transaction values could relate to more serious investors who might trade over longer periods.
Average_DPM	The average dollar per million value per account.	$\frac{\sum\left(\frac{\text{profit}}{\text{volume\_usd}/1{,}000{,}000}\right)}{\text{Total\_Trades}}$	Suggests profitability per unit of trade, which may motivate continued trading if returns are favourable.
Unique_Symbols_Traded	The monthly average number of unique shares traded per account.	$\frac{1}{N_{\text{months}}} \sum_{m=1}^{N_{\text{months}}} \text{unique symbols}_m$	Diversity of trading could indicate a more sophisticated trading approach, which might be linked to sustained engagement.
Peak_Trading_Times	The most frequent hour in which a client makes trades per account.	mode(trading hour per 'login')	Identifying peak trading times can help in understanding the client's trading habits, which could be useful for predicting active trading periods and their retention.

Ratio_Profitable_Trades	The ratio of profitable trades made to non-profitable trades per account.	$\frac{\text{Number of trades with profit}>0}{\text{Total\_Trades}}$	A higher ratio of profitable trades could encourage continued trading on the platform.
Profit_Loss_Variability	The variability of profit and loss per trade per account.	$\sigma_{\text{profit}} = \sqrt{\frac{\sum(\text{profit}-\mu_{\text{profit}})^2}{\text{Total\_Trades}-1}}$	Higher variability may indicate riskier strategies, potentially affecting client longevity negatively or positively depending on risk tolerance.
Average_Trade_Duration	The average duration of a trade per account.	$\frac{\sum(\text{close\_time}-\text{open\_time})}{\text{Total\_Trades}}$	Longer durations may indicate a more strategic, long-term approach to trading, possibly correlating with sustained engagement.
TP/SL Hit Ratio	The ratio between when a trader hits their take profit vs when a trader hits their stop loss per account.	$\frac{\text{TP\_hits}}{\text{SL\_hits}}$	This ratio might reflect the effectiveness of a trader's strategy, influencing their satisfaction and persistence on the platform.
Reward_Risk_Ratio	The ratio between profit taken when a trader hits their take profit vs the loss made when a trader hits their stop loss per account.	$\frac{\text{Average\_Profit\_TP}}{-\text{Average\_Loss\_SL}}$	A favourable reward-to-risk ratio could encourage continued trading activities.
Most_Common_Trading_Method	The most common trading method used per account.	mode(reason per 'login')	Indicates the client's preferred trading strategy, some preferred methods may have a higher correlation with client longevity.
Average_Commission	The average commission per trade per account.	$\frac{\sum \text{commission}}{\text{Total\_Trades}}$	Costs associated with trading could influence trading frequency and duration, higher costs might deter long-term engagement.
Average_Swaps	The average swaps per trade per account.	$\frac{\sum \text{swaps}}{\text{Total\_Trades}}$	Swap rates affect the cost/reward of holding positions overnight, potentially influencing continued trading decisions.
Average_Profit	The average profit per trade per account.	$\frac{\sum \text{profit}}{\text{Total\_Trades}}$	Directly correlates with the financial success on the platform, higher profits might motivate prolonged trading activities.

Table 6. Description of features generated from trade data

	login	country	account_currency	reg_date	Total_Trades	Buy_Percentage	Average_Volume	Average_Volume_USD	Average_DPM	Unique_Symbols_Traded	Peak_Trading_Times	Ratio_Profitable_Trades	Profit_Loss_Variability	Average_Trade_Duration	TP/SL Hit Ratio	Reward_Risk_Ratio	Most_Common_Trading_Method
0	524974	Switzerland	USD	2023-05-17 03:13:02	143	48.251748	0.046364	1.816077e+04	96.362040	1.000000	17	0.979021	1.805020	4123.216783	0.000000	0.000000	1
1	524978	Austria	EUR	2023-06-07 05:58:36	1392	47.485632	1.230632	4.044965e+08	-3.415139	9.333333	17	0.762931	316.119097	36404.811782	3.375000	0.086763	1
2	524979	France	USD	2023-06-07 06:17:30	2194	49.635369	0.013943	6.724644e+03	-264.410462	3.500000	17	0.718323	11.676818	57056.876937	0.087081	-0.549598	5
3	524984	Singapore	USD	2023-10-31 08:34:08	244	38.934426	0.104262	1.836608e+06	-53.917476	3.000000	16	0.606557	42.988956	20359.959016	1.923077	-0.155076	1
4	760487	Singapore	SGD	2023-01-04 08:48:24	69	15.942029	0.012609	2.783882e+03	-3403.656636	1.000000	17	0.463768	12.396847	46984.231884	0.000000	0.000000	1

Figure 10. Head of resultant dataset after merging trades.csv and login.csv

3.3.2 | Target Variable: Longevity Considerations and Calculations

Due to the complex nature of the dataset a specific definition of longevity is hard to close in on. Under normal circumstances longevity could be defined as the duration between when a given account becomes shut and when it was first registered. However, the data does not comprise this information, and instead consists of account registration dates and their trading activities. Therefore, after consultation with the client, an alternate definition of longevity was identified based on account trading activity. This is, the duration between the time of a given account's final trade minus the time of their first trade. This model of longevity is arguably a more accurate representation, as it is a measure of a client's interaction with the brokerage platform service, and not just the length of a time that their account is open for. It is however, important to consider that client longevity is only a small factor of customer engagement, and for further analysis other aspects such as customer lifetime value should be taken into account.

A secondary consideration for longevity calculations is the activity status of an account. If an account's 'final trade' in the dataset is very close to the maximum date range (most recent date in the dataset) it is fair to assume that this trade might not be the client's last, and that if the date range of the dataset was extended, more trades from this client would appear. These accounts are defined as 'active' accounts and don't have a true 'final trade', and therefore don't have a fixed longevity duration. And on the other hand, 'inactive' accounts are defined as accounts that have a final trade not close to the maximum date range, and therefore do have a fixed longevity. These accounts are much more useful for longevity analysis as they have closed bounds.

Definitions:

- *Inactive account:* Final trade of this account does not appear in the most recent month of the dataset.
- *Active account:* Final trade of this account does appear in the most recent month of the dataset.
- *Account longevity:* Final trade time - first trade time.
- *Longevity maximum range (based on merged dataset):* 3/01/2023 - 29/02/2024 = 14 months. Accounts that registered in the most recent month of the dataset are removed to avoid confusion regarding activity status.

Even though an active and inactive account may have the same longevity duration, we are unable to label them the same due to the active account having an unbounded upper range. Inactive accounts have been identified to be superior for longevity analysis, yet we do not wish to remove all active accounts from the dataset as they are also important for identifying outstanding clients.

Account longevity is binned into five categories (in days) based on longevity distributions and common practice: 0-30, 30-90, 90-180, 180-270, and 360+. Activity issues are addressed by only including 'inactive' accounts for the first four bins, and then both inactive and active for the final bin. In this way, we

have strict durations on accounts for bins that have a bounded upper range, while being able to include outstanding, target active accounts in the final bin.

The filtering strategy used for binning longevity based on activity status reduces unique accounts in the merged dataset from 8167 to 5725. A total reduction in included accounts from the beginning of pre-processing by 52.50%.

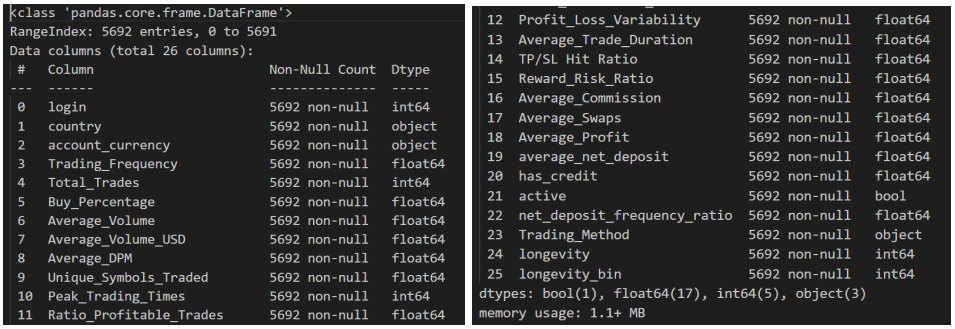
3.3.3 Feature Generation From longevity and daily_report.csv:

Both longevity and daily report dataframes are merged with the master dataframe on the unique account id in 'login'. In this process, seven more features are generated including the target variable and other variables related to the target variable from the longevity dataframe (see Figures 11, 12, and 13). This step is vital, as the dataset now has a defined target which can be analysed and predicted using a machine learning model. Three more features are also generated from merging daily reports with the master dataframe. These features will increase the dataset's depth, providing additional insights into trading behaviours for analysis and modelling.

New Feature	Description	Formula	Justification
longevity	The timeframe in days which a client actively trades on the brokerage platform.	$\frac{(\max(\text{close\_time}) - \min(\text{open\_time}))}{\text{days}}$	As the target variable for analysis, longevity provides a direct endpoint for predictive models focusing on client retention.
active	If an account is actively trading, defined by whether a trade appears in the most recent month of the dataset.	$\begin{cases} 1 & \text{if } \max(\text{close\_time}) \geq \text{most\_recent\_month\_start} \\ 0 & \text{otherwise} \end{cases}$	Active status is not used for analysis or modelling. This variable is used to filter out accounts based on the respective bin.
longevity_binned	Longevity, but binned into several categories per account.	$\begin{cases} 0 & \text{if longevity} < 30 \text{ days} \\ 1 & \text{if } 30 \leq \text{longevity} < 90 \text{ days} \\ 2 & \text{if } 90 \leq \text{longevity} < 180 \text{ days} \\ 3 & \text{if } 180 \leq \text{longevity} < 270 \text{ days} \\ 4 & \text{if } 270 \leq \text{longevity} < 360 \text{ days} \\ 5 & \text{if longevity} \geq 360 \text{ days} \end{cases}$	Longevity binned is not used for modelling, however it is used for ease of analysis and visualisation.
Trading_Frequency	Average frequency of trading activity per account.	$\frac{\text{Total\_Trades(login)}}{\text{Longevity(login)}}$	Trading frequency quantifies how often a client trades, which may affect a client's overall longevity.
average_net_deposit	Average net deposits per account.	$\frac{\sum(\text{net\_deposit})}{N}$	Higher average net deposits may indicate a client's financial commitment to trading on the platform, potentially correlating with prolonged trading activity.
has_credit	If an account has been given credit at any stage.	$\begin{cases} 1 & \text{if any } (\text{credit} > 0) \\ 0 & \text{otherwise} \end{cases}$	Accounts that have received credit may show enhanced trading activity due to increased financial leverage, potentially

			extending their longevity on the platform.
net_deposit_frequency_ratio	The ratio of non-zero net_deposit entries to zero net_deposit entries per account.	$\frac{\text{count}(\text{net\_deposit} \neq 0)}{\text{count}(\text{net\_deposit} = 0)}$	A higher net deposit frequency ratio suggests active financial engagement, serving as a key indicator of a client's potential longevity in trading.

Table 7. Description of features generated from longevity and report data



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5692 entries, 0 to 5691
Data columns (total 26 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   login                       5692 non-null   int64
 1   country                     5692 non-null   object
 2   account_currency            5692 non-null   object
 3   Trading_Frequency           5692 non-null   float64
 4   Total_Trades                5692 non-null   int64
 5   Buy_Percentage              5692 non-null   float64
 6   Average_Volume              5692 non-null   float64
 7   Average_Volume_USD          5692 non-null   float64
 8   Average_DPM                 5692 non-null   float64
 9   Unique_Symbols_Traded       5692 non-null   float64
 10  Peak_Trading_Times          5692 non-null   int64
 11  Ratio_Profitable_Trades     5692 non-null   float64
 12  Profit_Loss_Variability     5692 non-null   float64
 13  Average_Trade_Duration      5692 non-null   float64
 14  TP/SL Hit Ratio             5692 non-null   float64
 15  Reward_Risk_Ratio           5692 non-null   float64
 16  Average_Commission          5692 non-null   float64
 17  Average_Swaps               5692 non-null   float64
 18  Average_Profit              5692 non-null   float64
 19  average_net_deposit         5692 non-null   float64
 20  has_credit                  5692 non-null   float64
 21  active                      5692 non-null   bool
 22  net_deposit_frequency_ratio 5692 non-null   float64
 23  Trading_Method              5692 non-null   object
 24  longevity                   5692 non-null   int64
 25  longevity_bin               5692 non-null   int64
dtypes: bool(1), float64(17), int64(5), object(3)
memory usage: 1.1+ MB
```

Figure 11 & Figure 12. Information about the 2nd merged dataframe including newly generated features

Average_Commission	Average_Swaps	Average_Profit	longevity	active	longevity_bin	Trading_Frequency	average_net_deposit	has_credit	net_deposit_frequency_ratio
-6.935588	-1.481077	25.668214	142	False	2	9.802817	-148.587567	0.0	0.056225
0.000000	-0.111285	-0.837867	107	False	2	20.504673	6.950570	0.0	0.047809
0.000000	0.002172	-10.232797	13	False	0	5.307692	2.490144	0.0	0.004819
0.000000	-0.023176	-0.524588	9	False	0	9.444444	0.167482	0.0	0.007246
-0.101129	-0.009749	-0.210692	372	False	5	1.301075	2.798054	0.0	0.022388

Figure 13. Head of resultant dataset new features after merging trades.csv and login.csv

3.4 | Final Pre-processing and Final Dataset

3.4.1 | Variable Formatting Pre-processing

- **Datetime conversions:** reg_date from login.csv, open_time and close_time from trades.csv, and record_time from daily_reports converted to datetime.
- **Int conversions:** login from trades.csv converted to int.
- **Remapping of Trading Method:** Upon exploratory analysis and discussion with the client, the three most important trading methods that were identified were 'Client', 'Expert', and 'Mobile'. Alternate categories of trading methods were aggregated into a separate 'Other' category. This was done using reason.csv to translate the encodings into their respective trading method strings.
- **Categorisation:** Variables country, account_currency, Trading_Method, Peak_Trading_Times, has_credit are converted to 'category'.

3.4.2 | Data Cleaning

- **Imputation:** Upon inspection, it is noted that there are NaN values in the Ratio_Profitable_Trades, Profit_Loss_Variability, TP/SL Hit Ratio, and Reward_Risk_Ratio (see Figures 14 and 15). These values occur when either there is zero profitability among trades, or

the client does not engage in using take profits or stop losses in their trades. Therefore, these NaN values can be safely imputed with '0'.

```
Count of NaN values per column:
login                          0
country                        0
account_currency               0
Total_Trades                   0
Buy_Percentage                 0
Average_Volume                 0
Average_Volume_USD             0
Average_DPM                    0
Unique_Symbols_Traded          0
Peak_Trading_Times             0
Ratio_Profitable_Trades      756
Profit_Loss_Variability      556
Average_Trade_Duration         0
```

```
TP/SL Hit Ratio              1970
Reward_Risk_Ratio            1970
Average_Commission              0
Average_Swaps                   0
Average_Profit                  0
longevity                       0
active                          0
longevity_bin                   0
Trading_Frequency               0
average_net_deposit             1
has_credit                      1
net_deposit_frequency_ratio     1
Trading_Method                  0
dtype: int64
```

Figure 14 and Figure 15. NaN values before imputation

- Index and reg_date are dropped from the dataset, this is because there is already a unique identifier in login, and reg_date is encapsulated in longevity.
- **Outlier removal:** Extreme outliers are removed through visualisation techniques (see Figure 16). Firstly, all variables and their distributions to binned longevity are visualised in box plots. Next, a number of rules are defined for outlier removal based on observations (see Figure 17), and finally the outliers are removed from the master dataset.

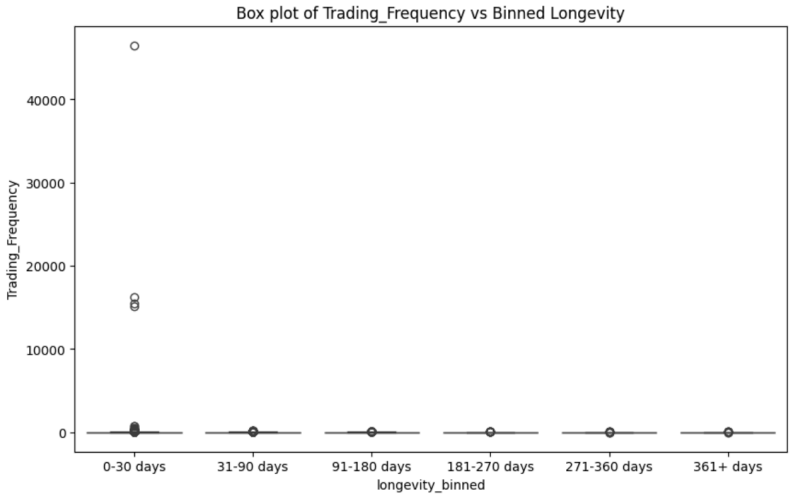


Figure 16. Example of box plot outlier visualisation for Trading_Frequency

```
Trading Frequency > 1000: 4
Profit Loss Variability > 10k: 10
Average Trade Duration > 3M: 5
TP/SL Hit Ratio > 400: 5
Reward Risk Ratio < -5000 or > 5000: 4
Average Profit < -20k: 1
Average Net Deposit not between -2k and 2k: 3
Net Deposit Frequency Ratio > 1: 2
```

Figure 17. Number of outliers to be removed from the master dataset based on created rules

3.4.3 | Final Dataset

There are a total of 26 variables in the final dataset (see Figures 18, 19, and 20)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5692 entries, 0 to 5691
Data columns (total 26 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   login                       5692 non-null   int64
 1   country                     5692 non-null   category
 2   account_currency            5692 non-null   category
 3   Trading_Frequency           5692 non-null   float64
 4   Total_Trades                5692 non-null   int64
 5   Buy_Percentage              5692 non-null   float64
 6   Average_Volume              5692 non-null   float64
 7   Average_Volume_USD          5692 non-null   float64
 8   Average_DPM                 5692 non-null   float64
 9   Unique_Symbols_Traded       5692 non-null   float64
 10  Peak_Trading_Times          5692 non-null   category
 11  Ratio_Profitable_Trades     5692 non-null   float64
```

```
 12  Profit_Loss_Variability     5692 non-null   float64
 13  Average_Trade_Duration      5692 non-null   float64
 14  TP/SL Hit Ratio             5692 non-null   float64
 15  Reward_Risk_Ratio           5692 non-null   float64
 16  Average_Commission          5692 non-null   float64
 17  Average_Swaps               5692 non-null   float64
 18  Average_Profit              5692 non-null   float64
 19  average_net_deposit         5692 non-null   float64
 20  has_credit                  5692 non-null   category
 21  active                      5692 non-null   bool
 22  net_deposit_frequency_ratio 5692 non-null   float64
 23  Trading_Method              5692 non-null   category
 24  longevity                   5692 non-null   int64
 25  longevity_bin               5692 non-null   int64
dtypes: bool(1), category(5), float64(16), int64(4)
memory usage: 935.0 KB
```

Figure 18 (left) and Figure 19 (right). Final dataset info

	login	country	account_currency	Trading_Frequency	Total_Trades	Buy_Percentage	Average_Volume	Average_Volume_USD	Average_DPM	Unique_Symbols_Traded	...	Average_Commission	Average_Swaps	Average_Profit	average_net_deposit	has_credit	active	net_deposit_frequency_ratio	Trading_Method	longevity	longevity_bin
0	524978	Austria	EUR	9.802817	1392	47.485632	1.230632	4.044965e+08	-3.415139	9.333333	...	-6.935588	-1.481077	25.668214	-148.587567	0.0	False	0.056225	Expert	142	2
1	524979	France	USD	20.504673	2194	49.635369	0.013943	6.724644e+03	-264.410462	3.500000	...	0.000000	-0.111285	-0.837867	6.950570	0.0	False	0.047809	Mobile	107	2
2	760487	Singapore	SGD	5.307692	69	15.942029	0.012609	2.783882e+03	-3403.656636	1.000000	...	0.000000	0.002172	-10.232797	2.490144	0.0	False	0.004819	Expert	13	0
3	804664	Malaysia	USD	9.444444	85	63.529412	0.018706	5.490400e+03	17.604088	6.000000	...	0.000000	-0.023176	-0.524588	0.167482	0.0	False	0.007246	Mobile	9	0
4	804687	Australia	AUD	1.301075	484	51.239669	0.019773	5.271755e+05	45.416080	2.769231	...	0.101129	-0.009749	-0.210692	2.798054	0.0	False	0.022388	Mobile	372	5

Figure 20. Head of final dataset

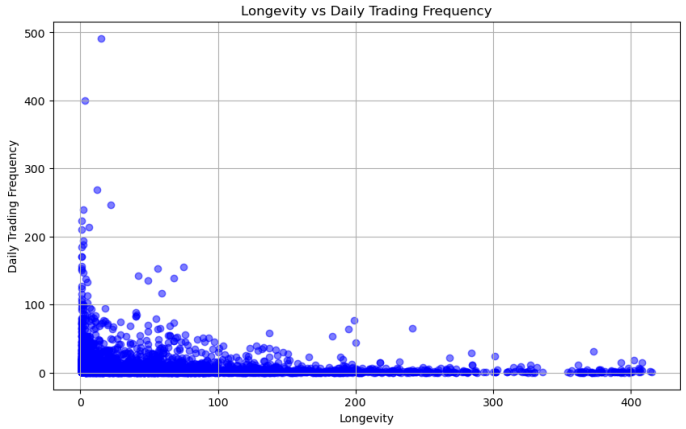
3.5 | Data Exploration

Initial data exploration assisted us in establishing a cursory understanding of the situation, as well as the shape of the data. Three phases of user growth were revealed, with a significant surge in registrations in 2023-2024 (see Appendices A and B). High attrition rates are observed among newer accounts, with most becoming inactive shortly after initial trade (see Appendix C). Overall trading activity has steadily increased, with noticeable surges and declines influenced by market conditions and seasonal trends (see Appendix D). Different countries show varied preferences in trading methods, influenced by local market dynamics (see Appendices E and F). Trading activity is dominated by expert methods, with a growing trend towards mobile trading (see Appendix G). While trading outcomes are evenly distributed, there's a slight edge towards losses (see Appendix H). Accounts typically operate around breakeven in terms of Daily Profit Margin, with some outliers exhibiting significantly higher profitability (see Appendix I).

From here, we conducted preliminary longevity analysis to identify what factors would likely have the greatest explanatory power for longevity. This revealed that user engagement declines over time, stabilising after 180 days, with a small group remaining active long-term (see Appendix J). Trading activity follows a similar pattern, with heavy initial trading diminishing gradually (see Appendices K and L). Expert trading becomes more prevalent with user experience, while mobile trading is popular among newcomers (see Appendix M). Profitability shows an initial loss for new users but improves with longer engagement, reaching a turning point around 270 days (see Appendix N). Regional trends indicate more short-term trading in emerging markets, while developed economies display stability and consistent activity across various time frames (see Appendix O).

We then leveraged our experiences with the dataset and these elementary insights to guide or primary analysis, the key points of which are discussed hereafter.

3.5.1 | Daily Trading Frequency and Longevity



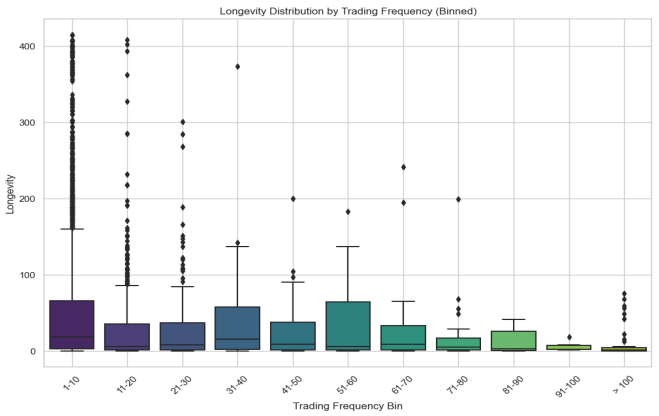
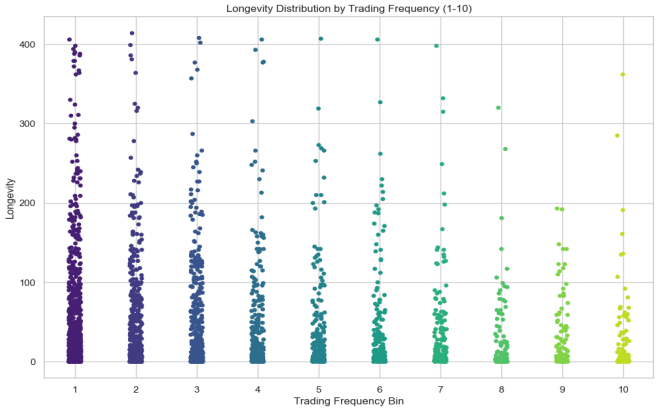


Figure 21. Daily Trading Frequency and Longevity (Scatter Plot) (left)
Figure 22. Daily Trading Frequency and Longevity (Box Plot) (right)

Figure 21 demonstrates the inverse relationship between trading frequency and longevity. As frequency increases, longevity is less likely to be high. The traders who execute numerous trades frequently may not consider all the risks associated with their trading plan, leading to risky behaviour and a shorter life span. The drop in trading frequency reflects the knowledge of the trader who is able to stay patient, avoid fear of missing out (FOMO) and adhere to their trading strategy to achieve a longer-term goal. Consistently, Menkhoff et al. (2008) study the impact of risk-taking behaviour on fund managers. Menkhoff and his colleagues figured out that even though inexperienced managers may achieve high profitability in the short run, they expose their funds to much higher risk, potentially leading to greater loss and reduced longevity.

We decided to take one step further and investigate the optimal trading frequency, or the rate at which the trader tends to have a higher longevity. As shown in Figure 22, the trading frequency of 11-20 still has some outstanding traders that can last over 100 to more than 400 days. This distribution is more consistent in the first bin. Figure 23 demonstrates that as the trading frequency rises to ten, the longevity gradually decreases. In the final step, we investigated the trading frequency distribution under one. Regardless, a reduction in trading frequency no longer greatly influences longevity after this threshold, illustrated by the random distribution in Figure 24.



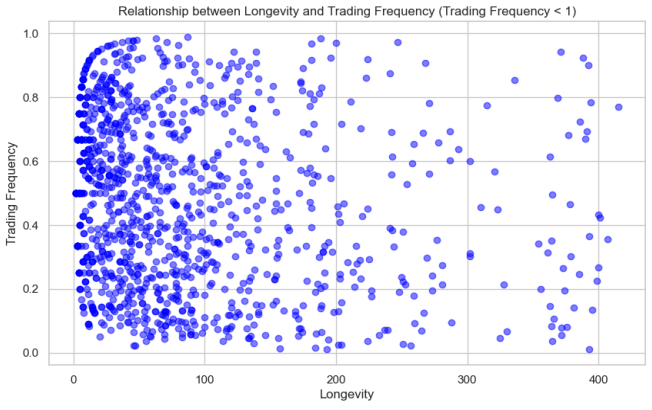


Figure 23. Daily Trading Frequency and Longevity (Strip Plot) (left)
Figure 24. Daily Trading Frequency and Longevity (Scatter Plot) (right)

In conclusion, traders who execute an average of less than 20 trades per day tend to have the highest longevity, with those with an average of less than 10 being consistently longer, likely spreading their trades over a longer period.

3.5.2 | Profitability and Longevity

3.5.2.1 | Average Profit

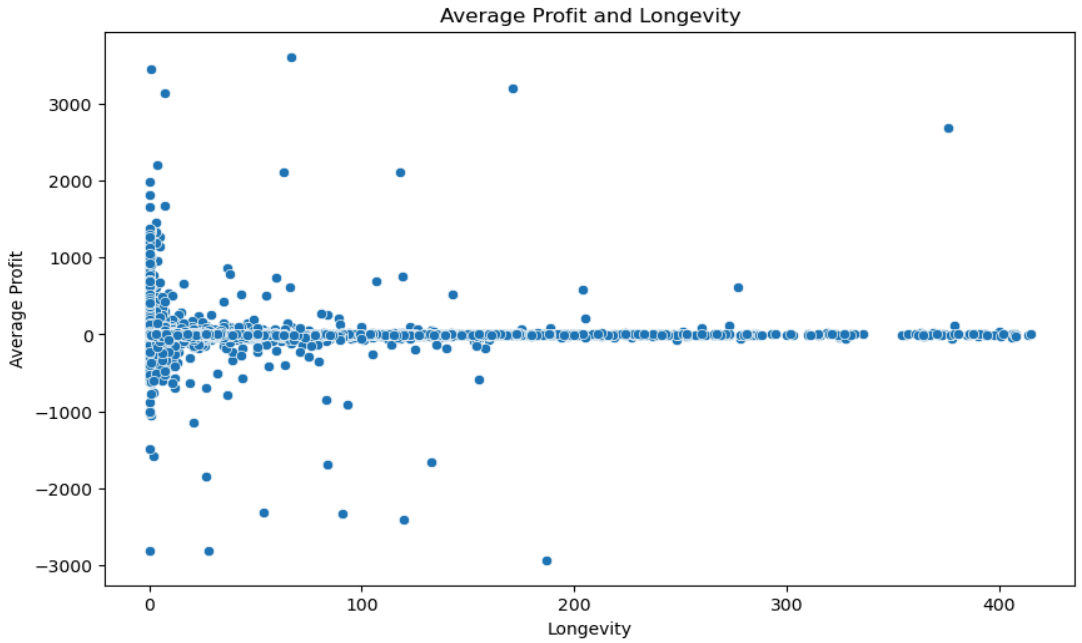


Figure 25. Average Profit and Longevity (Strip Plot)

Figure 25 indicates that higher profitability is not exclusively associated with higher longevity. Regardless, the trader's loss does have a great impact on longevity. Specifically, Figure 26 shows that a slight loss from (-100 < profit < 0) showed a better longevity distribution compared to the lower ranges. The largest short-term distribution is the group with a loss of -200 to -400. The data suggest that heavy-loss traders are more likely to exit the market and that profitability might not be the key determinant of a trader's longevity.

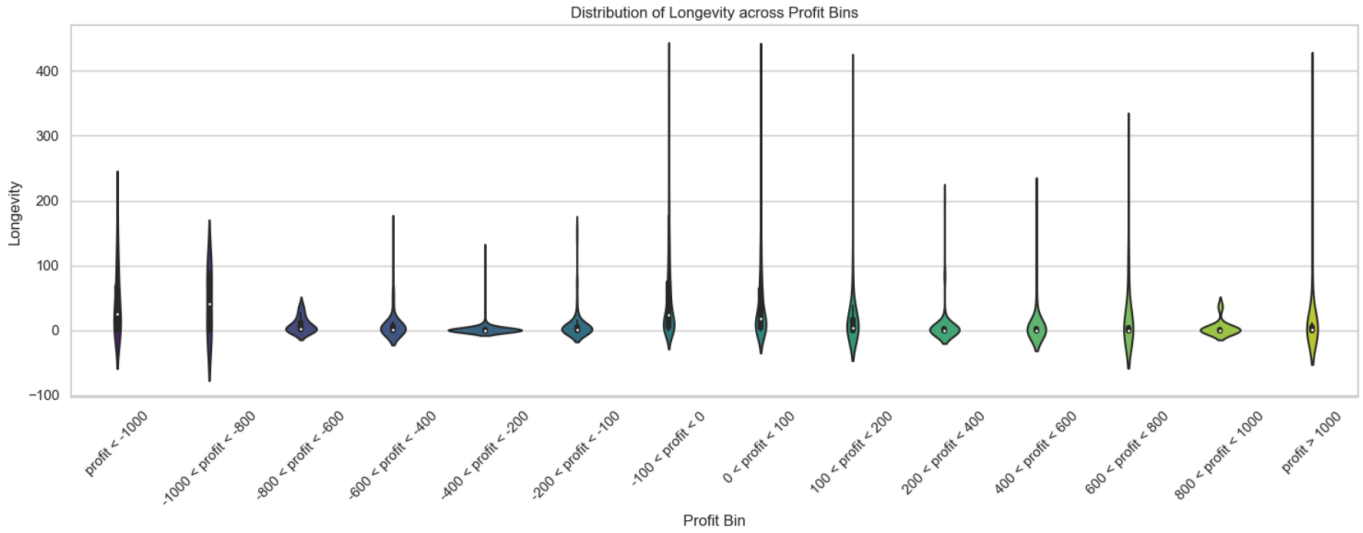


Figure 26. Distribution of Longevity across Profit Bins (Violin Plot)

With similar sample sizes to our model, a study from 2022 examining 5,164 traders from 2006 to 2012 has found a V-shaped relationship between the trader's profitability and their life span (Ma et al., 2022). Particularly, they found that the least profitable and the most profitable traders are more likely to stop trading than the average trader.

3.5.2.2 | Average DPM

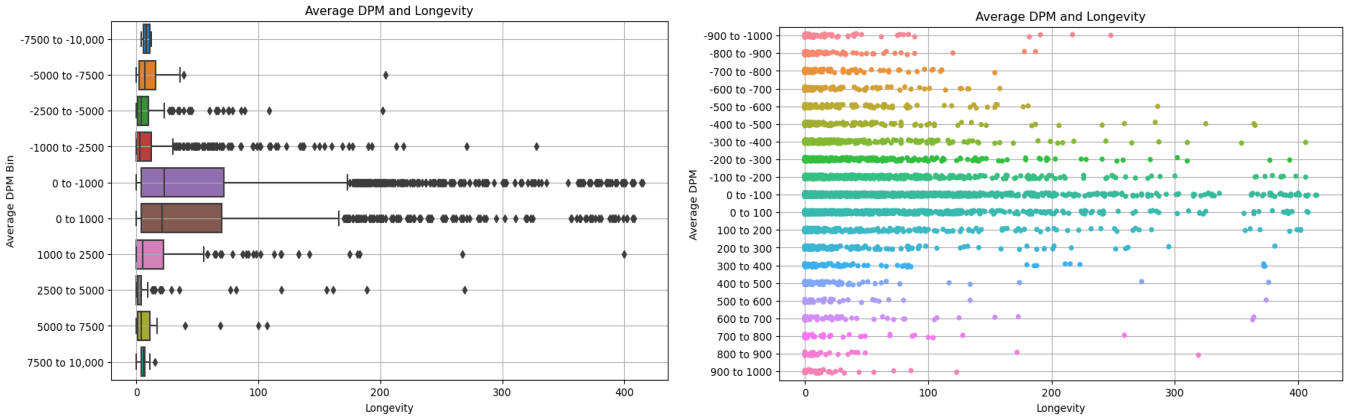


Figure 27. Average DPM and Longevity (Box Plot) (left)
Figure 28. Average DPM and Longevity (Strip Plot) (right)

The box plot from Figure 27 indicated that the more profit made per million value per account is not associated with higher longevity. Regardless, the stability in these metrics does have some correlation with the ability to survive of the trader, as the average DPM fluctuates around 0.

Taking one step further, we limit the average DPM to the range from -1000 to 1000 with 20 bins, as seen in the strip plot from Figure 28. The lower average DPM demonstrates a potential relationship with risk, as these traders may prioritise safety rather than excessive return. The concern for safety potentially influences the trader's ability to survive longer in the market.

3.5.2.3 | Ratio of Profitable Trades

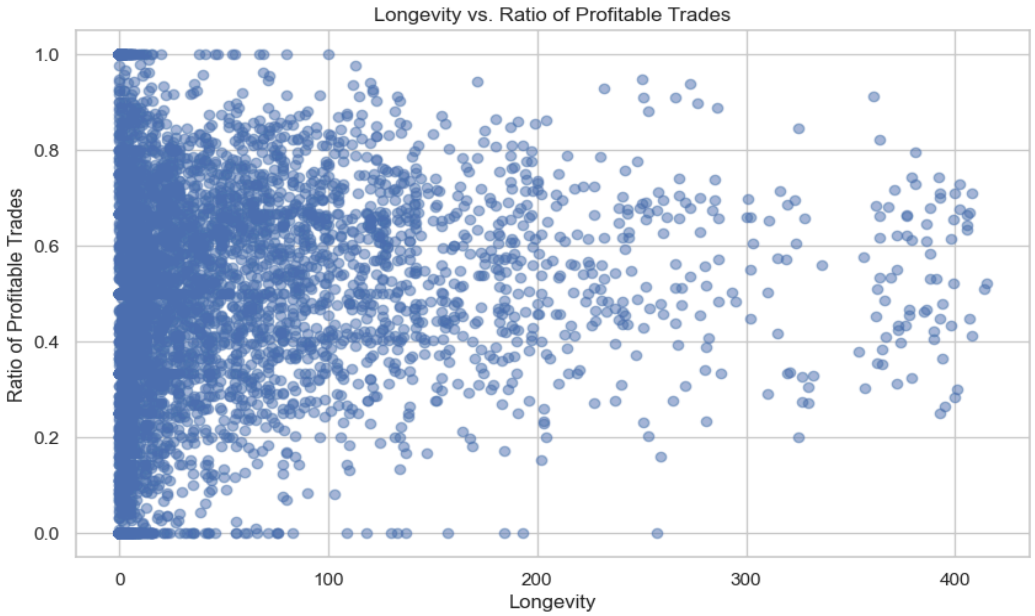


Figure 29. Longevity and Ratio of Profitable Trade

According to Figure 29, a higher ratio of profitable trades does not guarantee a higher longevity. Regardless, a certain threshold of winning ratio is required for traders to trade sustainably. The box plot from Figure 30 above shows a stiff increase in longevity as the Ratio of Profitable trades rises from 0 to 45%, then gradually plateaus before dropping again to 70%. This fluctuation suggested that a safe range of winning ratio is roughly from 40% to 70%. Going above this threshold, the traders may need to consider other factors that can impact the duration of their trading journey.

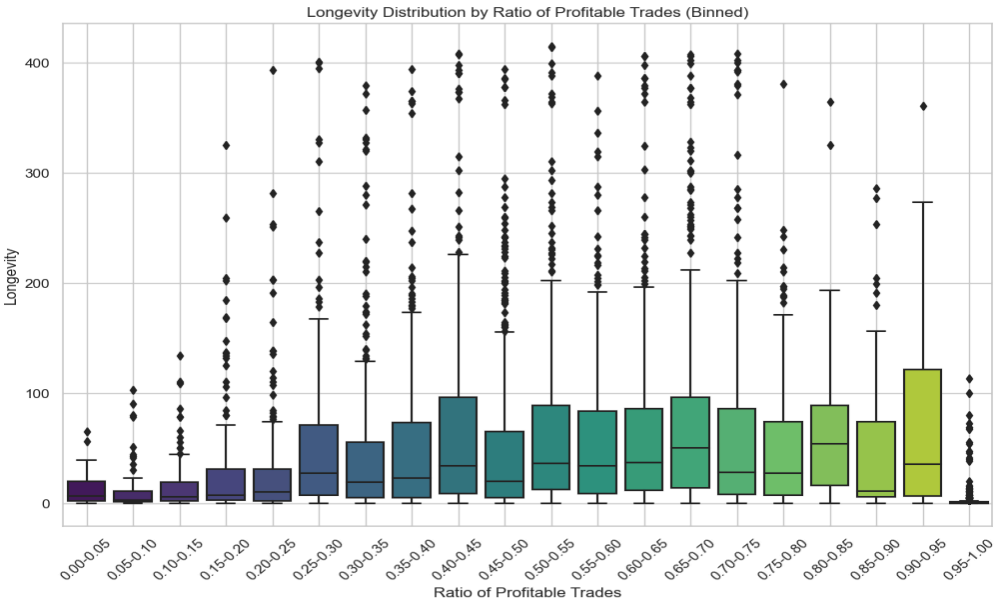


Figure 30. Longevity Distribution by Ratio of Profitable Trade

On the other hand, traders who exit the market quickly can yield a very high ratio of profitable trades purely due to luck. In the longer run, it would be very difficult for any trader to keep a winning percentage close to 100 all the time.

3.5.2.4 | Profit/Loss Variability

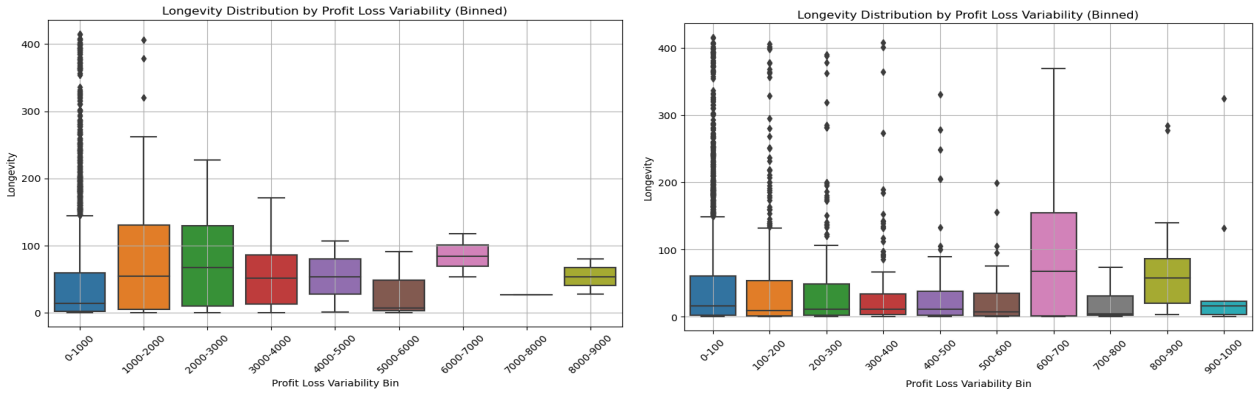


Figure 31. Longevity Distribution by Profit Loss Variability (10 bins from 0 -10,000) (left)
Figure 32. Longevity Distribution by Profit Loss Variability (10 bins from 0 -1,000) (right)

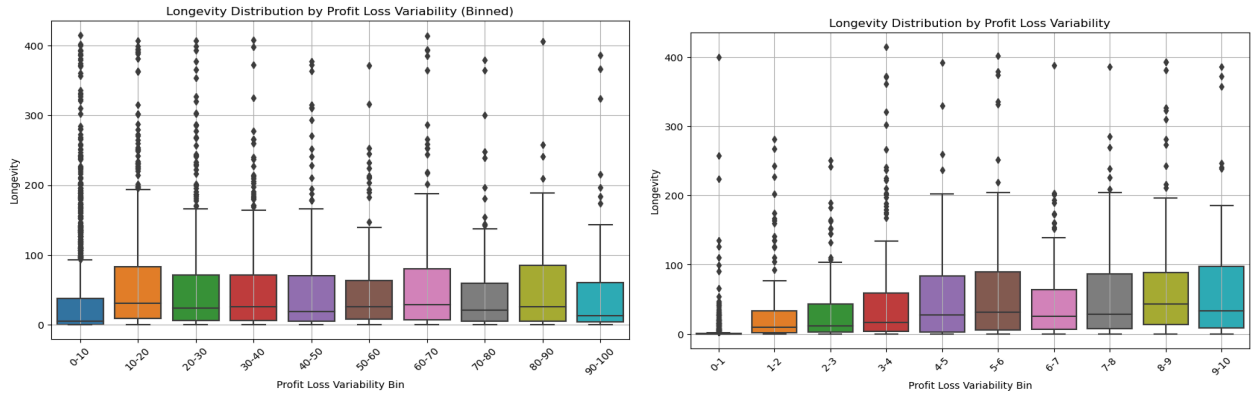
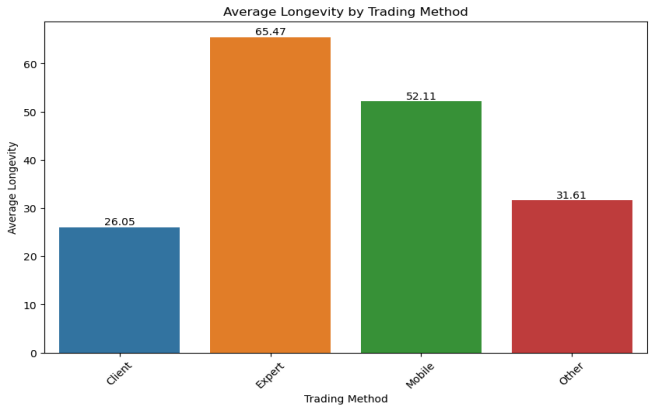


Figure 33. Longevity Distribution by Profit Loss Variability (10 bins from 0 -100) (left)
Figure 34. Longevity Distribution by Profit Loss Variability (10 bins from 0 -10) (right)

Accounts with higher variability in the profits and losses are seen to have lower longevity on average than those with lower variability (see Figures 31, 32, 33, and 34). In effect, this means that accounts that tend to take positions where they win or lose large sums tend to take those wins or losses and cease trading sooner than those with more conservative ones. This is more evident at the extremes, with more moderate variabilities being relatively flat, if still negative.

3.5.3 | Trading Method and Longevity



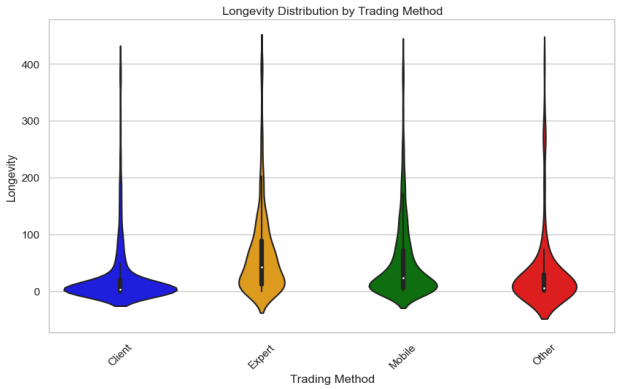


Figure 35. Average Longevity by Trading Method (left)
Figure 36: Longevity Distribution by Trading Method (right)

The Figures 35 and 36 show that the Client Trading Method performs worst among the four, with the lowest average of 30 days for longevity. A significant portion of the traders distribute around the lower end of the longevity scale. On the other hand, the Client method has the highest average longevity of 65 days while having a narrow distribution from 20,30 to above nearly 200. Mobile, even though having the second highest average longevity, has a considerable amount of traders' longevity distributed around 10 days. A possible explanation for the outperformance of the Expert trading method could be because they have access to expert knowledge, trading models and much more advanced support.

3.5.4 | Commission and Longevity

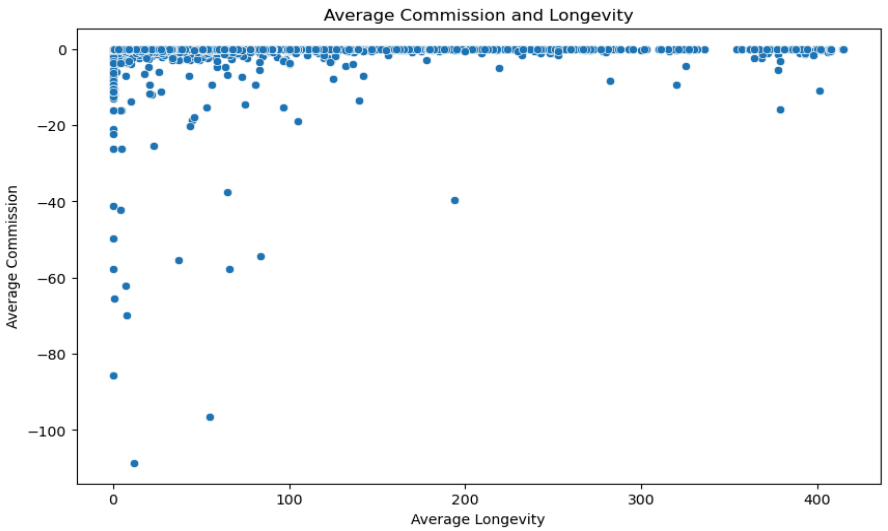


Figure 37. Average Commission and Longevity

The trading commission is highly correlated to the trading frequency. Traders that survive longer in the market tend to have lower commission fees, as seen in Figures 37 and 38. However, this could also be because these traders utilise the trading assets with lower commission costs or tighter spread between selling and buying. Either way, lower commission costs can conceivably improve the overall profitability of the trader, fostering sustained trading activity and enhancing the trader's longevity.

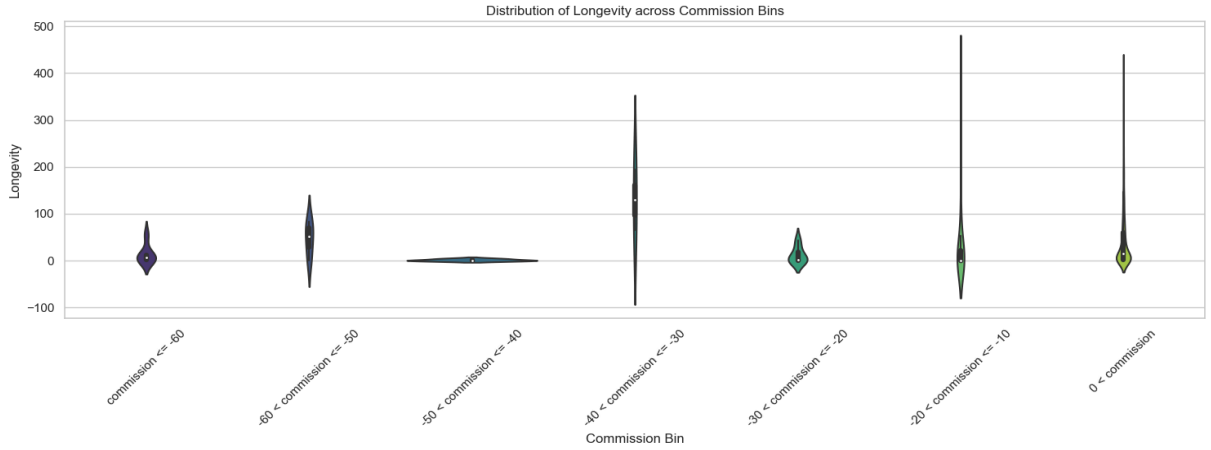


Figure 38. Distribution of Longevity across Commission Bins

3.5.5 | TP/SL Hit Ratio and Longevity

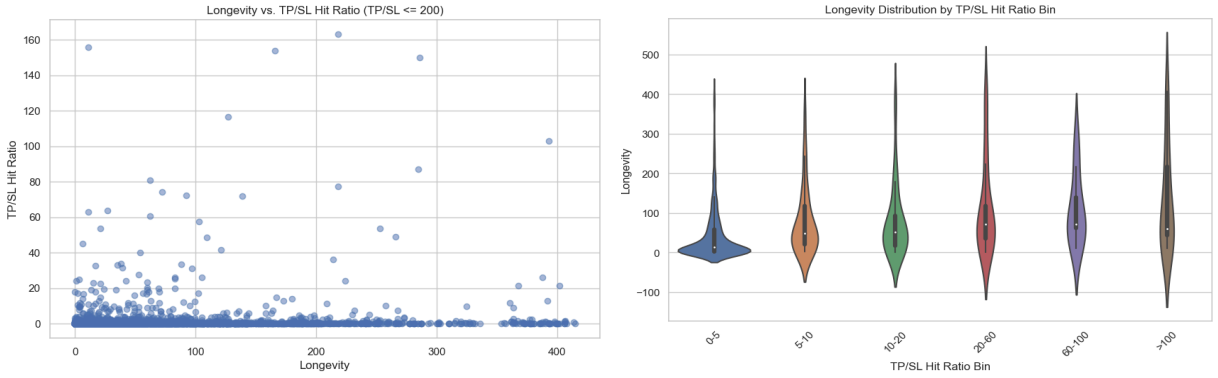


Figure 39. Longevity vs TP/SL Hit Ratio (left)
Figure 40. Longevity Distribution by TP/SL Hit Ratio (right)

The distribution of longevity increases as the TP/SL Hit ratio denotes a high correlation between these two variables (see Figures 39 and 40). Clients who frequently close a position as a result of hitting a TP ceiling tend to have a greater longevity than those that do so due to hitting a SL floor, or that don't use either measure at all (see Figure 41). The nature of this relationship suggests that, in addition to the ability to accurately determine winning positions, prudence in the form of being willing to accept a rational gain is an important determinant of longevity. Clients that regularly hit their SL limit more than their TP exhibit a similar prudence, but are not as likely to be long term users of the service, while those without TP or SL measures at all run the gamut from high to low longevity.

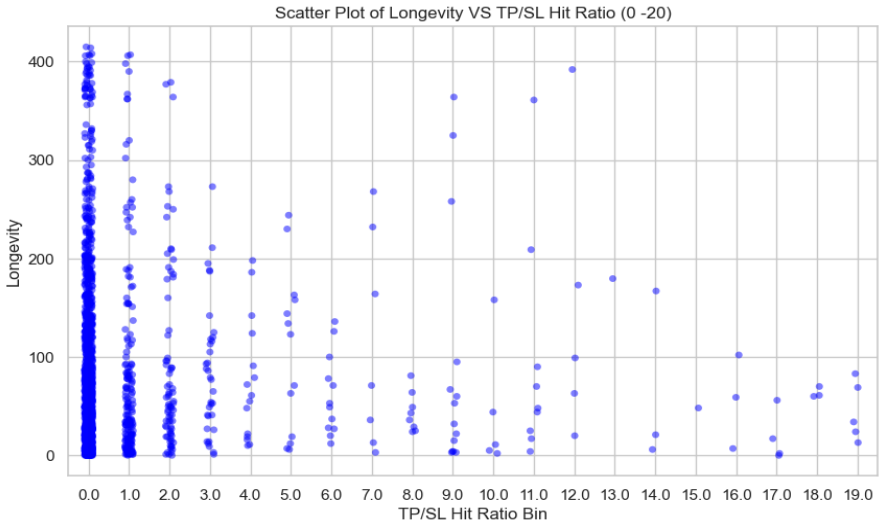


Figure 41. Distribution of Longevity across Commission Bins

3.5.6 | Net Deposit and Longevity

3.5.6.1 Average Net Deposit

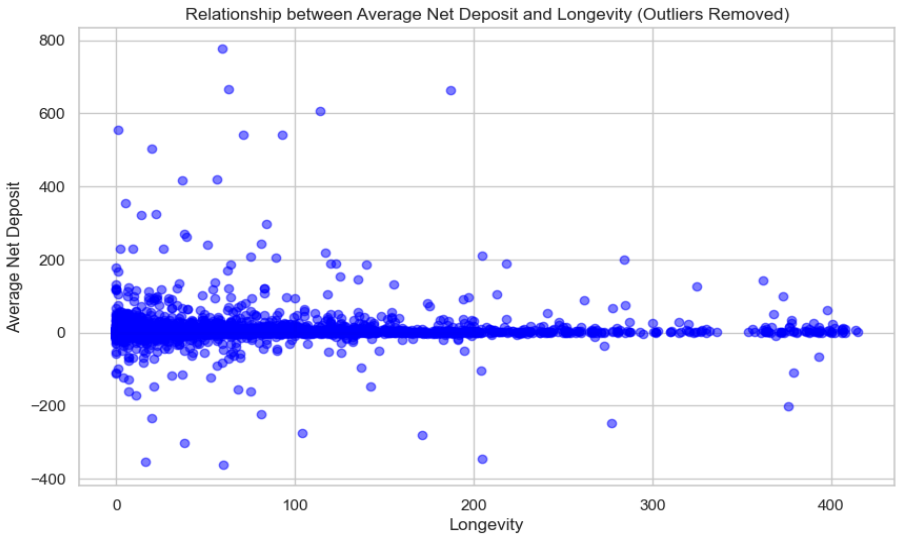


Figure 42. Average Net Deposit and Longevity

Most of the average net deposits are concentrated around zero, creating a negative V-shaped distribution see (Figures 42 and 43). Hence, the Average Net Deposit might not be a good indicator of longevity. Regardless, there are a few interesting insights that could be very helpful.

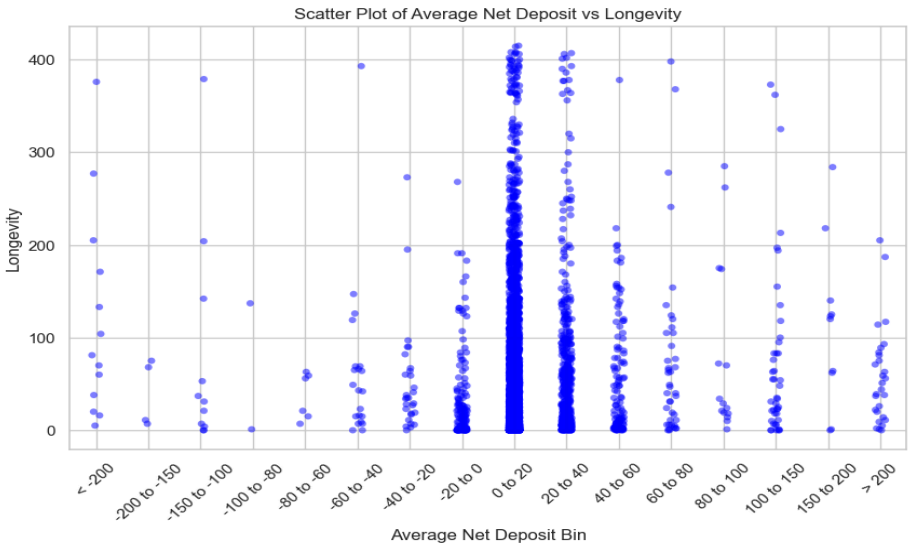


Figure 43. Average Net Deposit and Longevity (Scatter Strip Plot)

Although both sides decreased, the positive net deposits bin experienced a gradual reduction in longevity. The negative net deposit seems to have less high longevity compared to the positive side. The stiff decline in the net deposit indicates that traders make more profit than initial deposit, and tend to cease trading. This is consistent with the point we have mentioned in part 3.5.2.

The net deposits highly distributed around zero reflect a balance between deposits and withdrawals of the traders. This balance might signal a sustainable trading lifestyle where the trader generates enough profit to maintain a positive balance without the need of depositing money. On the other hand, this distribution around zero of lower longevity also suggests that there may be a lack of commitment among traders. Particularly, these traders may constantly withdraw profit as soon as they make any profit, not continue to use the profit to reinvest.

3.5.6.2 Net Deposit Frequency Ratio

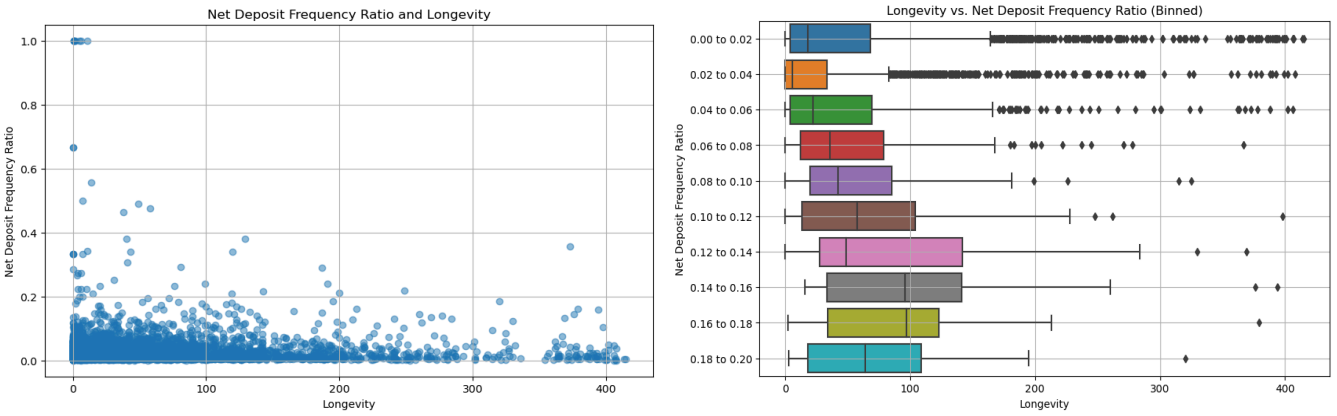


Figure 44. Net Deposit Frequency Ratio and Longevity (Scatter Plot) (left)
Figure 45. Net Deposit Frequency Ratio and Longevity (Box Plot) (right)

By investigating the relationship between how often a trader deposits or withdrawals from their account, however, we get a better picture of the relationship. As demonstrated in figures 44 and 45 above, longevity is loosely but positively correlated with the frequency with which an account makes a deposit or withdrawal. In other words, traders that frequently add funds to their account, likely to immediately or in

the near future execute a trade, and then withdraw profits earned from that transaction tend to have a higher longevity than those that do so less frequently, such as those that add larger amounts upfront to cover their needs for longer. This tendency also suggests that long term traders, whether due to necessity or proclivity, prefer their funds to be immediately accessible to them, rather than secured in a trading account.

3.5.7 | Average Swaps and Longevity

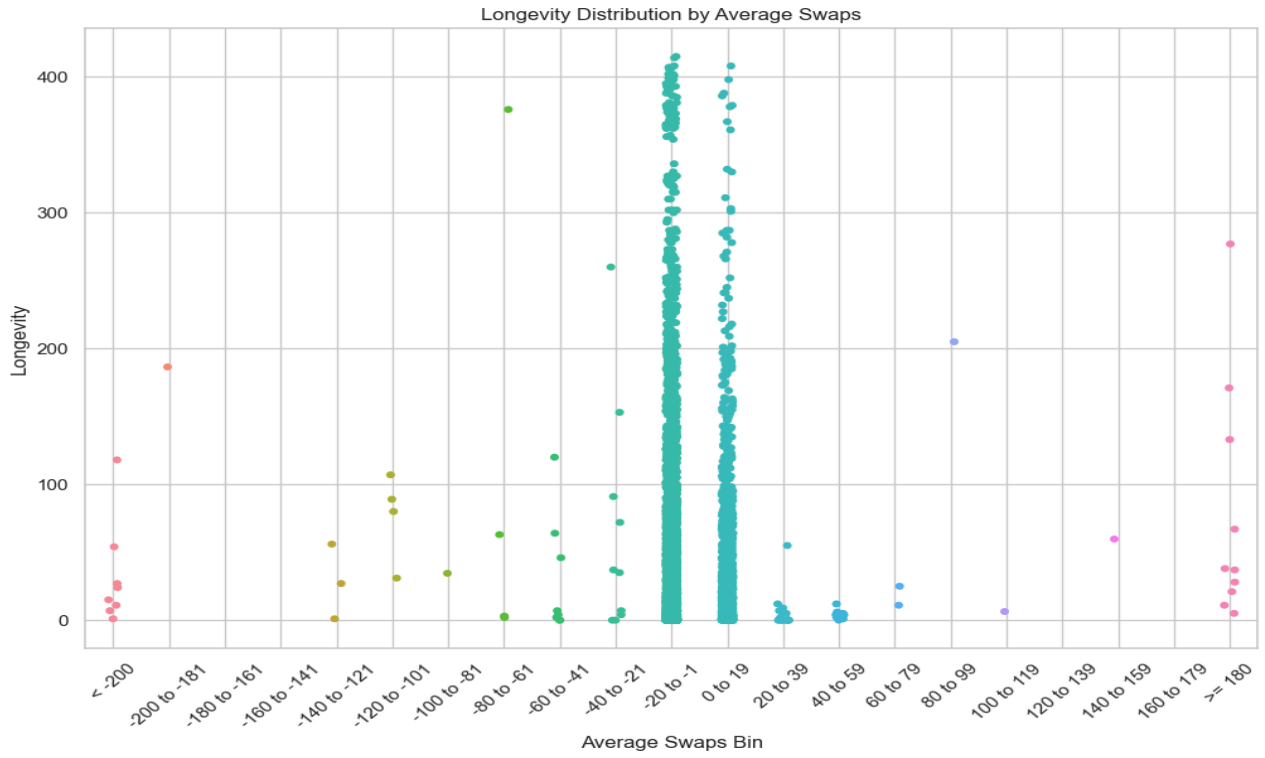


Figure 46. Longevity Distribution by Average Swaps

The average swaps accrued by an account over the durations of its constituent trades have a centralised, if slightly negative, relationship with longevity (see Figures 46 and 47). Except at the extremes, accounts that on average are unaffected or slightly hampered by swap rates tend to use the service for longer than those whose positions are more affected by them. Some accounts achieve this parity through sheer number of trades, which eventually average out, while others tend to make trades that aren't affected or aren't affected severely by the swap rate. Additionally, as usual, the central cluster of observations is home to both high and low longevity accounts, though, proportionally, high longevity traders are more prominent here than elsewhere.

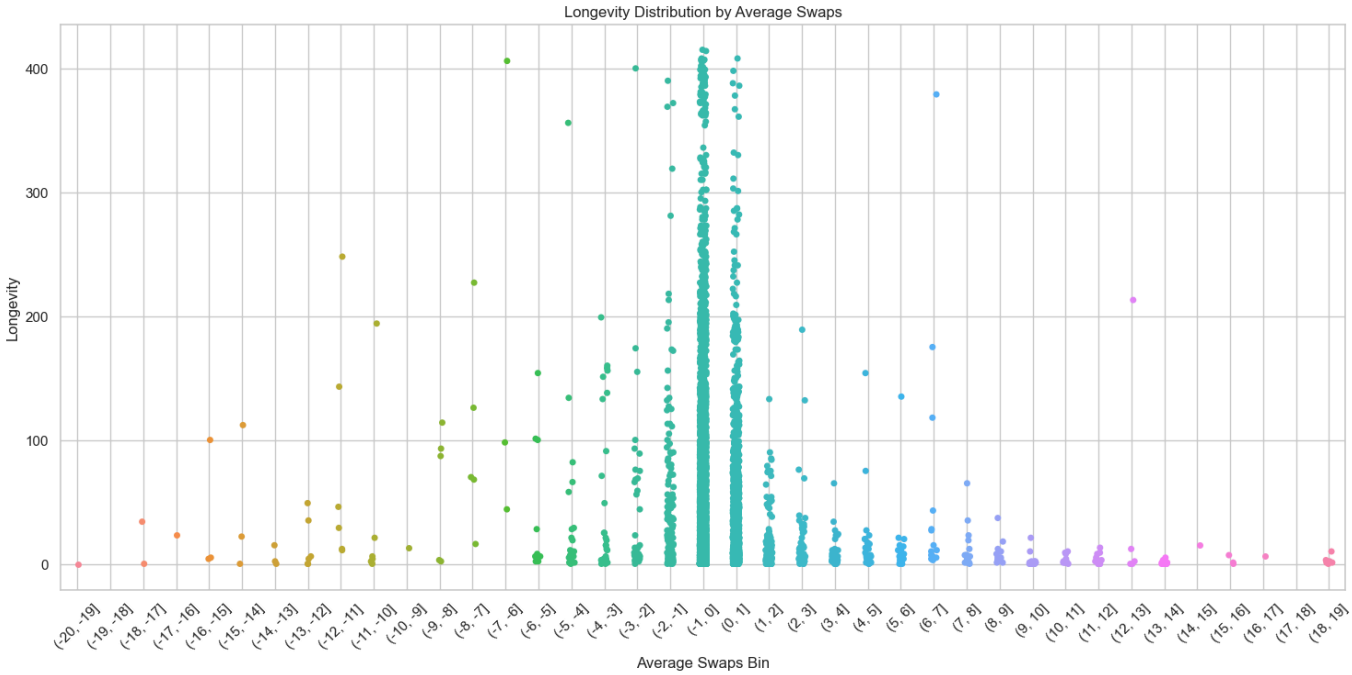


Figure 47. Longevity Distribution by Average Swaps

3.5.8 | Demographics and Longevity

3.5.8.1 | Country

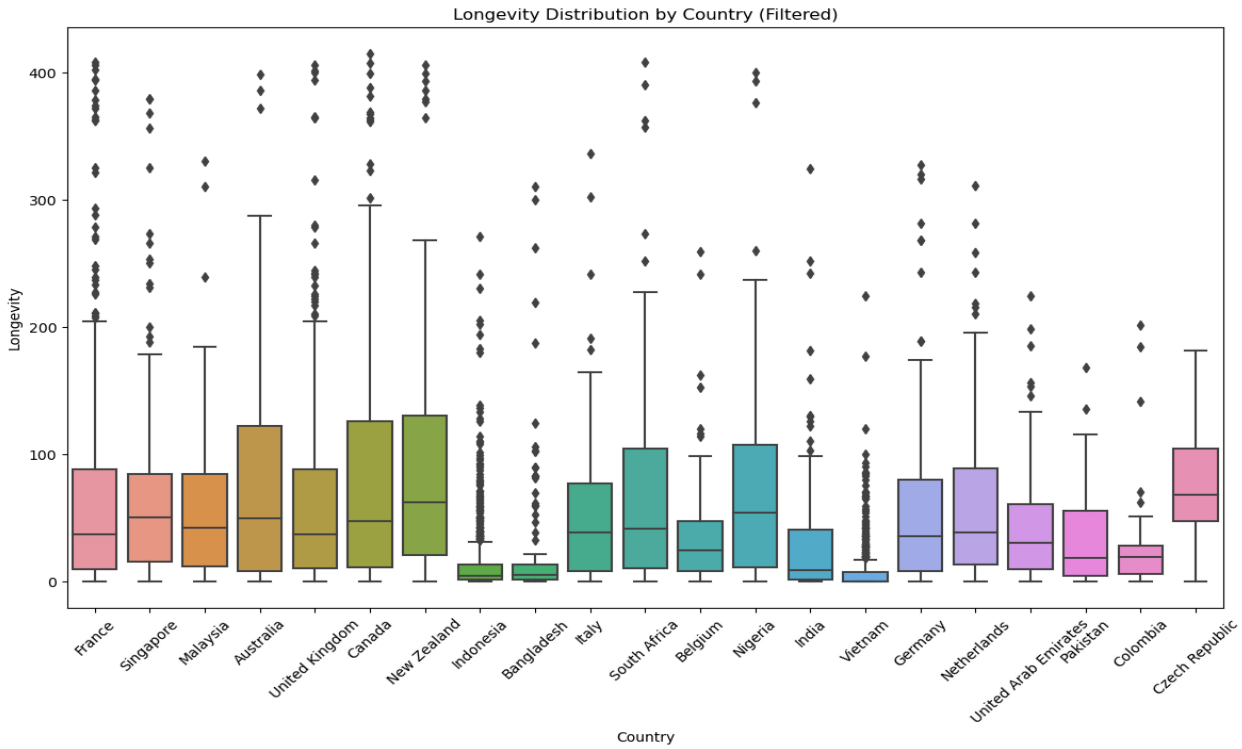


Figure 48. Longevity Distribution by Country

For countries with more than 50 accounts, as seen in Figure 48, several stood out as having a noticeably greater potential for longevity. These were Canada, Australia, and New Zealand due to having both a relatively high average and market share of accounts, with South Africa and Nigeria being honourable mentions. The longevity of accounts from the Czech Republic is notable as well given their high average

despite their smaller size and maximum, but whether this is a coincidence or a genuine pattern of higher longevity is difficult to ascertain without more data.

3.5.8.2 | Account Currency

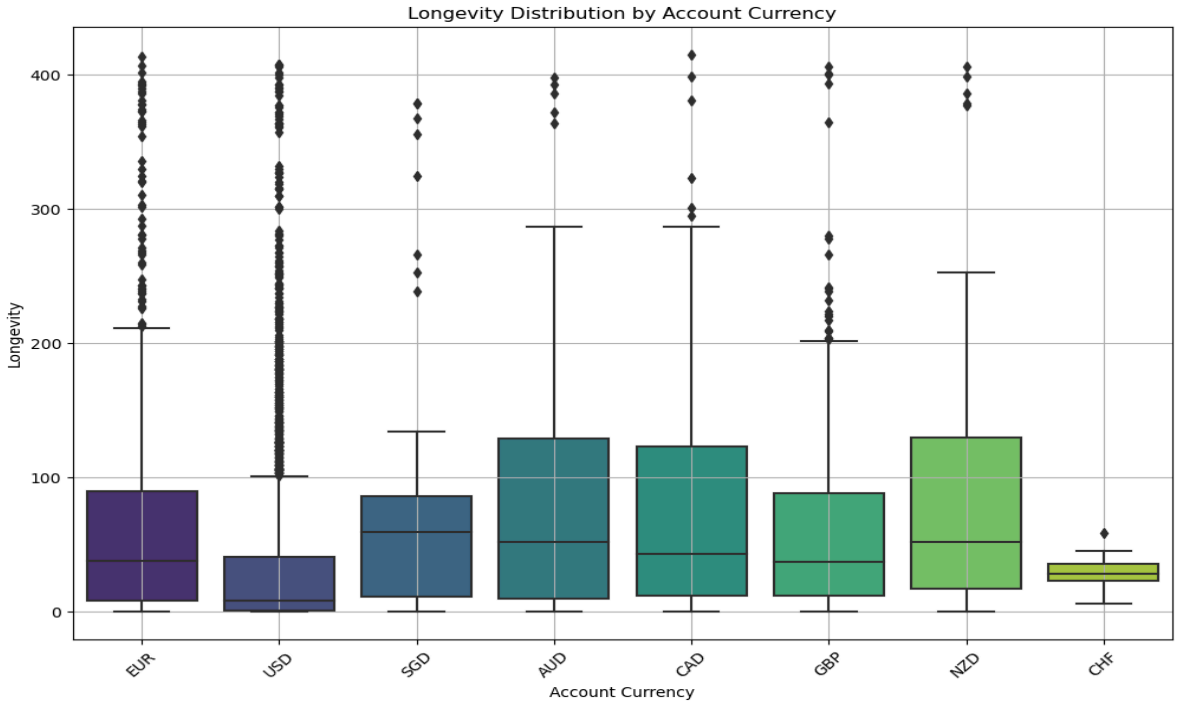


Figure 49. Longevity Distribution by Country

Similarly, accounts using the Canadian, New Zealand, and Australian dollars tend to have higher longevity, whilst the USD's popularity causes it to adopt a lower distribution (see Figure 49).

3.5.9 | Buy/Sell Ratio and Longevity

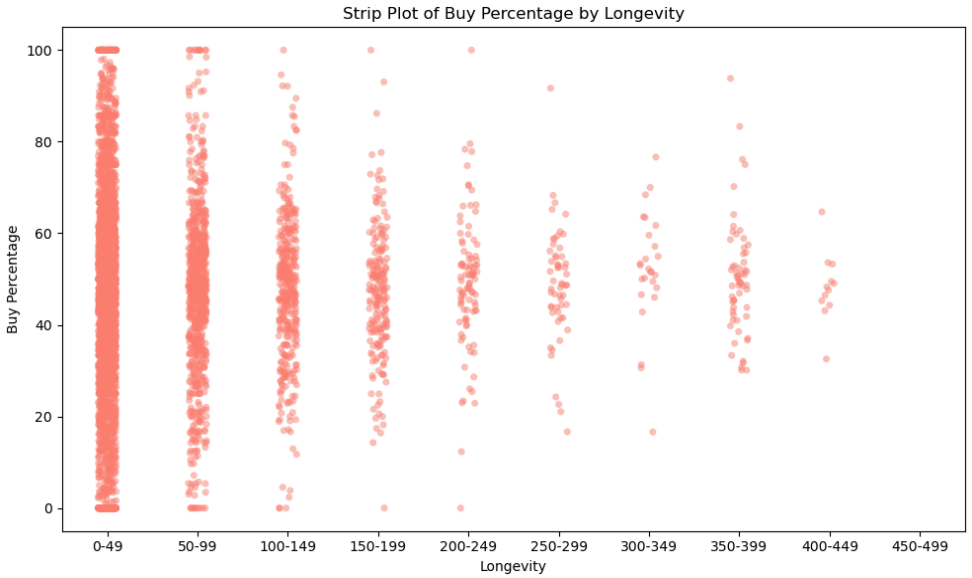


Figure 50. Strip Plot of Buy Perentage by Longevity

Buying seems to be more consistent with high longevity than selling, with little to no representation of high longevity accounts between 0-30% of trades being a buy order (see Figure 50). This data is also

consistent with the notion that many traders, particularly novice ones, are more familiar and comfortable with exclusively taking long positions, rather than a mix of buying and selling. That said, the most reliable indicator in this case is that of a balanced buy sell history, slightly favouring a higher buy percentage. In addition to potentially being a natural market phenomenon, this is consistent with typical algorithmic trading practices, whereby strategies go long on favourable buckets of stocks while going short on unfavourable ones, as well as savvy traders exercising any opportunity they are presented with, rather than solely buying or selling like their counterparts at the extremes.

3.5.10 | Average Volume and Longevity

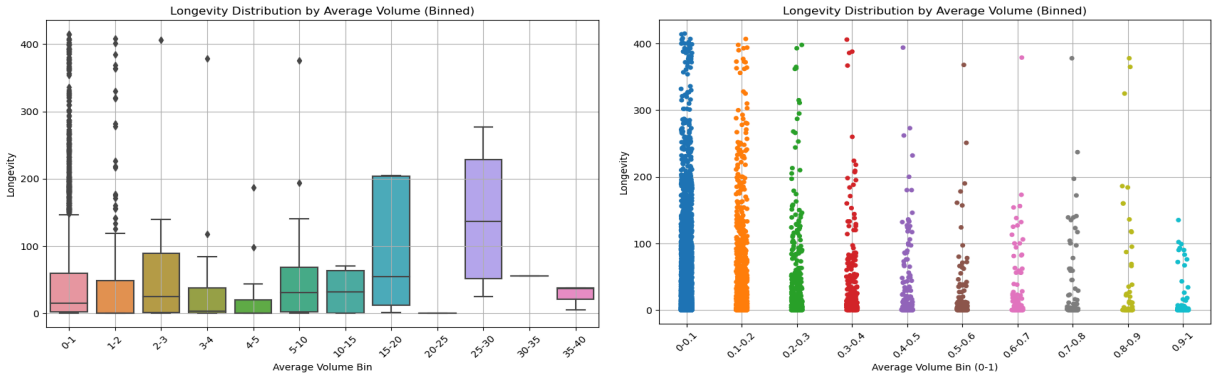


Figure 51. Longevity Distribution by Average Volume (Box Plot) (left)
Figure 52. Longevity Distribution by Average Volume (Strip Plot) (right)

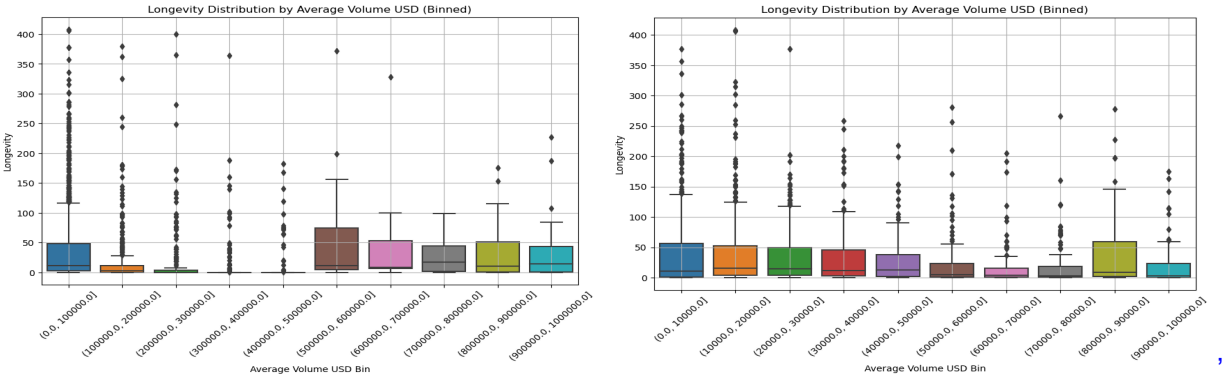


Figure 53. Longevity Distribution by Average Volume (10 bins from 0 - 1,000,000) (left)
Figure 54. Longevity Distribution by Average Volume (10 bins from 0 - 100,000) (right)

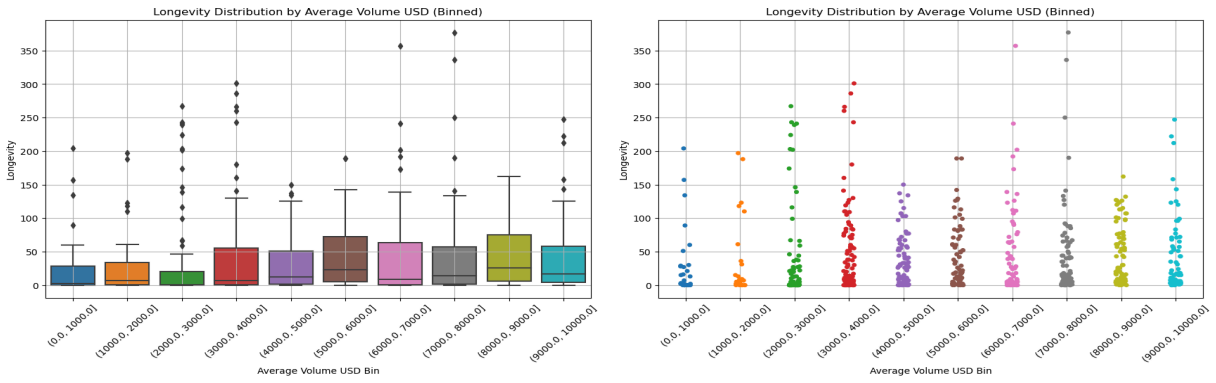
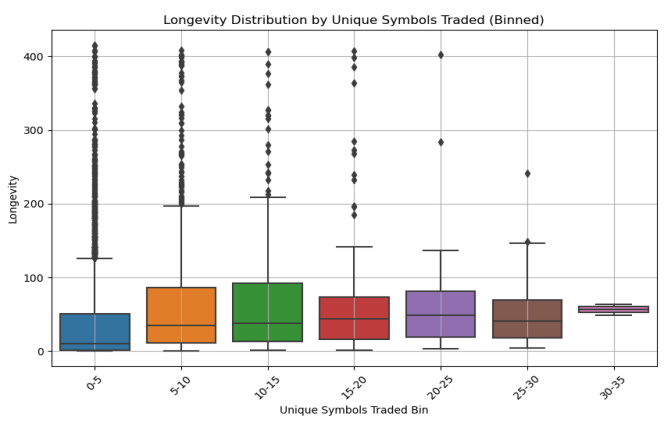


Figure 55. Longevity Distribution by Average Volume (Box Plot, 10 bins from 0 - 10,000) (left)
Figure 56. Longevity Distribution by Average Volume (Strip Plot, 10 bins from 0 - 10,000) (right)

Volume and Volume USD perform similarly against longevity, which is to be expected (see Figures 51, 52, 53, 54, 55, and 56). Most trading, and therefore account averages, take place around the lowest end of the distribution, where high longevity is the most concentrated. From here, these factors share a negative correlation with longevity, until they reach greater magnitudes of size, where the relationship ultimately shifts and becomes positive. Relative to the total sample size, however, there are very few accounts that trade in these upper extremes, and fewer still that have a high longevity to accompany doing so. As a result, the relationship of volume in lot can be thought of as practically flat for the purposes of estimating longevity, while the total notional volume of trades in USD can be positively linear for values less than 10000, negatively linear for those between 10000 and 100000, and then positively correlated again up to values in the hundreds of millions.

3.5.11 | Unique Symbols Traded

The number of unique symbols traded by an account during its lifetime, as an indicator of an at least somewhat diversified strategy, is positively correlated with longevity, at least until around 15, after which observations decrease along with average longevity (see Figure 57).



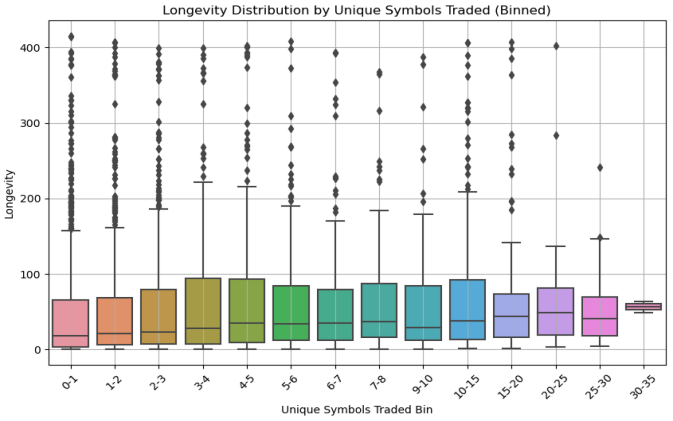


Figure 57. Longevity Distribution by Unique Symbols Traded (7 bins from 0 to 35) (left)
Figure 58. Longevity Distribution by Unique Symbols Traded (7 bins from 0 to 35) (right)

Meanwhile, longevity is relatively equal between accounts with symbols between 1 and 10 (see Figure 58).

3.5.11 | Peak Trading Time

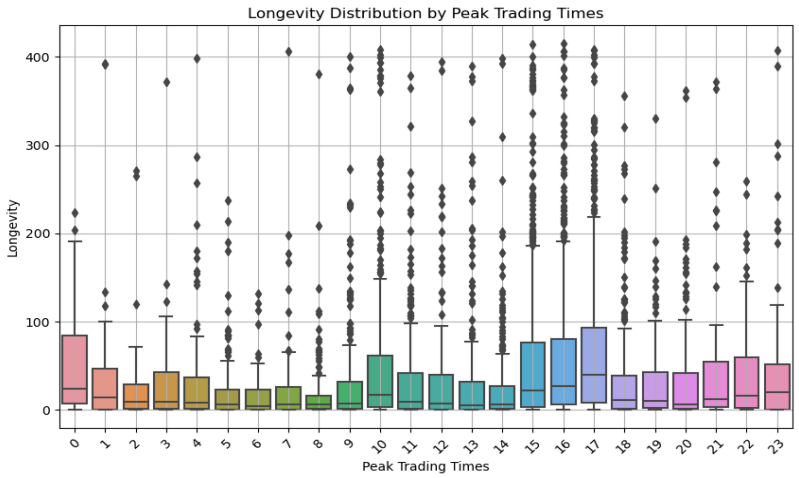


Figure 59. Longevity Distribution by Peak Trading Time (24 hours)

As an aggregation of each account's most popular trading time, an analysis of peak trading time suggests that longer-term traders tend to trade between the hours of 1pm and 5pm, with the periods of 9am to 11am and 9pm to 11pm being the next most popular (see Figure 59).

3.5.12 | Trade Duration

Trade duration is naturally correlated with longevity, in the sense that the ability to make trades of a certain length requires that an account be active be at least that long, usually longer

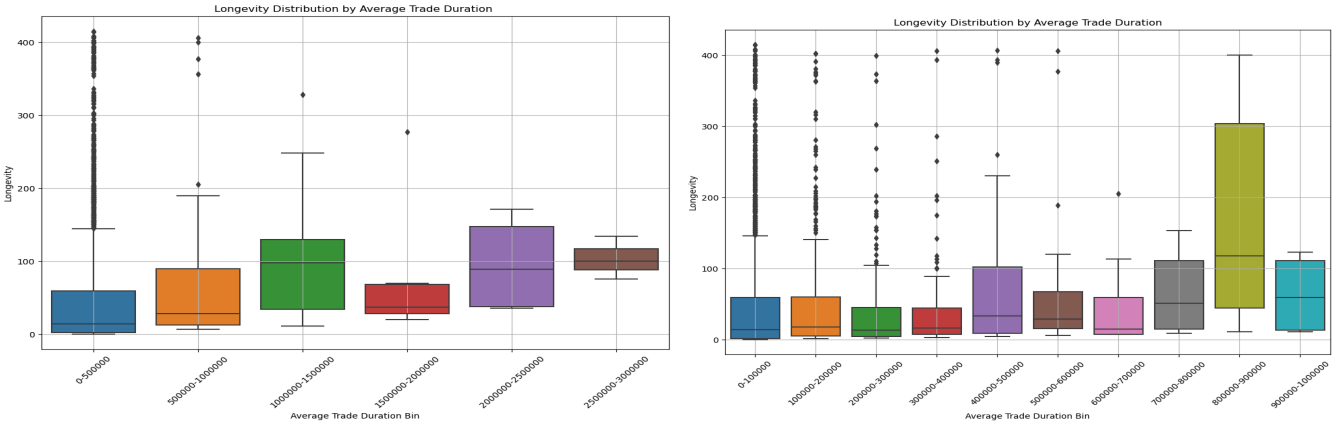


Figure 60. Longevity Distribution by Average Trade Duration (6 bins from 0 to 3,000,000) (left)
Figure 61. Longevity Distribution by Average Trade Duration (10 bins from 0 to 10,000,000) (right)

That said, as an account aggregate, average trade duration also meaningfully serves to characterise accounts by their trading duration tendency more so than their inherent longevity. As such, the fact that average trade duration remains positively correlated (see Figure 60), if more loosely at lower durations (see Figure 61), suggests that accounts that repeatedly make longer duration trades are advantaged in longevity beyond the mere opportunity to do so.

3.5.13 | Credit and Longevity

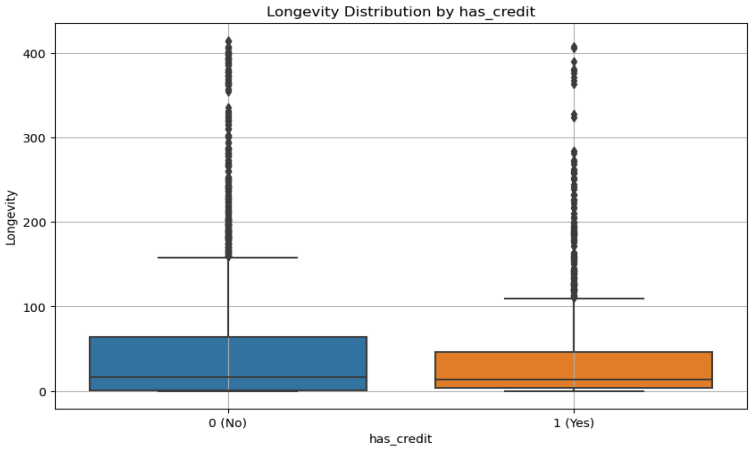


Figure 62. Longevity Distribution by has_credit

Accounts that had been provided credit exhibited a lower overall longevity than those that had never been provided credit, resulting in a negative correlation (see Figure 62). This is surprising, as credit was, appropriately, used as a means of incentivising clients to make trades. While the average longevity of clients provided credit is only slightly shorter than that of their non-incentivised counterparts, around 3 days less, they are also disproportionately less present at higher longevity distributions. That said, we

can also observe that clients provided with credit are also somewhat less clustered around lower longevities, indicated by a slightly higher lower quartile range. As such, it seems that credit remains an effective means of incentivising trades, but as far as incentivising longevity in clients it is only particularly effective in short-lived clients. This is likely because if a client intends to adopt a strategy for long periods, rather than capitalising on an immediate time sensitive opportunity, they will likely be more prepared and have access to the funds they need in advance, either through a successful trading history or ready to go deposits.

4 | Modelling

4.1 | Modelling Approach

This section goes into an in-depth examination of the modelling approach used on the trading dataset with the goal of predicting users' trading longevity. The optimal goal is to create strong regression models that can accurately predict the duration between a user's first and last trade, the core definition of Longevity within this task. Gaining insight into user longevity is substantially helpful for strategic decision-making, as it reveals user behaviour and engagement strategies, thus evaluating the business strategies of the trading management company.

4.1.1 | Synopsis of Modelling Approach

A fundamental part of the modelling strategy started with establishing the target variable of longevity for regression analysis. Longevity was calculated per user, which is represented as y throughout the modelling process. It serves as the lifespan of the user in the trading management company and is a vital metric in understanding what affects user engagement.

4.1.2 | Selection of Regression Models over Classification Models

Initially, predictions were calculated by classification, where longevity was categorised into a range of days (0-30 days, 31-90 days, etc). However, after multiple tests it became obvious that treating outcomes as a smooth spectrum rather than separate groups might manufacture higher accuracy in predictions. Thus, to further optimise higher accuracy in predictions, the choice to pick regression models over classification models was made. This was vital as longevity clearly holds the properties of having a continuous nature and there is a greater need for precise estimation over categorical classification. Furthermore, the logic being that regression models are well-suited for numerical outcome predictions, making them ideal for predicting longevity days. By implementing regression models, there’s a higher chance that precise predictions are developed, a crucial requirement in this data mining problem

4.1.3 | Split Model System

After much trial and error in developing the ideal regression models, a large challenge kept arising of extreme inaccuracies due to a lack of users having a longevity lifespan of greater than 110 days. This affair being the top 10% of users always caused inaccuracies within the models, because of their low population but unique user behaviour per user. Thus, a split model was called into question which involved dividing users into two groups:

- **Bottom 90%** of Users accounting their longevity lifespan (Approximately 0 to 110 days)
- **Top 10%** of Users accounting their longevity lifespan (Approximately 110 to 415 days)

Creating the split model system for the two longevity splits allows to capture unique patterns developed by the two distinct groups. Additionally, it gives more authority to the bottom 90% model, which is less mitigated by constant outliers. The top 10% split also benefits as it isn't negated by the significantly large population of the shorter longevity users. All in all, the split model was perfect to optimise predictive accuracy.

4.1.4 | Regression Models Employed

Several regression models were employed; however, three main models were clearly ahead of the rest. They are all known for being powerful regression models, especially in terms of capturing complex relationships and performing accurate predictions. All three regression models possess their own strengths and weaknesses, thus comparing and contrasting their results is vital to exploring the characteristics of the dataset and the data mining problem at hand. The three being:

4.1.4.1 | Decision Tree Regressor

Description:
Decision tree regression is a method of supervised learning that is non-parametric and utilised for tasks involving both classification and regression. It operates by dividing the input area repeatedly into sections and incorporating a basic model within each section. Decision tree regressors are crucial when the connection between features and the target variable is intricate and not difficult to interpret, making it ideal for the longevity problem.

Strengths:
- Ability to capture non-linear relationships between features and the target variable.
- Easily visualised, allowing for easy interpretability of the decision-making process.
- Handles both numerical and categorical data.

Weaknesses:
- Prone to overfitting, especially with deep trees.
- Can be sensitive to small variations in the training data.

Implementation:
The Decision Tree Regressor was implemented using the base DecisionTreeRegressor class without any specific tuning parameters.

4.1.4.2 | Random Forest Regressor

Description:
Random Forest is an ensemble learning method based on decision trees. It constructs a multitude of decision trees during training and outputs the mean prediction (regression) of the individual trees. Random forest regressors are effective for predicting outcomes when there are many features, and the relationship between features and target variable is complex, resulting in it being paragon for the longevity problem.

Strengths:
- Reduces overfitting compared to single decision trees by averaging predictions from multiple trees.
- Robust to outliers and noise in the data.
- Handles high-dimensional data well.

Weaknesses:

- Less interpretable compared to individual decision trees.
- Computationally more expensive, especially with many trees.

Implementation:
The Random Forest Regressor was implemented using the base RandomForestRegressor class with the criterion set to 'squared_error', which minimises the mean squared error.

4.1.4.3 | XGBoost Regressor

Description:
XGBoost (Extreme Gradient Boosting) is a powerful implementation of gradient boosting algorithms. It builds multiple decision trees sequentially, where each tree corrects the errors of the previous one. It uses a technique called gradient boosting, which focuses on diminishing the loss function. XGBoost is often the model of choice in structured/tabular data problems where the dataset is not too large, and high predictive accuracy is desired. Thus, perfect for the longevity problem.

Strengths:
- Exceptional performance on a wide range of problems.
- Handles missing data well.
- Regularisation techniques to prevent overfitting.

Weaknesses:
- Sensitive to hyperparameter tuning.
- Can be computationally expensive.

Implementation:
The XGBoost Regressor was implemented using the XGBRegressor class. XGBoost is highly tunable. The target variable was set to 'reg:squarederror', indicating squared error regression loss. The 'max_depth', 'eta', 'alpha', and 'reg_lambda' parameters control tree depth, learning rate, L1 and L2 regularisation, respectively. To avoid overfitting, regularisation parameters like 'alpha' and 'reg_lambda' are crucial. Tuning these hyperparameters optimises the trade-off between bias and variance, leading to better model generalisation.

4.1.5 | Alternative Regression Models

Alongside the three main regression models, several alternatives were explored to solve the data mining problem. Although the models weren't considered to be powerful in solving the regression problem, they can be an immensely useful asset in understanding the relationship between the dataset and the models. Ultimately, showcasing strengths and weaknesses of the dataset features. This set of models included:

- MLP Regressor
- Linear Regression
- Ridge Regression
- Lasso Regression
- KNeighbors Regressor
- Support Vector Regressor

Further reinstating the major weaknesses of these models are:
- **Poor Performance:** The alternative models fail to demonstrate competitive performances, plus they lack scalability and robustness to compete with the main models.

- **Complexity and Interpretability:** Models like MLP Regressor and Support Vector Regressor are overly complex and less interpretable compared to the selected models. Their intricate nature could obscure insights and hinder the interpretation of results, making them less suitable for analysis.
- **Model Suitability:** The main models have a superior ability to handle the characteristics of the trading dataset. Their performance, interpretability, and scalability are more optimal with the analysis goals.

While these alternative models were considered for comparison purposes, they were excluded from the focus. This was due to the weaknesses aforementioned related to performance, complexity, and suitability for the task.

4.1.6 | Dataset Partitioning Strategy

As a part of the modelling approach, the trading dataset underwent partitioning into three distinguishable sets. The allocations were:

- **Training Set: 70%**
  - Training involves iteratively adjusting model parameters to minimise errors and enhance predictive accuracy, thus for the large selection for this set.
- **Validation Set: 15%**
  - Crucial in preventing overfitting by providing an independent dataset for evaluating model training.
- **Testing Set 15%**
  - Pivotal in evaluating the final model's performance.

Dividing the dataset into these three sets facilitates the training, validation and testing of models on separate datasets. It is vital especially when performing calculations involving feature importance, metrics, and visualisations on the validation and testing sets. By invoking unseen data, the sets offer a reliable measure of a model's generalisation capability, gauging its effectiveness in real-world scenarios.

4.1.7 | Dataset Modelling Specific Preprocessing

Before the data splits were passed into the model, they underwent some final preprocessing. This firstly involved dropping ‘login’, ‘Total_Trades’, ‘active’, ‘Average_Volume’, ‘longevity’, ‘longevity_bin’. Login was dropped as the variable was just a unique identifier and not relevant to longevity. Total Trades was dropped due to its cumulative and non averaging calculation methods, not suited for modelling. Average_Volume was dropped due to its similarity to Average_Volume_USD. Active, longevity, and longevity bins were dropped due to their relationship to the target variable.

Categorical Encoding: Categorical variables were encoded into numerical values so that they were in a suitable format for all selected models. Initially, one hot encoding was attempted, however this method increased the dimensionality of the dataset immensely, and therefore was not chosen. Label Encoding was the finalised encoding method, which involved converting categories to numbers from 0 - n, where n represented the number of subcategories. This method doesn’t introduce the same dimensionality issues as one hot encoding, but does give an incorrect numerical ordering to subcategories which may affect modelling.
Robust Scaling: The Robust Scaler from sklearn.preprocessing was chosen to scale all non-categorical variables to accommodate for significant outliers that may have still been present in the dataset. This scaling should mitigate the impact of variables of varying ranges on modelling.

4.1.8 | Feature Importance

Feature importance is critical in understanding which factors wield the most influence in determining the longevity of a user in the data mining problem. It is a technique that assigns a score to a data feature, based on the importance they serve in predicting the target variable. So, accordingly, it provides insights into the modelling approach and crucially serves as a guide for feature selection.

In an instance where the most relevant features are highlighted, it assists in increasing the interpretability of the regression model, aiding in focusing on the key variables that solve the data mining problem. On the other hand, by uncovering unusual feature importance values, hidden data or model issues can be found and resolved. All in all, feature importance allows for a greater understanding of where the model may be underperforming.

Additionally, feature importance authorises the comparison of the regression models, facilitating the identification of trends across the different models. Where, if features are consistently important in different models, they clearly must serve a significant impact on the data mining problem.

It is important to note that, it is important to use evaluation metrics to support feature importance for a better comprehensive understanding of model performance. Additionally, in instances where feature importance was unavailable, coefficients were computed to gauge the impact of input variables on the target. However, all three main models allowed for feature importance.

4.1.9 | Evaluation Metrics

Evaluation metrics are vital in judging the performance of regression models. They provide crucial insight into how accurate the model prediction results are compared to the true values. This data mining problem used four key metrics to comprehensively evaluate the effectiveness of predicting longevity with our regression models. These metrics encompass:

- **Mean Absolute Error (MAE):**

  Mean absolute error measures the average absolute difference between the predicted values and the real values and is used to assess the effectiveness of a regression model. It is effectively able to provide an easy to interpret indication of the closeness of the predicted values to the real values.

- **Mean Squared Error (MSE):**

  Mean squared error quantifies the average squared difference between the predicted values and the real values. Squaring the differences penalises larger errors more heavily than smaller ones. Thus, uncovering large discrepancies between predicted and real values.

- **Root Mean Squared Error (RMSE):**

  Root mean squared error is the square root of the MSE and is commonly used because it provides a measure of the average magnitude of the errors in the predicted values in the same units as the target variable. It offers a more interpretable measure compared to MSE, making it easier to understand the magnitude of errors.

- **R-squared (R2)**

R-squared is a measure that provides information about the goodness of fit of the regression model. It is a great statistical measure in understanding how well the regression line approximates the real data.

These aforementioned evaluation metrics comprehensively provide a satisfactory assessment of the accuracy of the regression models. By analysing and comparing these metrics, we gain valuable insights into the quality of the regression models in predicting longevity. Allowing to make educated decisions regarding their performance and potential areas for improvement, thus solving the data mining problem.

4.1.10 | Visualisations

For modelling only one visualisation was significant in assessing the performance of the predictive regression models. This was a scatter plot which was employed to visually compare the true values against the predicted values. A scatter plot is the perfect visualisation tool in this case as it allows readers to quickly identify any discrepancies or patterns in the model's performance.

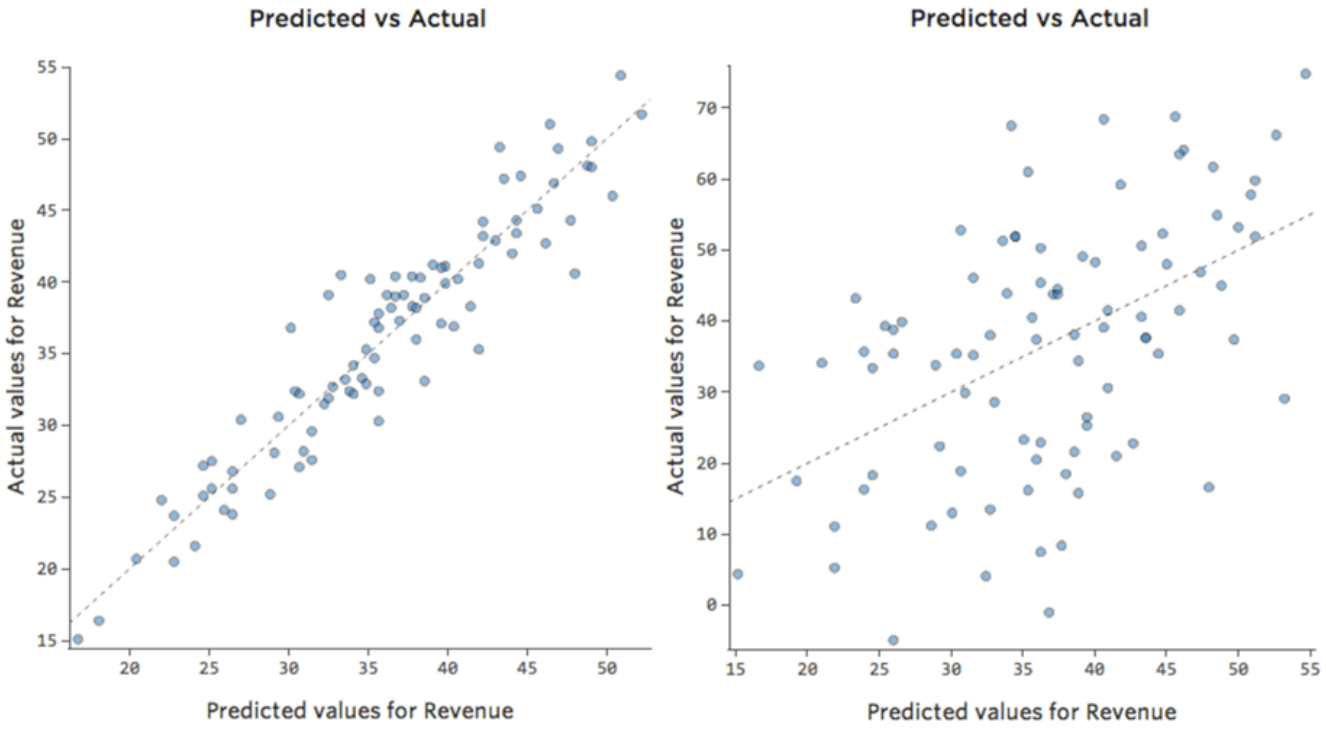


Figure 63. Qualtrics Accuracy Example: Observed vs Predicted

Breaking Figure , it is quickly evident the scatter plot on the left is vastly superior to the scatter plot on the right. This is evident because of factors including:

1. **Linearity:**
   Ideally, the points should form a relatively straight line with a slope of 1, indicating that the model's predictions closely match the actual values. Deviations from this linearity suggest that the model may be under or over-predicting certain values.

2. **Outliers:**
   Identifying outliers in the scatter plot is crucial as they may indicate instances where the model performs poorly. These outliers could be due to anomalies in the data or areas where the model needs improvement.

3. **Clustering:**
   Reading the clustering around the line of perfect predictions allows the reader to understand the accuracy of the model. Where a wider dispersion suggests variability in the model's performance across many data features, whilst a strong following to the line suggest an accurate model.

These three factors alone highlight how easily it is to interpret the difference between the two scatter plots.

Thus, scatter plots are the only employed visualisation in modelling as it allows for easy comparison between plots and aids in qualitative assessment. Authorising the making of educated decisions on refining the model or adjusting their approach to better meet their target of longevity.

4.1.11 | Modelling Approach Summary

To summarise, the extensive modelling approach undertaken to predict users' trading longevity represents a multifaceted effort to leverage advanced regression techniques, comprehensive evaluation methodologies, and insightful visualisations to gain a deep understanding of user behaviour and strategies in the trading domain. The findings and methodologies presented in this analysis provide a solid foundation for strategic decision-making and future research endeavours aimed at enhancing user engagement and profitability in the dynamic landscape of the trading data mining problem.

4.2 | Data Mining Problem

The fundamental business problem of sustaining client longevity and experience in SIGMA trading management translates into a data mining problem by focusing on predicting users' trading longevity. This challenge involves leveraging data mining techniques to extract key findings from the extensive dataset collected by SIGMA trading. Thus, guiding strategic decision-making and enhancing user retention ultimately increasing profit.

The process begins with extensive data preprocessing on customer information, trading information from 2023 and onwards, and transactional data found from daily aggregated reports. This ensures data accuracy and readiness for both analysis and modelling, with users categorised by their trading strategies and behaviours.

Data mining techniques are critical for analysing patterns and trends of users. By applying predictive regression modelling, brokers can forecast customer behaviour and identify factors impacting client longevity. This predictive capability supplies SIGMA with the opportunity to optimise customer experience and profitability.

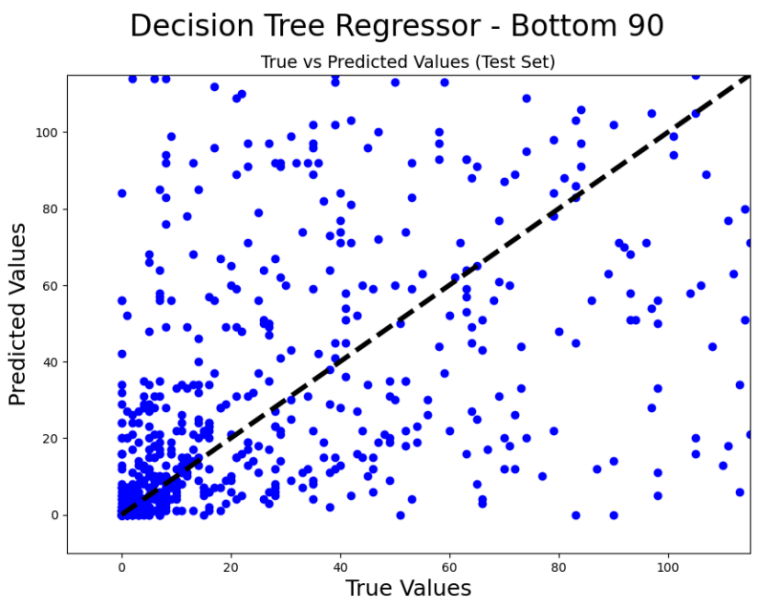
The ultimate goal is to provide brokers with additional leverage through actionable insights. Identifying segments of users with different longevity profiles and uncovering hidden patterns empowers SIGMA trading management to make informed decisions that enhance client engagement and retention.

In summary, the data mining problem involves systematically applying data and modelling techniques to extract meaningful insights from the complex datasets compiled. This process is essential for understanding user behaviours, predicting lifespan and retention, and optimising resources to maximise engagement and longevity. Embracing data mining is crucial for SIGMA trading management to thrive in the emerging machine learning landscape.

4.3 Modelling Results

4.3.1 | Bottom 90% Split

4.3.1.1 | Decision Tree Regressor



Validation	Value
Mean Absolute Error (MAE)	16.9077
Mean Squared Error (MSE)	761.061
Root Mean Squared Error (RMSE)	27.5873
R-squared (R2)	0.2639

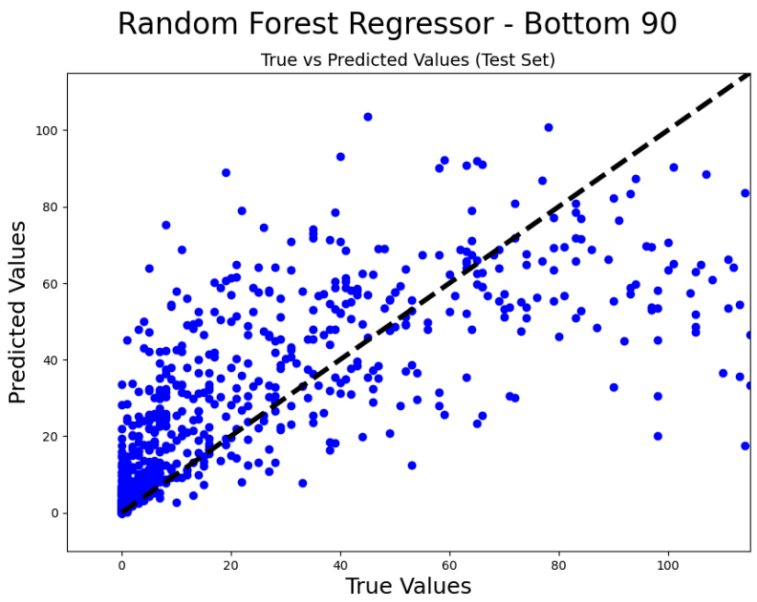
Testing	Value
Mean Absolute Error (MAE)	19.2341
Mean Squared Error (MSE)	976.573
Root Mean Squared Error (RMSE)	31.2502
R-squared (R2)	0.0552

Figure 64. Decision Tree Regressor - Bottom 90 Validation (left)
Figure 65. Decision Tree Regressor - Bottom 90 Testing (right)

As can be seen in Figure 64, the decision tree regressor for the bottom 90% split performs quite poorly in its predictions. There is no demonstration of linearity between the predicted and true values and therefore a significant number of outliers considering the wide range of outcomes. Between 0-10 longevity there is a cluster of predictions that are fairly close to the true values, indicating that the model performs reasonably well at this lower longevity range, however not at higher longevity.

The validation and testing metrics (see Figure 65) further underscore the model's performance issues. The Mean Absolute Error (MAE) is 16.9077 for validation and 19.2341 for testing, Mean Squared Error (MSE) is 761.061 for validation and 976.573 for testing, and Root Mean Squared Error (RMSE) is 27.5973 for validation and 31.2502 for testing. All are higher on the test set compared to the validation set, indicating that the model may be a bit overfitted with a decline in accuracy on unseen data. The lack of linearity in the scatter plot is also reflected in the R Squared metrics (R2) at 0.2639 on the validation set and 0.0552 on the test set, suggesting the model explains very little of the variance and underlying patterns in the data.

4.3.1.2 | Random Forest Regressor



Validation	Value
Mean Absolute Error (MAE)	13.593
Mean Squared Error (MSE)	409.384
Root Mean Squared Error (RMSE)	20.2332
R-squared (R2)	0.604

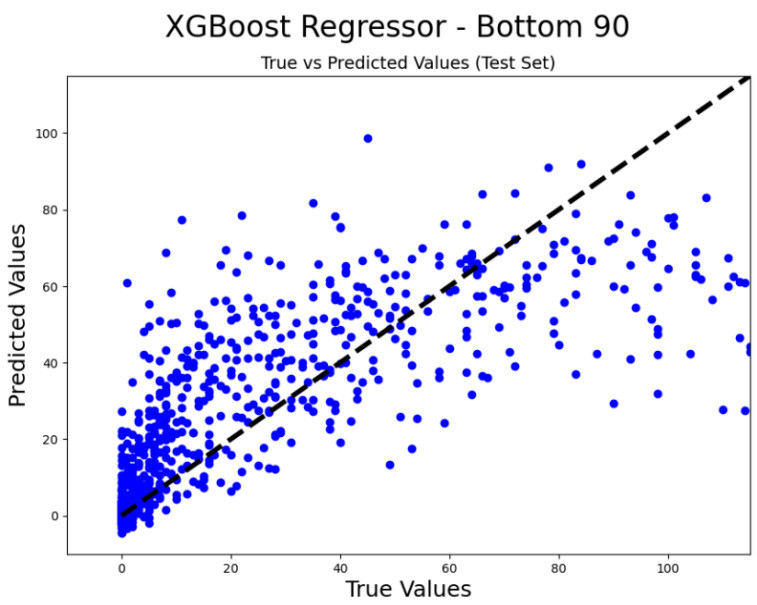
Testing	Value
Mean Absolute Error (MAE)	13.8995
Mean Squared Error (MSE)	453.385
Root Mean Squared Error (RMSE)	21.2928
R-squared (R2)	0.5613

Figure 66. Random Forest Regressor - Bottom 90 Validation (left)
Figure 67. Random Forest Regressor - Bottom 90 Testing (right)

The scatter plot for the Random Forest Regressor for the bottom 90% split (see Figure 66) shows improved linearity compared to the Decision Tree Regressor, with points closer to the dashed line, indicating better predictive accuracy. While there are some outliers particularly among higher true values, their frequency and deviation from the ideal line are reduced. Like the decision tree regressor, the random forest regressor does have a cluster of predictions at the lower range of longevity, however this is less pronounced. These factors extracted from the graph visualisation suggest that the random forest regressor is superior in performance across a much wider range of longevity. However, it should be noted that the model does tend to underpredict towards the upper end of longevity.

The validation and testing metrics for the Random Forest Regressor (see Figure 67) also indicate significant improvement over the Decision Tree Regressor. The Mean Absolute Error (MAE) is 13.593 for validation and 13.8995 for testing, the Mean Squared Error (MSE) is 409.384 for validation and 453.385 for testing, and the Root Mean Squared Error (RMSE) is 20.2332 for validation and 21.2928 for testing. These values are substantially lower and much closer together for the validation and testing sets, suggesting that the model is not overfitted. The R-squared (R^2) values are also higher at 0.604 for validation and 0.5613 for testing, indicating that the Random Forest model explains a much larger proportion of variance in the data.

4.3.1.3 | XGBoost Regressor



Validation	Value
Mean Absolute Error (MAE)	13.3647
Mean Squared Error (MSE)	394.438
Root Mean Squared Error (RMSE)	19.8605
R-squared (R2)	0.6185

Testing	Value
Mean Absolute Error (MAE)	13.3473
Mean Squared Error (MSE)	407.576
Root Mean Squared Error (RMSE)	20.1885
R-squared (R2)	0.6057

Figure 68. Random Forest Regressor - Bottom 90 Validation (left)
Figure 69. Random Forest Regressor - Bottom 90 Testing (right)

The scatter plot for the XGBoost Regressor for the bottom 90% split (see Figure 68) shows a similar moderate linear correlation between true and predicted values, with most points aligning closely with the dashed line, indicating higher predictive accuracy. There are some outliers, increasing at higher longevity ranges, however this is still less than the Decision Tree model. The distribution of clusters is also similar to the Random Forest model, with a smaller cluster at the lowest range of longevity, which evens out as longevity increases. These factors suggest the model has a high robustness in providing more consistent prediction over a wide range of values, similar to the Random Forest model.

The validation and testing metrics for the XGBoost Regressor (see Figure 69) are slightly improved compared to the Random Forest model, with MAE of 13.3647 for validation and 13.3473 for testing. These low error metrics highlight the model's accuracy. The MSE is 394.438 for validation and 407.576 for testing, and the RMSE is 19.8605 for validation and 20.1885 for testing, indicating minimal outlier prediction errors. The R-squared (R^2) values are notably high, with 0.6185 for validation and 0.6057 for testing, also slightly edging out the Random Forest model. This demonstrates that the XGBoost model explains a substantial portion of variance in the data to a slightly higher degree.

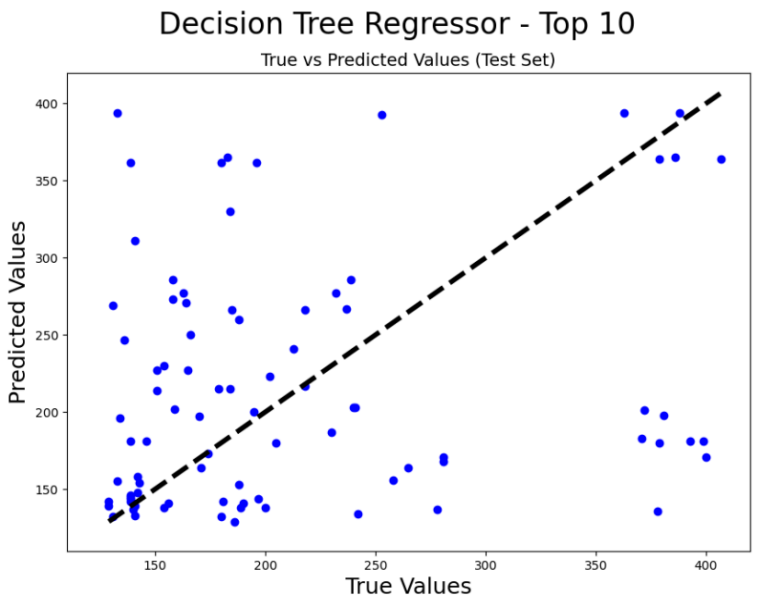
4.3.1.4 | Best Performing Model for the Bottom 90% Model Split

Overall the best performing model for the bottom 90% split is the XGBoost model that demonstrates slightly improved metrics in all categories to the Random Forest model, and significantly improved metrics in all categories to the Decision Tree model. These are 13.3473 MAE, 407.576 MSE, 20.1885 RMSE, and 0.6057 R squared. The MAE of 13.3473 indicates that, on average, predictions are within 13.35 days of the true values, which is about 12.13% of the total longevity range of 110, highlighting the model's precision. The RMSE of 20.1885, representing 18.35% of the range, confirms the model's ability to keep larger errors in check. This value is a bit higher than the MAE, so there still are a small amount of larger errors as can be seen in the scatter plot. And finally, the R^2 value of 0.6057 indicates that the model explains approximately 60.57% of the variance in longevity, demonstrating a reasonable robustness and reliability in capturing the underlying patterns of the data.

These metrics indicate that the best performing model performs reasonably well for the bottom 90% of longevity accounts, however does decrease in performance as longevity increases.

4.3.2 | Top 10% Split

4.3.2.1 | Decision Tree Regressor



Validation	Value
Mean Absolute Error (MAE)	72.593
Mean Squared Error (MSE)	10049.7
Root Mean Squared Error (RMSE)	100.248
R-squared (R2)	-0.3264

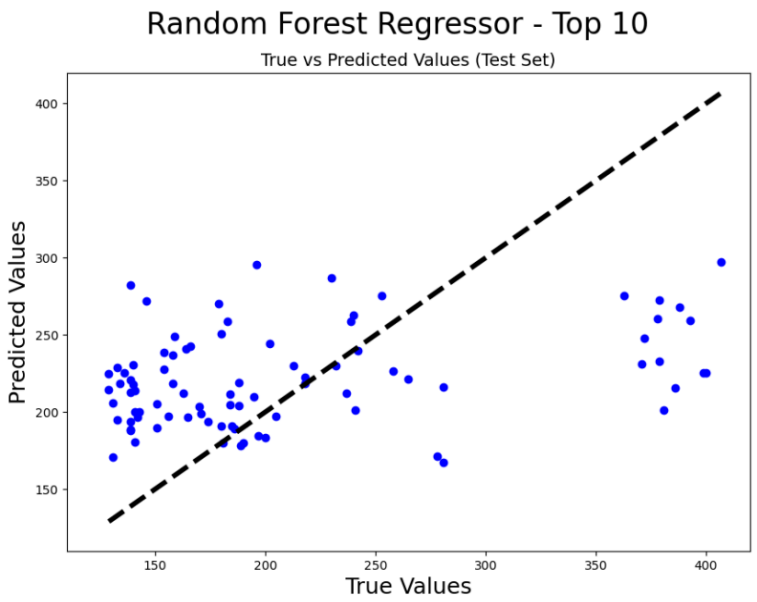
Testing	Value
Mean Absolute Error (MAE)	73.5
Mean Squared Error (MSE)	10165.4
Root Mean Squared Error (RMSE)	100.824
R-squared (R2)	-0.478

Figure 70. Decision Tree Regressor - Bottom 90 Validation (left)
Figure 71. Decision Tree Regressor - Bottom 90 Testing (right)

The scatter plot for the Decision Tree Regressor (see Figure 70), applied to the top 10% longevity accounts, shows a noticeable divergence from the ideal line, indicating significant prediction errors. The points are widely dispersed, especially as true values increase, highlighting the model's struggle to accurately predict higher longevity values. The lack of linearity and presence of numerous outliers suggest that the model is not effectively capturing any underlying patterns for this subset of data.

The performance metrics for the Decision Tree Regressor (see Figure 71) reflect the poor performance visualised in the scatter plot. The MAE is 72.593 for validation and 73.5 for testing, indicating substantial average prediction errors. The RMSE is also high, at 100.248 for validation and 100.824 for testing, reflecting the model's difficulty in managing larger prediction errors. The negative R^2 values of -0.3264 for validation and -0.478 for testing signify that the model performs worse than a simple mean prediction, failing to explain the variance in the data. In comparison to the bottom 90% split, where the Decision Tree model also struggled but with lower error rates and positive R^2 values, the performance here is significantly worse, emphasising the model's inadequacy in predicting the top 10% longevity accounts.

4.3.2.2 | Random Forest Regressor



Validation	Value
Mean Absolute Error (MAE)	69.5802
Mean Squared Error (MSE)	7023.44
Root Mean Squared Error (RMSE)	83.8059
R-squared (R2)	0.073

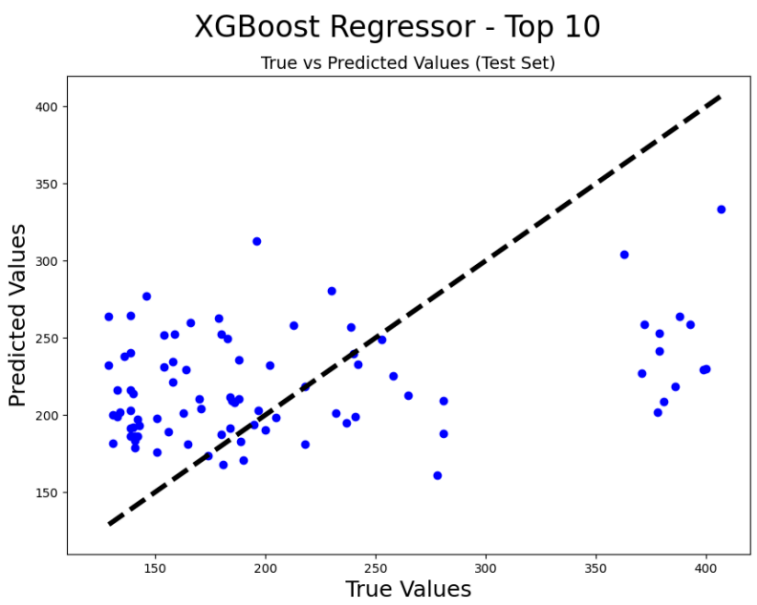
Testing	Value
Mean Absolute Error (MAE)	63.7078
Mean Squared Error (MSE)	6105.4
Root Mean Squared Error (RMSE)	78.137
R-squared (R2)	0.1123

Figure 72. Decision Tree Regressor - Top 10 Validation (left)
Figure 73. Decision Tree Regressor - Top 10 Testing (right)

The scatter plot for the Random Forest Regressor (see Figure 72), applied to the top 10% longevity accounts, shows a significant deviation from the ideal line, indicating notable prediction errors. The points are clustered in certain areas, particularly between true values of 150, 250, and above 350, showing that the model struggles to generalise across the entire range of true values. The linearity is weak, and there are several outliers, suggesting the model's limited effectiveness in accurately predicting higher longevity values for the top 10% of accounts.

The validation and testing metrics for the Random Forest Regressor (see Figure 73) on the top 10% longevity accounts reveal several performance issues. The MAE is 69.5802 for validation and 63.7078 for testing, and the RMSE is 83.8059 for validation and 78.137 for testing. These high error values indicate substantial prediction inaccuracies. The R^2 values are 0.073 for validation and 0.1123 for testing, suggesting that the model explains very little of the variance in the data. Compared to the bottom 90% split, where the Random Forest model showed much lower error rates and higher R^2 values, the performance for the top 10% split is significantly worse, emphasising the model's inadequacy in handling this subset of data.

4.3.2.3 | XGBoost Regressor



Validation	Value
Mean Absolute Error (MAE)	70.2343
Mean Squared Error (MSE)	7120.28
Root Mean Squared Error (RMSE)	84.3817
R-squared (R2)	0.0602

Testing	Value
Mean Absolute Error (MAE)	63.7532
Mean Squared Error (MSE)	6200.85
Root Mean Squared Error (RMSE)	78.7455
R-squared (R2)	0.0984

Figure 74. Decision Tree Regressor - Top 10 Validation (left)
Figure 75. Decision Tree Regressor - Top 10 Testing (right)

The scatter plot for the XGBoost Regressor (see Figure 74) on the top 10% longevity accounts shows significant prediction errors, with points dispersed widely around the ideal line. While there are clusters between true values of 150 and 250, and above 350, the model struggles to generalise across the entire range, similar to the Random Forest Regressor. The weak linearity and numerous outliers suggest that the model faces challenges in accurately predicting higher longevity values in this subset.

The validation and testing metrics for the XGBoost Regressor (see Figure 75) on the top 10% longevity accounts are as follows: MAE is 70.2343 for validation and 63.7532 for testing, with RMSE of 84.3817 for validation and 78.7455 for testing. These metrics are comparable to those of the Random Forest Regressor for the same subset, however showing slightly high error rates. The R^2 values are 0.0602 for validation and 0.0984 for testing, indicating the model explains very little variance. In comparison to the XGBoost Regressor's performance on the bottom 90% split, where it demonstrated low error rates and high R^2 values, the performance for the top 10% split is significantly worse, emphasising the difficulty in accurately predicting this subset.

4.3.2.4 | Best Performing Model for the Bottom 90% Model Split

Overall, the best performing model for the top 10% split is the Random Forest Regressor, but it falls short compared to the XGBoost model's performance on the bottom 90% split. For the top 10% split, the Random Forest Regressor's metrics are as follows: 63.7078 MAE, 6105.4 MSE, 78.137 RMSE, and 0.1123 R-squared. The MAE of 63.7078 indicates that, on average, predictions are within 63.71 days of the true values, which is significantly higher than the MAE for the XGBoost model on the bottom 90%, where predictions were within 13.35 days. This suggests that the Random Forest model lacks precision for the top 10% subset. The RMSE of 78.137, which represents a larger portion of the longevity range, confirms the model's struggle with larger prediction errors. The R^2 value of 0.1123 indicates that the model explains only 11.23% of the variance in longevity for the top 10% subset, compared to 60.57% for the XGBoost model on the bottom 90%.

These metrics indicate that while the Random Forest Regressor is the best performing model for the top 10% split, it is not adequate for this subset when compared to the XGBoost model's performance on the

bottom 90% split. The significant drop in precision and increase in error rates highlight the model's inadequacy in capturing the underlying patterns of the data for higher longevity accounts.

4.3.3 | Best Performing Model Feature Importance Extraction

4.3.3.1 | XGBoost Regressor Bottom 90% Split

Sorted:

Feature	Importance
TP/SL Hit Ratio	0.228337
Average_Swaps	0.0788747
Ratio_Profitable_Trades	0.0709618
Trading_Frequency	0.0705397
Buy_Percentage	0.064017
Trading_Method	0.0639011
Average_Profit	0.0523329
Average_DPM	0.0496698
country	0.0486316
average_net_deposit	0.0442499
Profit_Loss_Variability	0.0418685
Average_Trade_Duration	0.027824
net_deposit_frequency_ratio	0.0268137
Peak_Trading_Times	0.0231574
Unique_Symbols_Traded	0.0228409
Reward_Risk_Ratio	0.0211816
Average_Volume_USD	0.0209346
account_currency	0.0152497
has_credit	0.0143914
Average_Commission	0.0142231

Figure 76. XGBoost Regressor Bottom 90% Split ‘s Feature Importance

Extracting feature importance from the top performing model on the 90% split helps reveal key insights into both factors influencing the model’s predictions as well as factors influencing longevity in general (see Figure 76). The most important feature is the TP/SL Hit Ratio, with a high importance score of 0.228337, indicating that the ratio of take-profit to stop-loss hits significantly impacts the model's accuracy. This is followed by Average_Swaps (0.0788747), Ratio_Profitable_Trades (0.0709618), and Trading_Frequency (0.0705397), suggesting that the frequency and profitability of trades, as well as the cost of swaps, are also crucial. Other notable features include Buy_Percentage (0.064017) and Trading_Method (0.0639011), highlighting the importance of trading strategies and behaviours. Average_Profit (0.0523329) and Average_DPM (0.0496698) also play significant roles, indicating that consistent profitability and daily performance metrics are essential. Ranking these features as the XGBoost model has done, provides a comprehensive understanding of diverse and important trading features that differentiate longevity for accounts with life spans between 0-110 days.

4.3.3.2 | Random Forest Regressor Top 10% Split

Sorted:

Feature	Importance
Average_Volume_USD	0.0844325
Average_Trade_Duration	0.0825748
net_deposit_frequency_ratio	0.0777481
TP/SL Hit Ratio	0.071528
Reward_Risk_Ratio	0.0714847
Average_Commission	0.0659334
Average_DPM	0.0603108
Average_Swaps	0.0586937
Unique_Symbols_Traded	0.054617
Buy_Percentage	0.0531182
Trading_Frequency	0.0514871
Ratio_Profitable_Trades	0.046455
Average_Profit	0.0399249
Profit_Loss_Variability	0.0385884
average_net_deposit	0.0380322
country	0.0351194
Peak_Trading_Times	0.0261455
Trading_Method	0.0245896
account_currency	0.0133188
has_credit	0.00589816

Figure 77. Random Forest Regressor Top 10% Split's Feature Importance

The feature importance analysis for the Random Forest model on the top 10% split identifies different key factors compared to the best performing XGBoost model for the bottom 90% (see Figure 77). The top features for the top 10% subset include Average_Volume_USD (0.0844325), Average_Trade_Duration (0.0825748), and net_deposit_frequency_ratio (0.0777481). These features differ significantly from the bottom 90% model, where the TP/SL Hit Ratio was the most important. The importance of Average_Volume_USD and Average_Trade_Duration suggests that the overall trading volume and duration of trades are critical. Other notable features for the top 10% include Reward_Risk_Ratio (0.0714847), Average_Commission (0.0659334), and Average_DPM (0.0603108), indicating that risk management, trading costs, and daily performance metrics are vital. This contrasts with the bottom 90% model, which placed higher importance on features like Trading_Frequency and Buy_Percentage. However, despite identifying these features, the Random Forest model did not perform well enough on the top 10% accounts to justify these as reliably relevant to predictions, given the model's high error rates and low R^2 values.

4.3.4 | Relation To Original Business Problem

The core target was to provide actionable insights into client longevity, leveraging data analytics to improve SIGMA's strategies, which is consummated by utilising modelling. By analysing client behaviours and categorising them according to trading styles, habits, and approaches, we aimed to reveal potential patterns that influence client retention. This analysis led to the development of various

predictive models to determine factors affecting client longevity and to identify clients who are likely to remain engaged over long periods.

4.3.4.1 | Comparison of Model Performance

Our modelling efforts focused on two primary subsets of clients: the bottom 90% in terms of longevity and the top 10%. We employed Decision Tree, Random Forest, and XGBoost regressors to predict client longevity, aiming to provide SIGMA with the most reliable models for both segments.

Bottom 90% of Longevity Accounts
The XGBoost model emerged as the best performer for the Bottom 90% of accounts, showing superior metrics across all categories compared to the other models. The metrics of the Bottom 90% models indicate that the XGBoost model can predict client longevity with reasonable precision, with the average prediction error being within 13.35 days, representing about 12.13% of the total longevity range of 110 days. This level of accuracy provides SIGMA with a reliable tool to predict client behaviour and tailor engagement strategies accordingly.

Top 10% of Longevity Accounts
For the Top 10% of clients, the Random Forest model was the best performing, but it still fell far short of the accuracy achieved by any of the Bottom 90% models. The metrics of the Top 10% models reveal significant prediction inaccuracies, with the MAE indicating that predictions are, on average, within 63.71 days of the true values. The RMSE of 78.137 and the R^2 of 0.1123 underscore the model's limited ability to explain the variance in longevity for this subset, highlighting the complexity and variability in predicting the behaviours of these long-term clients.

4.3.4.2 | Key Implications

The contrasting performance of models between the bottom 90% and the top 10% of clients suggests key implications for the SIGMA trading management:

1. The more accurate XGBoost model for the bottom 90% can help SIGMA identify which clients are likely to have shorter trading relationships and develop targeted interventions to extend their engagement. By understanding the key features influencing this group's behaviour, such as the TP/SL Hit Ratio and Trading Frequency, SIGMA can refine its engagement tactics.

2. The lower accuracy of models for the top 10% highlights the difficulty in predicting behaviours of long-term clients. This suggests that long-term engagement may be influenced by more complex and less quantifiable factors. SIGMA might need to complement predictive models with qualitative insights, such as direct client feedback and more personalised service approaches, to better retain these high-value clients.

3. By effectively using the XGBoost model's predictions, SIGMA can prioritise resources towards clients with high potential for longevity in the bottom 90%, enhancing retention rates. For the top 10%, SIGMA should consider investing in more personalised relationship-building efforts, acknowledging the limitations of current predictive models.

Through the thorough modelling endeavour SIGMA can formulate various strategies to enhance user experience and longevity strategies. The regression models like XGBoost show promise for 90% of the user base, while the Top 10% models may present challenges in prediction accuracy, the comprehensive insights provided by the Bottom 90% models empower SIGMA to tailor their user engagement strategies effectively. Thus, highlighting the significance of leveraging modelling to create actionable insights and ultimately enhance user longevity.

5 | Findings

5.1 | Analysis

Addressing the challenge of maintaining client engagement and longevity is paramount for SIGMA Trading Management. At Sevenfold Consultants, our analysis delved into the factors influencing trader longevity, aiming to provide actionable insights to enhance customer retention and engagement.

Our investigation into trading frequency, illustrated in Figures 41 and 42, revealed that higher daily trading frequencies are inversely related to trader longevity. Traders who engage in numerous trades daily often exhibit riskier behaviour and shorter lifespans in the market. Conversely, those with a reduced trading frequency tend to exhibit patience and adherence to their trading strategies, resulting in longer-term engagement. Specifically, the optimal trading frequency for longevity appears to be less than 20 trades per day, with those averaging fewer than 10 trades consistently exhibiting longer trading relationships. This finding aligns with Menkhoff et al.'s (2008) research on fund managers, which underscores the risks associated with high-frequency trading and the benefits of a more measured approach.

Profitability analysis, as seen in Figures 45 and 46, shows that higher profitability does not necessarily correlate with longer longevity. Traders experiencing slight losses, specifically in the range of -100 to 0, tend to stay active longer than those with significant losses. This suggests that moderate, consistent performance is more sustainable than extreme profitability. Supporting this, Ma et al.'s (2022) study found a V-shaped relationship between profitability and longevity, with both the least and most profitable traders more likely to exit the market early.

Further examination of profitability through Average DPM (Figures 47 and 48) indicated that stability, rather than high profitability, is associated with longer trading lifespans. Lower average DPM suggests that traders prioritising safety and stability over excessive returns tend to survive longer in the market. Similarly, the ratio of profitable trades (Figures 49 and 50) revealed that a winning ratio between 40% and 70% is optimal for sustainable trading. Traders with excessively high win ratios often rely on luck and are less likely to maintain long-term engagement.

The analysis of profit/loss variability (Figures 51 and 52) showed that higher variability is associated with lower longevity. Traders with more conservative, consistent returns tend to remain active longer, highlighting the importance of promoting stable trading practices. Additionally, trade duration analysis (Figures 82 and 84) indicated that longer trade durations correlate with longer trading lifespans, suggesting that clients engaging in longer-term trades are more committed and strategic in their approach.

Examining trading methods (Figures 47 and 3.5.3.1.2) revealed that the Expert Trading Method, characterised by access to advanced advice and trading models, demonstrated the highest average longevity. This implies that providing clients with high-quality resources and support can significantly enhance their trading lifespan. Conversely, the Client Trading Method, with the lowest average longevity, underscores the need for tailored guidance and support to improve retention in this segment.

Trading commissions (Figures 48 and 49) emerged as another critical factor. Lower commissions are associated with longer market survival, likely because they reduce the cost burden on traders and enhance overall profitability. Encouraging clients to choose assets with lower costs or tighter spreads can thus foster longer-term engagement.

The TP/SL hit ratio (Figures 51 and 52) demonstrated a strong correlation between hitting Take Profit (TP) targets and trader longevity. Traders who frequently close positions at TP levels exhibit greater longevity than those hitting Stop Loss (SL) limits or using neither measure. This finding suggests that prudent risk management practices, such as setting and adhering to TP/SL limits, are vital for sustaining trading activity.

Net deposit patterns (Figures 53 and 54) further illustrated sustainable trading behaviours. Traders maintaining a balance between deposits and withdrawals, with net deposits around zero, tended to have longer trading relationships. This balance reflects a sustainable trading lifestyle, where traders generate enough profit to maintain a positive balance without needing frequent additional deposits. It also suggests that quick profit withdrawals may indicate a lack of long-term commitment.

The analysis of average swaps (Figures 57 and 58) showed that accounts with average swaps close to zero tend to have higher longevity. Moderate swap rates contribute to sustained trading activity, suggesting that minimising swap costs can enhance client retention.

The buy/sell ratio (Figure 61) revealed that a balanced buy/sell ratio slightly favouring buy orders correlates with higher longevity. This indicates that traders more comfortable with long positions, or those employing balanced strategies, tend to remain active longer. Encouraging balanced trading strategies could therefore improve client engagement.

Trading volume (Figures 63, 65, and 67) demonstrated that most trading occurs at lower volumes, where high longevity is concentrated. The relationship between volume and longevity becomes more complex at higher volumes, suggesting that volume alone is not a clear predictor of longevity. However, promoting moderate trading volumes might support sustained engagement.

Diversification in traded symbols (Figures 68 and 69) also positively correlates with longevity up to a certain point. Traders engaging with a broader range of symbols exhibit longer trading lifespans, although this effect diminishes beyond a certain level of diversification. Encouraging diversification within optimal limits can therefore enhance client retention.

Peak trading times (Figure 71) were found to be between 1pm and 5pm, with secondary peaks in the morning and late evening. This suggests that longer-term traders prefer these trading windows, possibly due to market conditions or personal schedules. Understanding and accommodating these preferences can help tailor services to client needs.

Finally, credit provided to accounts (Figure 85) exhibited a negative correlation with longevity. While credit effectively incentivizes short-term trades, it appears less effective for promoting long-term engagement. Clients intending to adopt long-term strategies are likely to have access to sufficient funds without needing additional credit, highlighting the importance of tailoring credit offerings to client needs and trading behaviours.

Promoting disciplined and strategic trading practices, providing high-quality resources and support, and encouraging sustainable trading behaviours are key to retaining clients. By understanding and leveraging these insights, SIGMA can optimise its marketing strategies, risk management, and client services to build enduring relationships with traders.

5.2 | Modelling

The analysis reveals distinct differences in model performance between the bottom 90% and the top 10% of client accounts. The XGBoost model’s superior performance in predicting the bottom 90% of client longevity indicates that trading behaviours and strategies significantly influence client retention.

Features such as the TP/SL Hit Ratio and Trading Frequency are crucial, suggesting that clients who manage their trades effectively and engage frequently tend to have longer trading relationships.

In contrast, the predictive models struggled with the top 10% of clients, where the Random Forest model, despite being the best in this subset, still showed high error rates and low explanatory power. This suggests that the longevity of high-value clients is influenced by more complex factors that are not fully captured by the available data or current modelling techniques. These factors may include personal motivations, external market conditions, or other qualitative aspects that require a different analytical approach.

For SIGMA, these findings imply a dual strategy: leveraging the XGBoost model to enhance retention strategies for the majority of clients by focusing on key trading behaviours, and adopting more personalised, qualitative approaches for the top 10% of clients to address the nuanced factors influencing their long-term engagement. Additionally, while factors affecting client longevity are varied and often subtle, monitoring the top 5 variables for each segment split provides a good starting point for determining when intervention may be required to increase client longevity.

6 | Recommendations

Based on our findings, we have a slew of recommendations for addressing the business problem.

Firstly, when making considerations for optimising marketing strategies to attract clients with high longevity tendencies, we'd recommend focusing on key categorical variables. Clients that are predominately algorithmic or mobile traders have higher longevity on average, so marketing that highlights these trading methods or is positioned to reach users of these avenues would help to cultivate a higher longevity consumer base.

Similarly, we'd recommend focusing marketing at least somewhat disproportionately towards the Canadian, New Zealand, and Australian markets, based primarily on the national currency of these countries (CAD, NZD, and AUD respectively) exhibiting slightly higher average longevity tendencies than others.

That said, it is difficult to make confident recommendations regarding marketing, such as those above, without considering the efficacy and viability of doing so. While our findings suggest a focus on these features would help create a high-longevity client base, deploying a campaign based on them would require a comprehensive cost-benefit analysis.

Then, as far as identifying and boosting longevity in new and existing clients, we'd broadly recommend monitoring their activity based on the key variables identified and promoting conservative trading practices using marketing, recommendations, and incentives. In particular, due to the nature of credit as a longevity indicator, we'd suggest providing clients, namely low longevity clients identified by exhibiting patterns at the extremes, with credit as an incentive to adopt a more conservative position, such as purchasing a diverse asset or one that typically requires a longer holding period to realise a return.

In the same vein, promoting or educating users in the application of effective take profit and stop loss limits should promote greater longevity. This is especially useful for addressing the issue of low longevity accounts, who tend to operate at the extremes of both account and trade metrics. By incentivising lower risk, longer term, more conservative positions, clients who are turned away from the service due to immediate short term gains or losses may be converted to longer term users, leading to further opportunities for conversion and incentivisation.

In practice, for example, this might involve providing a new trader who has recently made multiple profitable, foreign exchange trades in a short period by shorting credit to make further trades, on the stipulation that part or all of the money be put towards taking a position in one or more indices of their choice, or your recommendation. Alternatively, a client that has not seen success taking short almost exclusively short positions could be incentivised through credit and recommendation to try taking a long position of their choice, so long as they include reasonable take profit and stop loss limits.

Finally, as suggested by the findings, the behaviours and priorities of extremely high longevity traders differ substantially from their lower longevity counterparts. As a result, the more general suggestions provided thus far are likely to be more effective with the later rather than the former. Recognising that, while specific parameters have been identified in our analysis as a means of monitoring and assessing longevity for long-term class clients, it is highly recommended to craft a tailored approach to managing clients that have or seem to be entering this class.

7 | Reflection

7.1 | Difficulties Encountered

Throughout the project, our team faced several significant challenges, each of which we addressed with targeted strategies and collaborative problem-solving.

Scope Change and CLV Analysis: Initially, we proposed expanding our analysis to include Customer Lifetime Value (CLV) due to its potential to optimise client value. However, we discovered that commission alone was insufficient for comprehensive CLV analysis. Consequently, we refocused on client longevity, supplemented with limited customer value measurements to guide future CLV analyses.

Data Consolidation and Longevity Definition: The complexity of consolidating diverse historical data, especially given the varying intervals and the limited timeframe of comprehensive data, posed a significant hurdle. To address this, we adopted a descriptive approach and implemented thorough pre-processing techniques to accurately represent historical significance. This involved meticulous validation and aggregation methods to ensure data integrity, as well as thorough discussions with the client.

Leveraging Financial Insights: Applying financial terminology and concepts effectively during pre-processing was initially challenging. Consultations with industry experts and extensive internal discussions helped clarify these concepts. We incorporated these insights to enhance our feature selection and analysis processes.

Literature Review: Finding relevant literature to support our unique project requirements proved difficult. To overcome this, we continuously reviewed and adapted our literature search strategies, incorporating insights from various sources to inform our feature engineering and analysis.

Outlier Management: The presence of data skewness necessitated robust outlier management techniques. We implemented advanced detection and handling methods, complemented by sensitivity analyses, to ensure our models were not disproportionately influenced by outliers.

Pre-processing and Validation: Errors in pre-processing, particularly with currency conversions, required the implementation of multi-factor validation processes. Rigorous validation procedures, peer reviews, and quality assurance checks were established to mitigate similar issues in the future.

Missing Data: Unexplained instances of missing data raised concerns about data integrity. We intensified our efforts to investigate and address the root causes, employing data imputation techniques and conducting sensitivity analyses to mitigate the impact of missing data on our analysis.

By addressing these challenges with strategic interventions and collaborative efforts, we were able to navigate the complexities of the project and ensure robust and reliable analysis outcomes. Several difficulties encountered by us have also identified possible additional avenues of research that should be conducted, and additional data that may be needed, to better address the business problem faced by the client, as well as adjacent problems related to the issue of client longevity and profitability to a brokerage firm.

7.2 | Important Lessons

Conducting this analysis for SIGMA has yielded several important lessons that can inform future efforts and strategies for enhancing client engagement and retention.

Importance of Data Quality and Preparation: Extensive data preparation was integral to the success of this project. Due to the sheer quantity of the data, it was important to seriously consider and iterate on the best way to transform the data to achieve our desired needs. This was especially poignant given the complexity of the data, as mistakes would not have been immediately obvious, and could have threatened the integrity of our analysis. This experience underscores the need for meticulous data management practices, including regular audits and updates to maintain data integrity over time.

Significance of Feature Engineering: Similarly, intelligent feature engineering was crucial, highlighting the value of creating meaningful features that can enhance predictive models. Categorising clients based on their trading style, approach, and habits was difficult, but by identifying the most descriptive elements of a trader's strategy, we were able to uncover significant patterns between these features and client longevity. This process demonstrated that thoughtful feature engineering can significantly improve the explanatory power of models, leading to more actionable insights.

Value of Mixed-Method Approaches: The analysis reinforced the value of combining quantitative and qualitative data to gain a holistic understanding of client behaviour. While quantitative data provided concrete metrics and patterns, qualitative data offered context and deeper insights into client motivations and satisfaction. This mixed-method approach proved essential in painting a complete picture of the factors influencing client retention, suggesting that future analyses should continue to integrate both types of data.

Predictive Modelling Nuances: Another lesson learned is the nuanced nature of predictive modelling, especially when dealing with different client segments. The varying performance of models across the bottom 90% and top 10% of client accounts highlighted that different segments may require tailored modelling approaches. This insight suggests that a one-size-fits-all model is often insufficient and that segmentation and customization are key to improving predictive accuracy.

Long-Term Engagement Strategies: The findings revealed the importance of fostering long-term engagement through stable and sustainable trading practices. High-frequency trading and excessive profitability were linked to shorter trader lifespans, while moderate, consistent performance and prudent risk management were associated with longer engagement. This lesson underscores the need for SIGMA to promote disciplined and strategic trading among its clients, possibly through educational initiatives and tailored support services.

Balancing Costs and Benefits: Examining trading commissions and swap rates revealed that lower costs are associated with longer client lifespans. This lesson emphasises the need for SIGMA to balance costs and benefits effectively, ensuring that trading conditions are competitive while still maintaining profitability. Encouraging clients to choose assets with lower costs or tighter spreads could enhance overall client satisfaction and retention.

Continuous Monitoring and Adaptation: Finally, the analysis underscored the importance of continuous monitoring and adaptation in client engagement strategies. The dynamic nature of financial markets and client behaviours necessitates an ongoing commitment to data analysis and strategy refinement. Regularly updating models and strategies based on new data will help SIGMA stay responsive to changing conditions and client needs, ensuring sustained success in client retention efforts.

8 | Future of Work

To address SIGMA's business problem more comprehensively, several avenues for future work should be considered. These recommendations focus on expanding the data scope, incorporating qualitative insights, and conducting a detailed Customer Lifetime Value (CLV) analysis.

Expanding Dataset Size and Duration: A key suggestion is to gather a more substantial and expansive dataset that encompasses a broader timeframe. Presently, the analysis, although insightful, is constrained by the limited scope and duration of available data. A more extensive dataset spanning multiple years would offer deeper insights into trader behaviour and sustainability. This prolonged timeframe facilitates the recognition of enduring trends and patterns that might elude detection in shorter datasets. Moreover, a larger dataset enhances the viability of rigorous statistical analysis and model training, thereby enhancing the precision and dependability of predictive models.

Integration of Qualitative Insights: Sole reliance on quantitative data fails to encompass the entirety of factors shaping client engagement and retention dynamics. Thus, there is a paramount need to incorporate a broader range of qualitative data sources, including client satisfaction surveys, feedback garnered from customer support interactions, and marketing analytics. These qualitative inputs offer a nuanced perspective, enriching the understanding of the underlying reasons behind clients' decisions to either stay with the service or disengage. For example, insights from satisfaction surveys illuminate specific areas of strength or weakness within the service, guiding targeted enhancements. Furthermore, analysis of marketing data unveils the resonance of various campaigns and messaging strategies with clients, facilitating the formulation of more impactful marketing approaches.

Customer Lifetime Value (CLV) Assessment: A comprehensive evaluation of Customer Lifetime Value (CLV) is imperative to synergize the discoveries of this study with comprehensive data elucidating the value derived from individual clients. CLV analysis stands as a cornerstone in comprehending the enduring profitability of clients and pinpointing the segments that wield the greatest influence on the company's revenue stream. By amalgamating insights gleaned from our examination of trading behaviours, longevity, and profitability with CLV metrics, SIGMA can strategically allocate resources and focus on clientele with higher value. Furthermore, this analysis facilitates refined segmentation, empowering tailored retention tactics aimed at optimising returns on investment in client engagement endeavours.

Personalised Engagement Strategies: Developing personalised engagement strategies based on the insights from a more extensive dataset and qualitative data is another critical area for future work. These strategies should be tailored to different client segments identified through the CLV analysis and predictive modelling. Personalised approaches could include targeted educational resources, customised

trading advice, and exclusive offers for high-value clients. This should be done in service of creating an individualised service so as to foster a more meaningful client-service relationship, and therefore greater longevity.

Continuous Monitoring and Adaptation: Establishing a structured framework for ongoing monitoring and iteration of client engagement and retention strategies is essential. This entails a perpetual process of gathering and scrutinising data to gauge the efficacy of implemented strategies, making data-driven modifications as necessary. Regular updates to models and analyses with fresh data ensure that SIGMA's strategies remain pertinent and potent within a dynamic market landscape.

9 | References

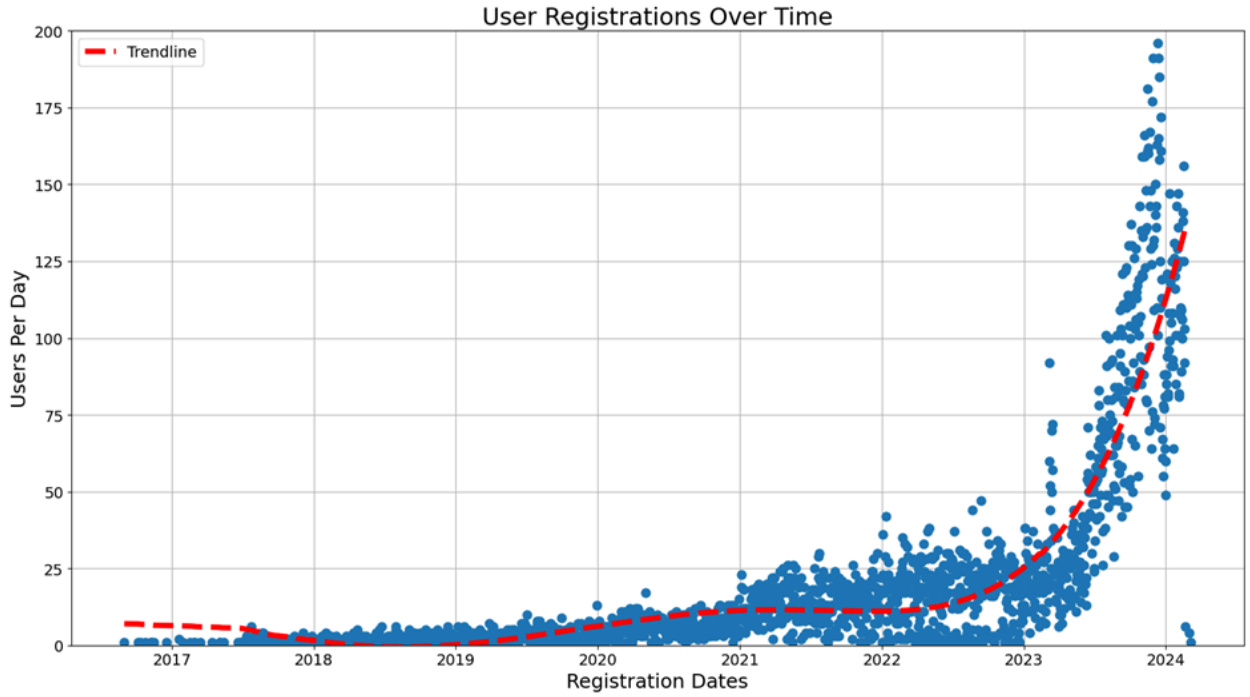
Ma, T., Fraser-Mackenzie, P. A. F., Sung, M., Kansara, A. P., & Johnson, J. E. V. (2022). Are the least successful traders those most likely to exit the market? A survival analysis contribution to the efficient market debate. *European Journal of Operational Research*, 299(1), 330-345. https://doi.org/10.1016/j.ejor.2021.08.050.

Brozynski, T., Menkhoff, L., & Schmidt, U. (2004). The Impact of Experience on Risk Taking, Overconfidence, and Herding of Fund Managers: Complementary Survey Evidence (Diskussionsbeitrag No. 292). *Universität Hannover, Wirtschaftswissenschaftliche Fakultät, Hannover*. https://hdl.handle.net/10419/22404.

Qualtrics Support. (2017). Interpreting Residual Plots to Improve Your Regression - Qualtrics Support. [online] Available at: https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/.

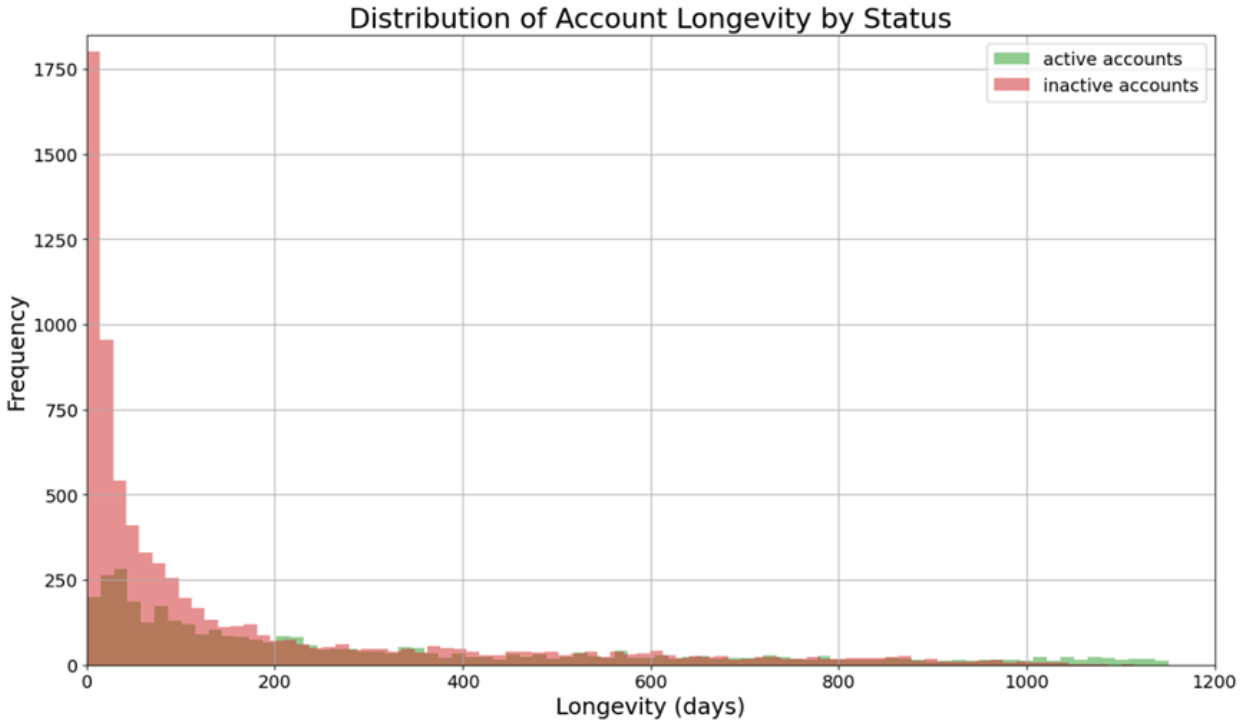
10 | Appendices

Appendix A: User registrations Over Time



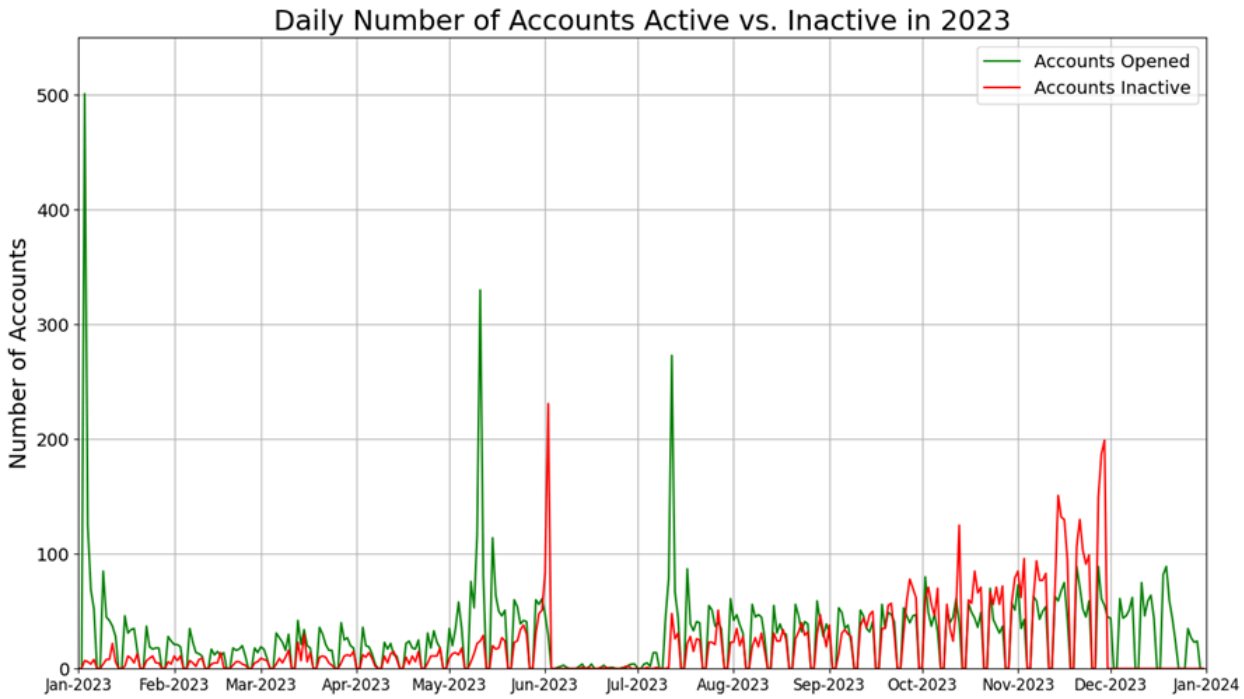
Note: This graph plots the number and trend of users who registered on a particular date.

Appendix B: Distribution of Account Longevity by Status



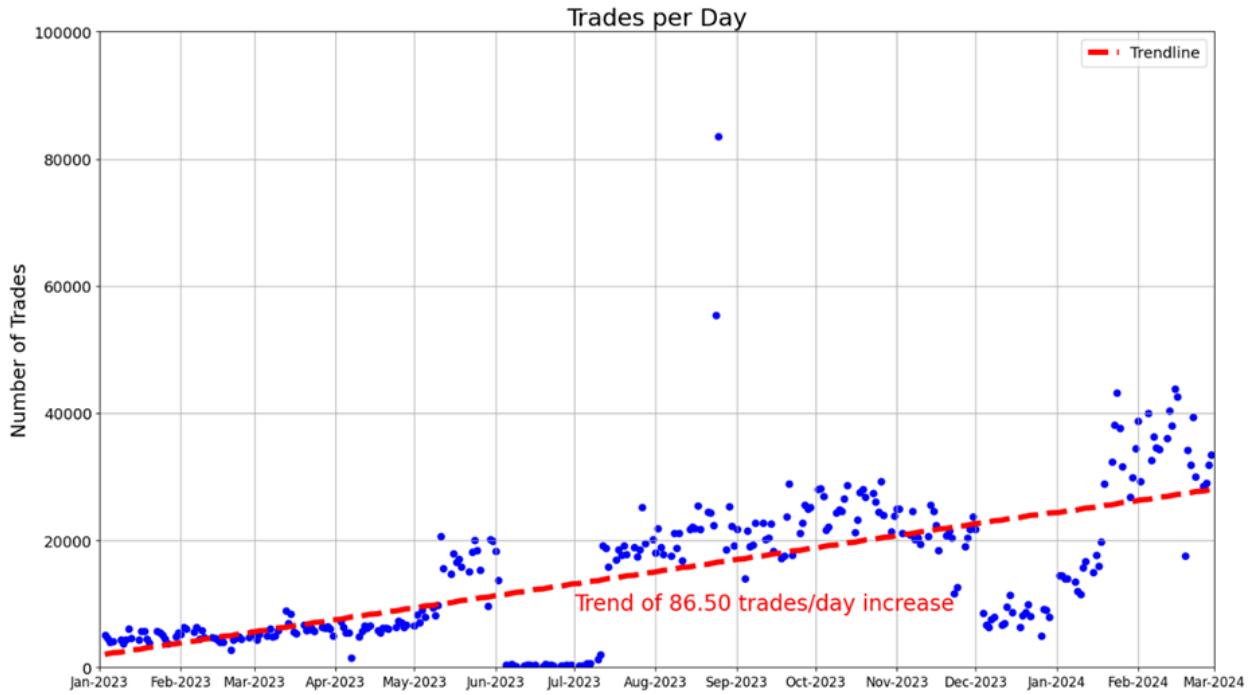
Note: This graph plots the distribution of active and inactive accounts according to their longevity. Inactive accounts in this case are accounts that have not traded within the past 2 months.

Appendix C: Daily Number of Accounts Active vs. Inactive in 2023



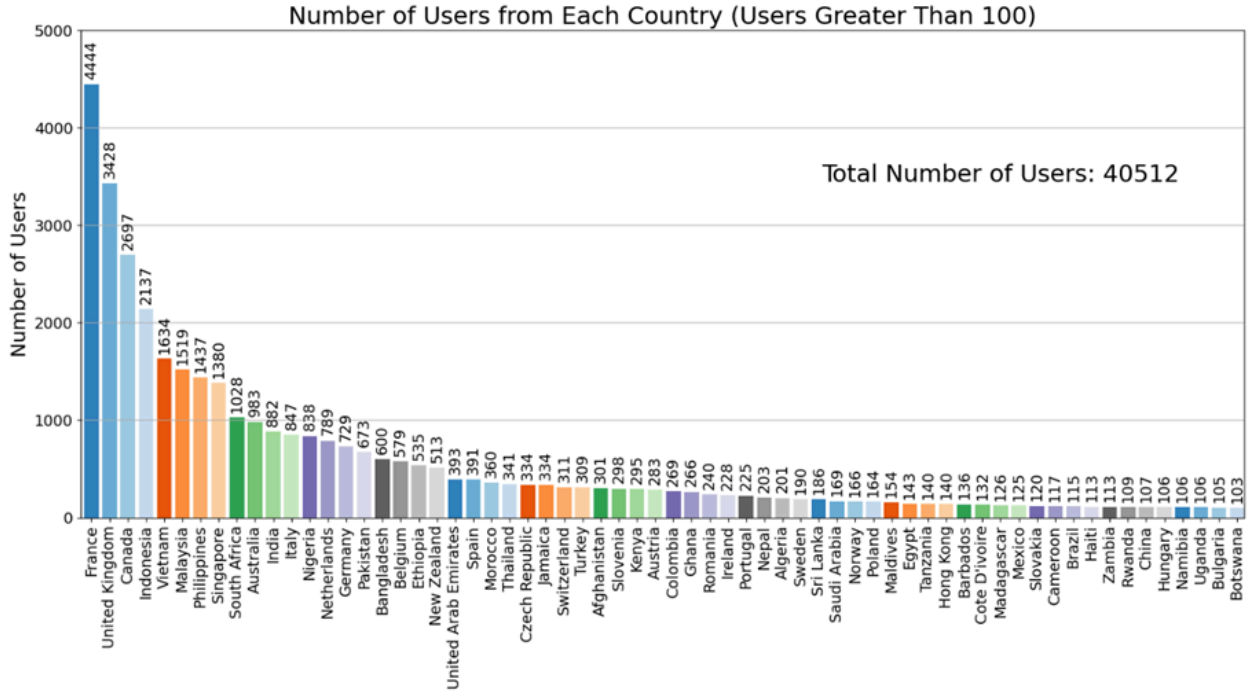
Note: This graph plots the number of accounts that become active during the period against the number of accounts that enter inactivity. Inactive accounts in this case are accounts that have not traded within the past 2 months.

Appendix D: Trades per Day



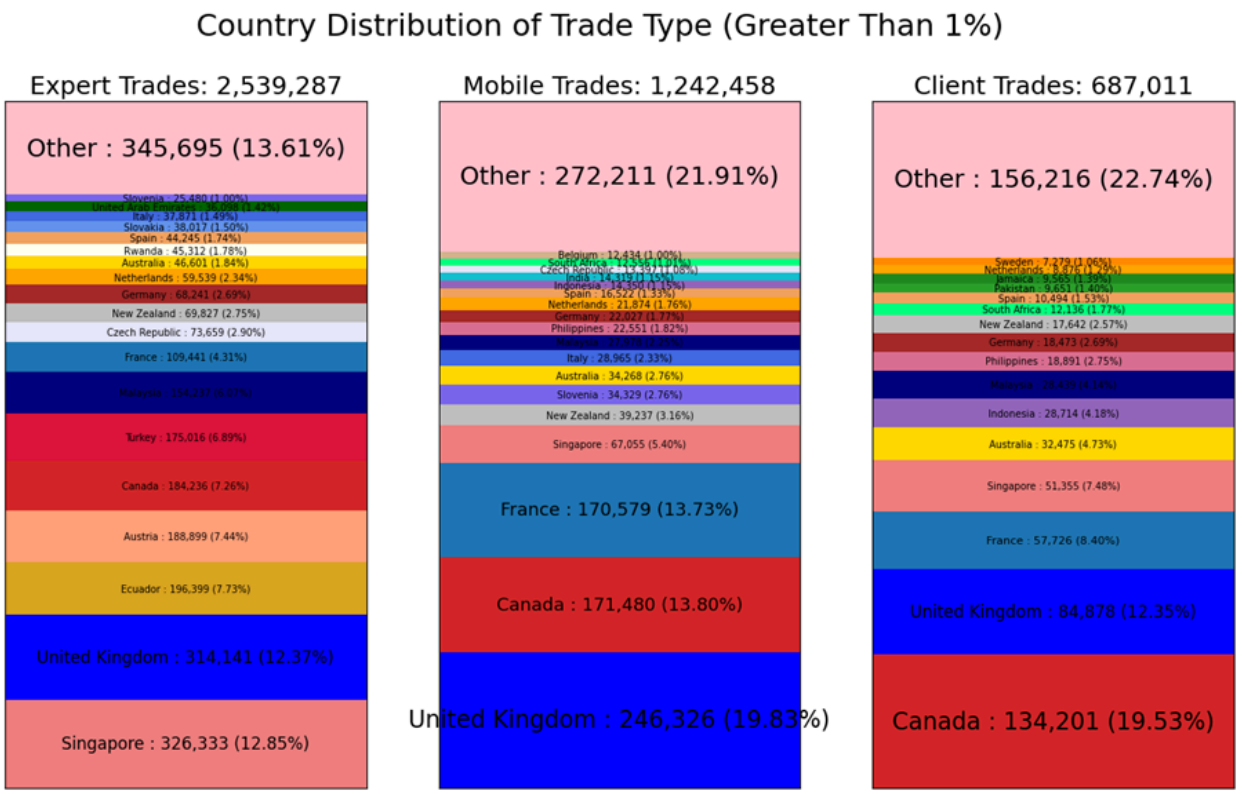
Note: This graph plots the number and trend of trades that are commenced each day, according to their open datetime.

Appendix E: Number of USers from Each Country (Users Greater than 100)



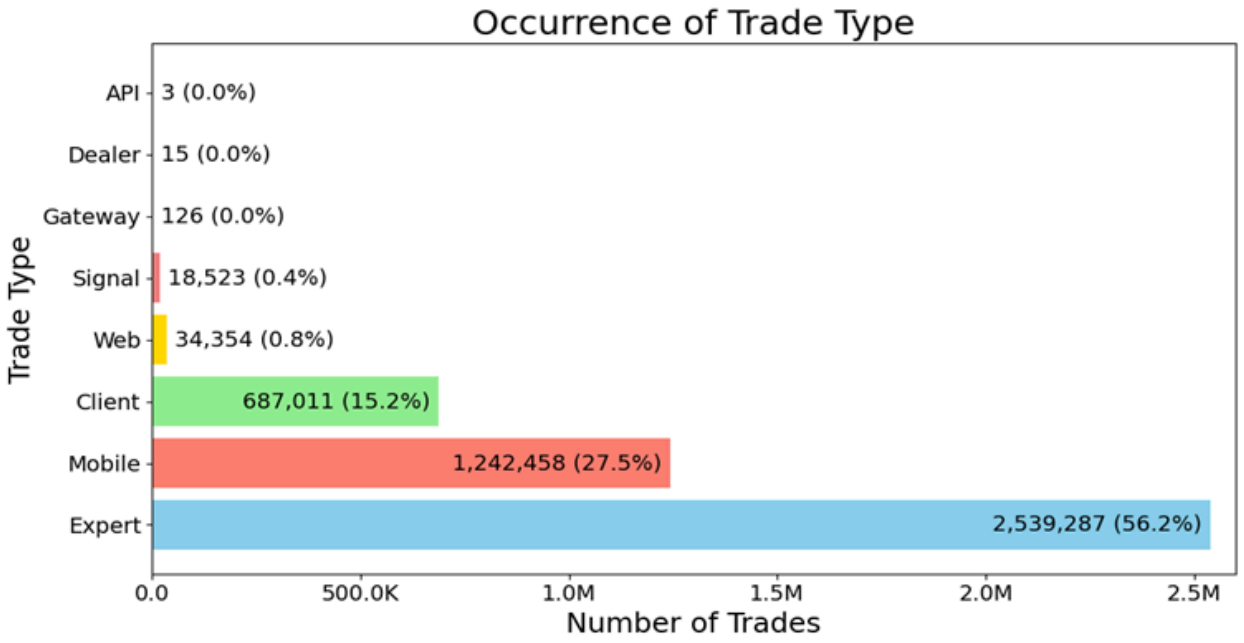
Note: This graph plots the number of accounts registered to each country where the number of users for that country is greater than 100.

Appendix F: Country Distribution of Trade Type (Greater Than 1%)



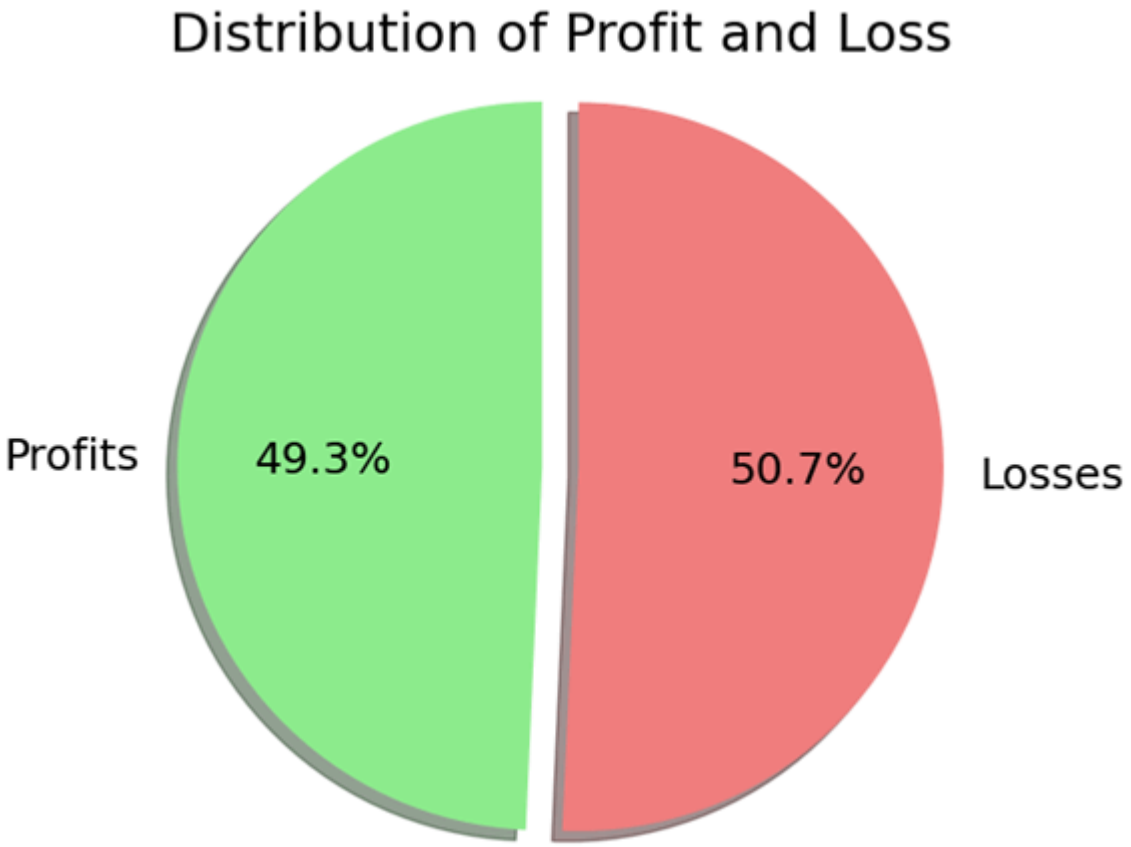
Note: This graph shows the distribution of each country across the three most prominent trading types recorded in the dataset.

Appendix G: Occurrence of Trade Type



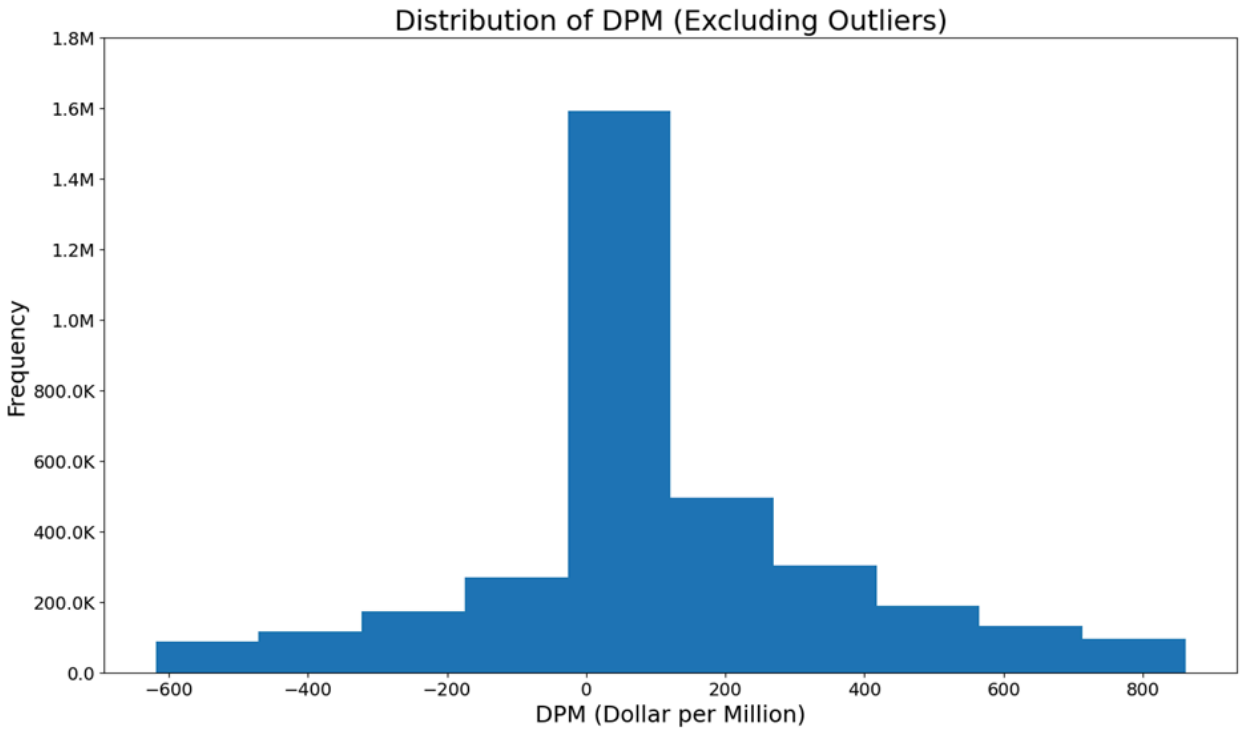
Note: This graph plots the number of trades that are conducted using a specified trading method.

Appendix H: Distribution of Profit and Loss



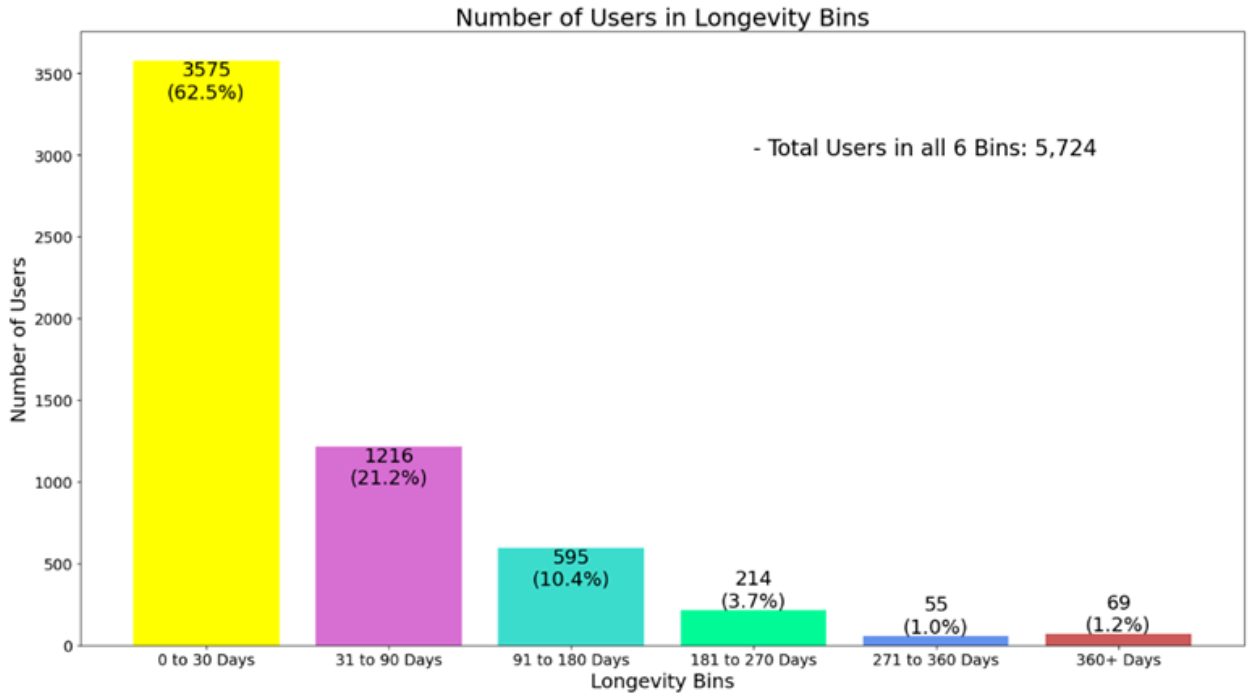
Note: This graph plots the distribution of total trades based on whether they were profitable or not. In this case, profitable trades are those with profit greater than 0, and losing trades are those with a profit less than 0.

Appendix I: Distribution of DPM (Excluding Outliers)



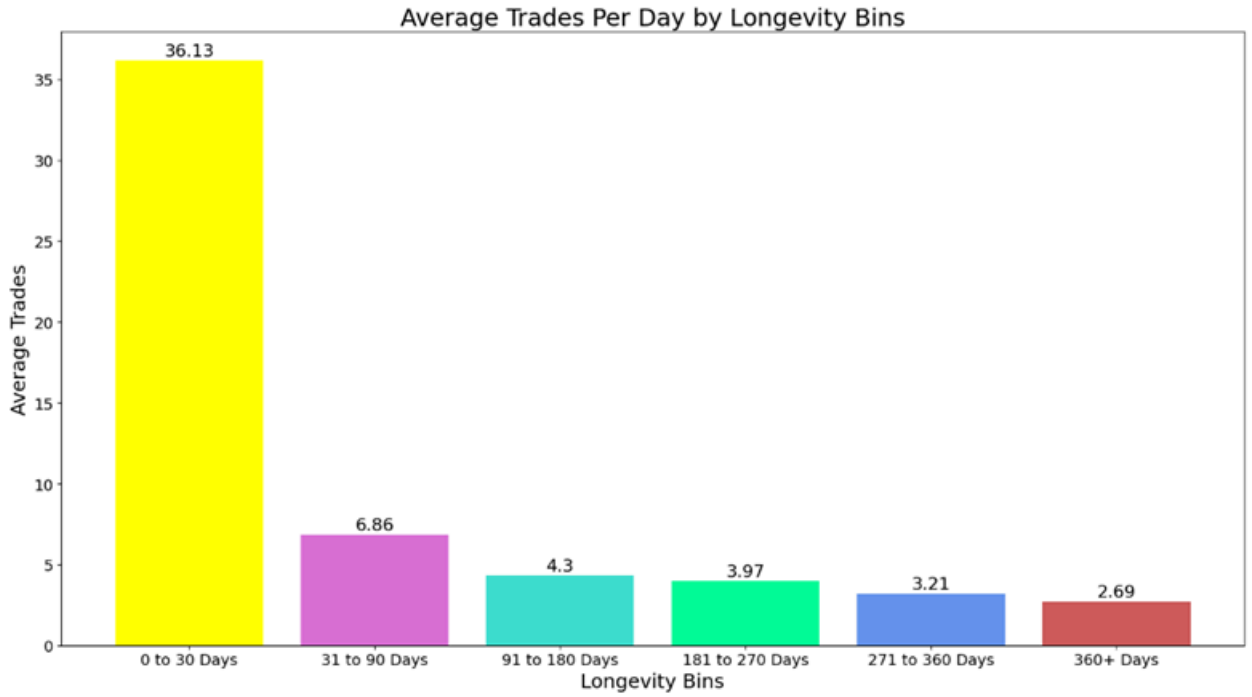
Note: This graph plots the distribution of trades according to their DPM, or dollar per million, where DPM equals total profit/loss in USD over volume in millions.

Appendix J: Number of Users in Longevity Bins



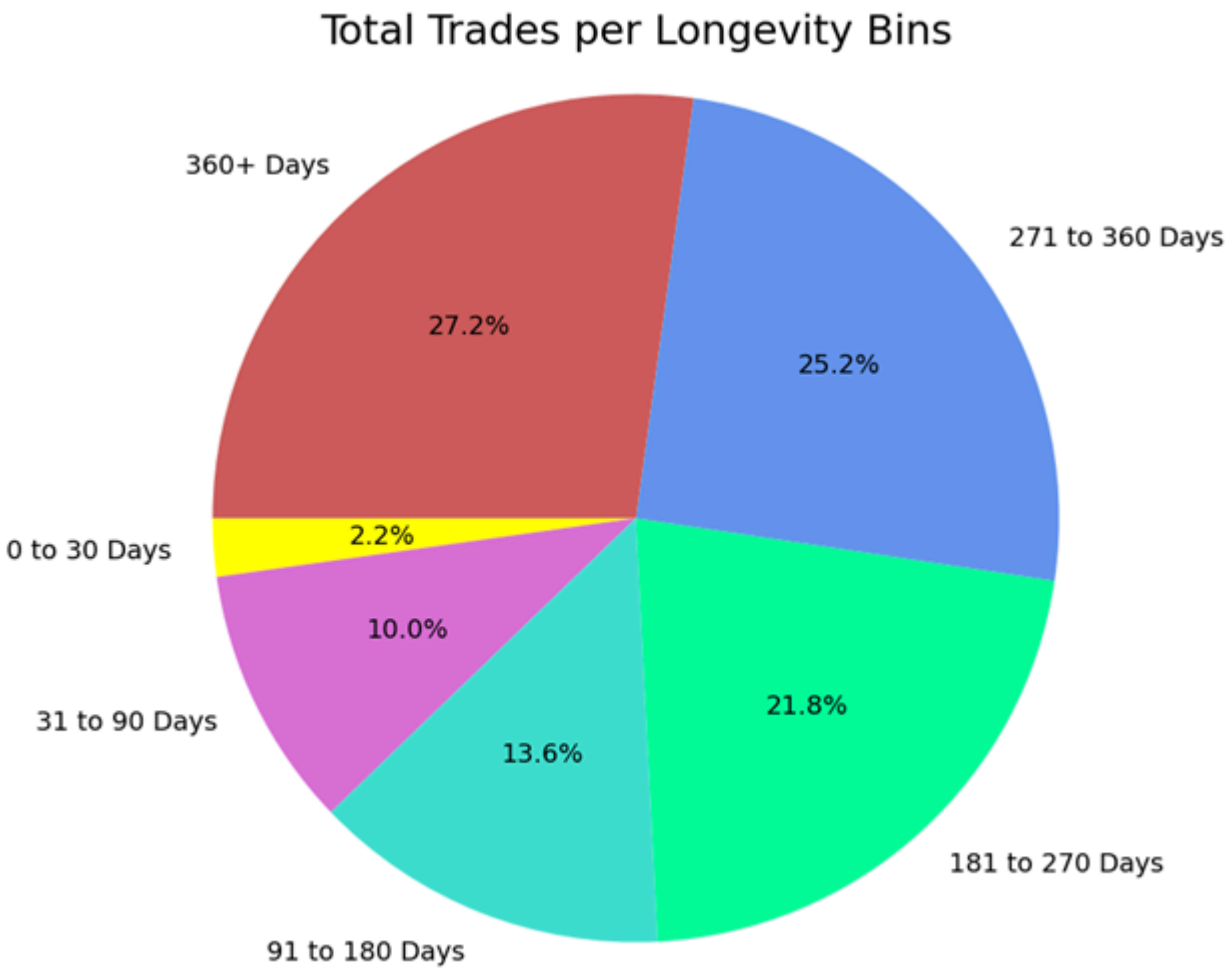
Note: This graph plots the distribution of accounts across each longevity bin.

Appendix K: Average Trades Per Day by Longevity Bins



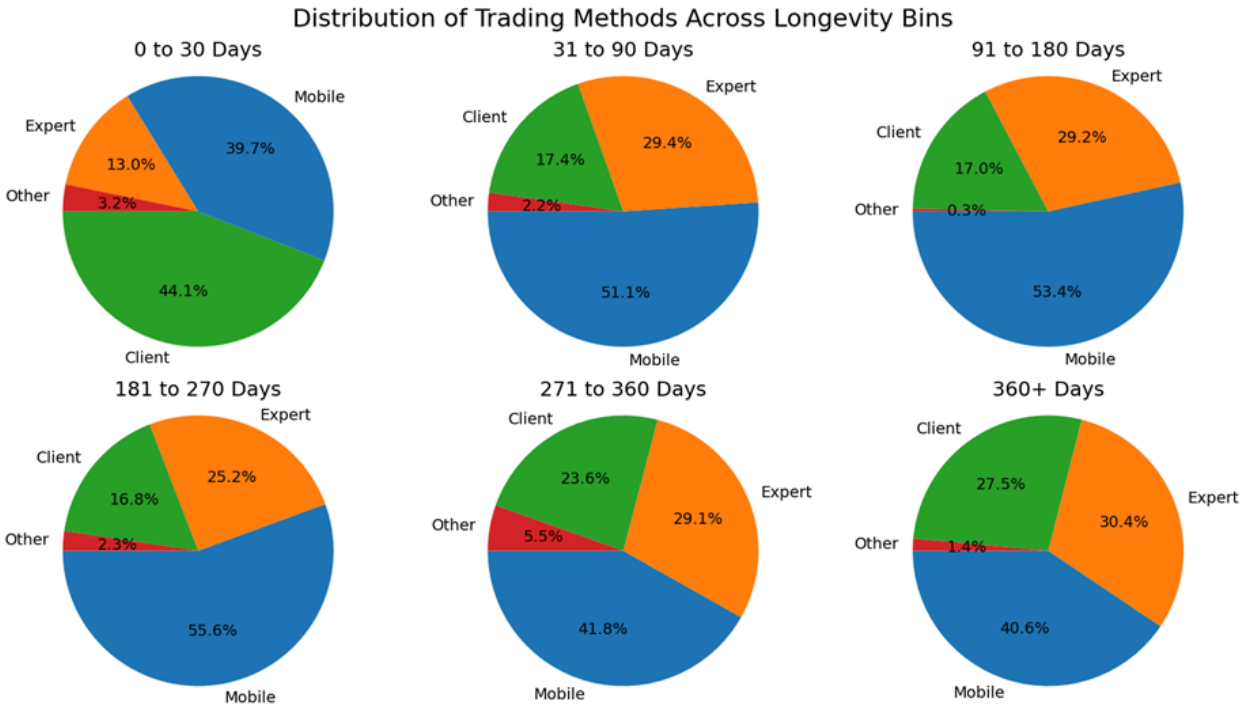
Note: This graph plots the distribution of trades per day across accounts in each longevity bin.

Appendix L:Total Trades per Longevity Bins



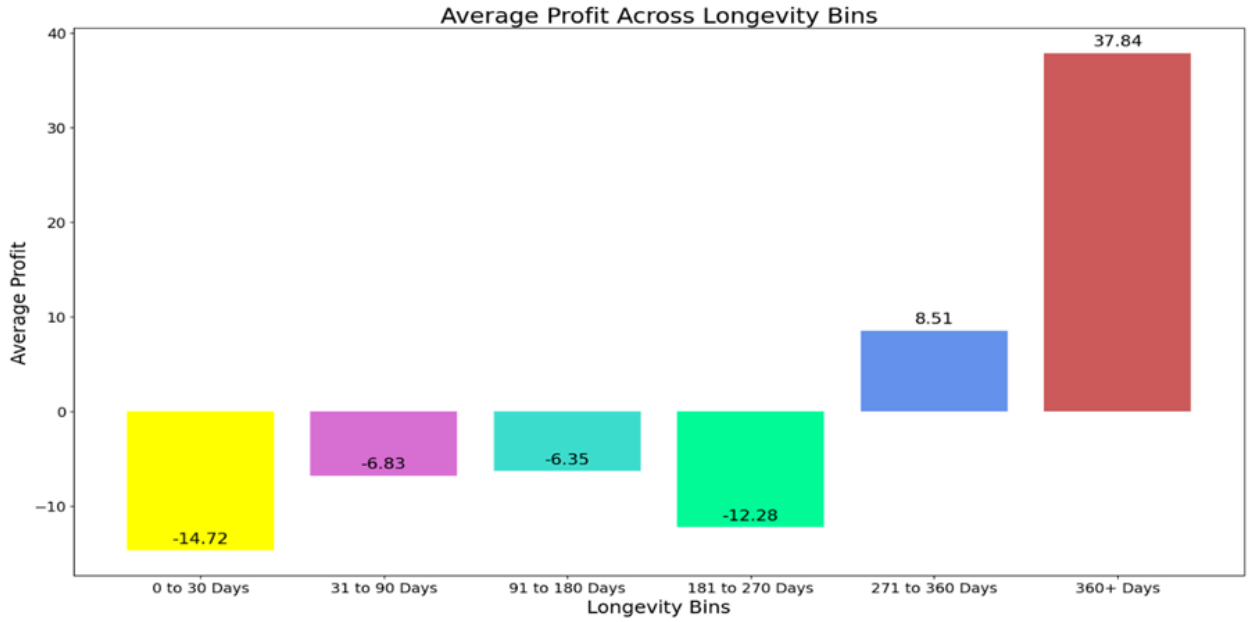
Note: This graph plots the distribution of the total number of trades made by each account across longevity bins relative to the total number of trades.

Appendix M: Distribution of Trading Methods Across Longevity Bins



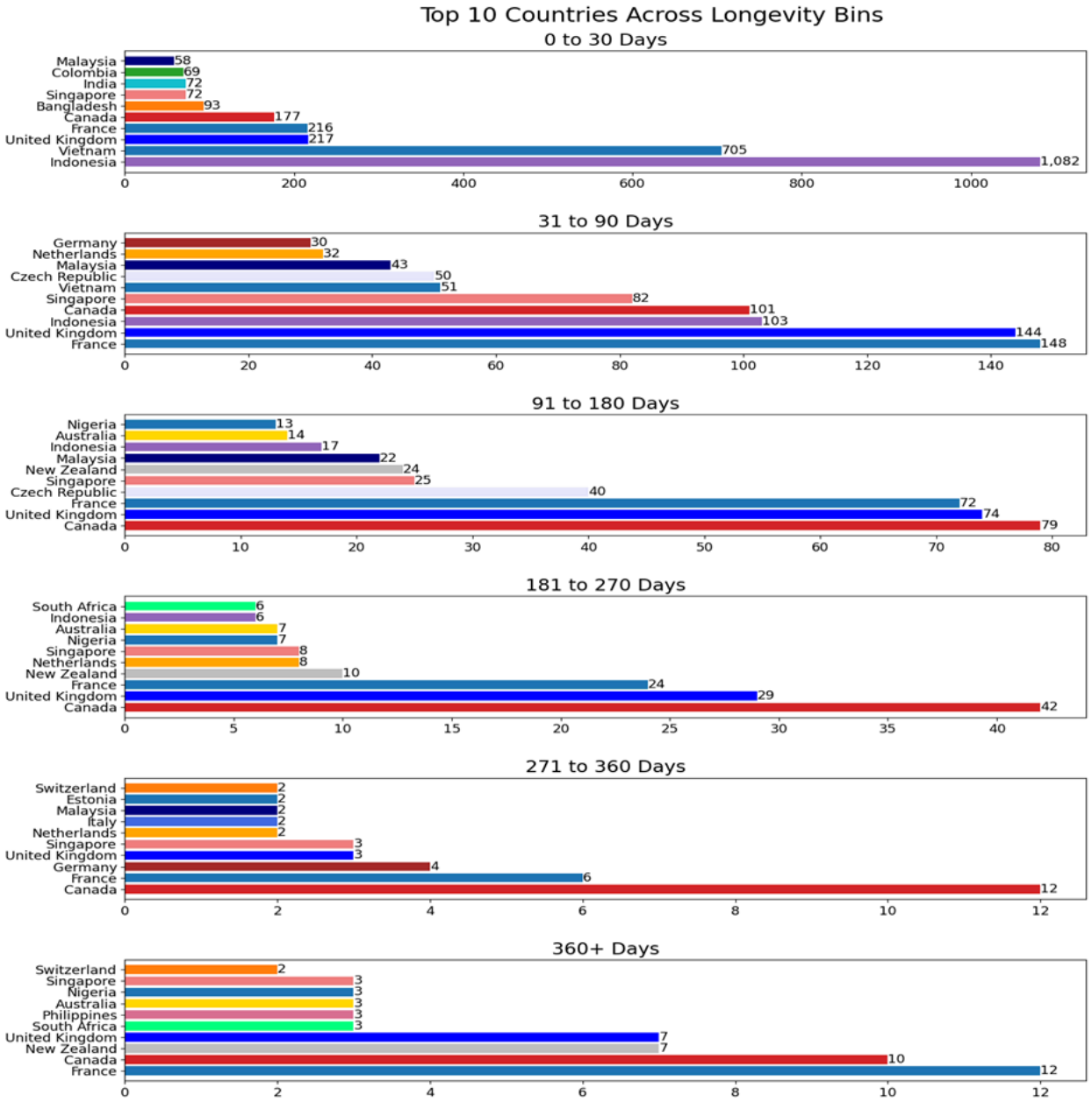
Note: This graph plots the distribution of accounts according to their most popular trading method across each longevity bins relative to the total number of accounts in that bin.

Appendix N: Average Profit Across Longevity Bins



Note: This graph plots the distribution of accounts across each longevity bin according to the average profit of accounts in that bin.

Appendix O: Top 10 Countries Across Longevity Bins



Note: This graph plots the top 10 most popular countries had by accounts across each longevity bin.