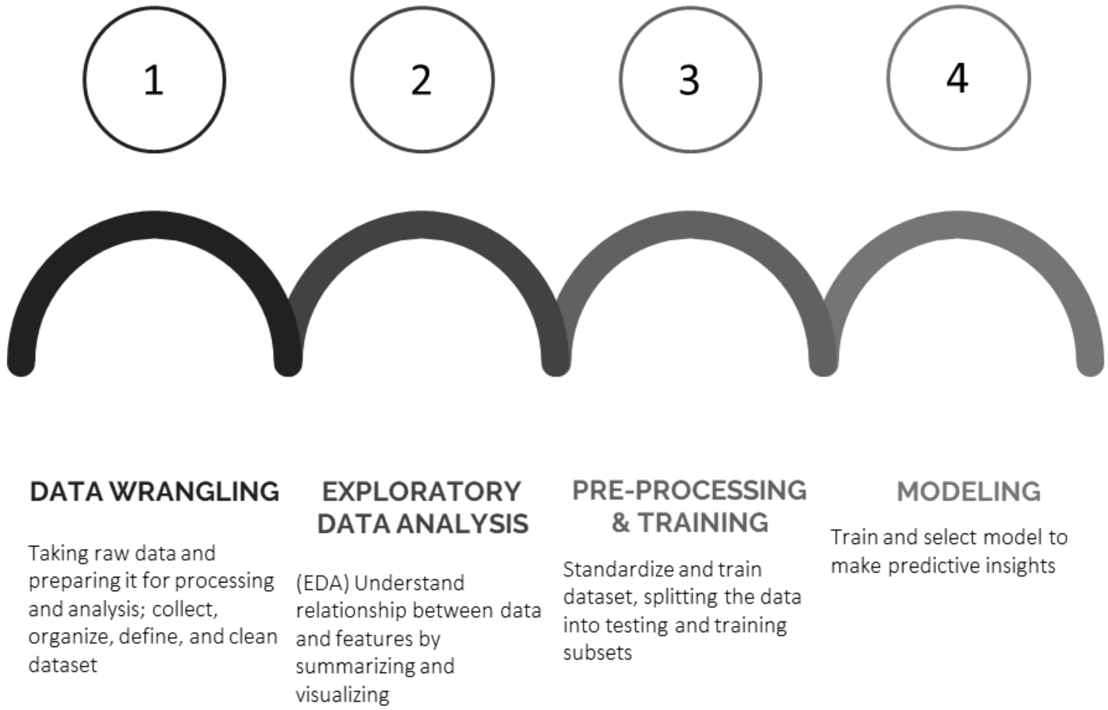
Katarina Delos Santos

PREDICTING MOBILE PAYMENT ADOPTION IN EMERGING ASIAN MARKETS:
A MACHINE LEARNING APPROACH FOCUSED ON ACCURATE FORECASTING OF DIGITAL PAYMENT TRENDS IN SOUTHEAST ASIA

Mobile payment adoption is crucial for businesses and financial institutions operating in Southeast Asian markets, where digital transformation is rapidly reshaping the financial landscape. This study employs advanced machine learning techniques to forecast adoption patterns.



<Figure 1: Data Science Method (DSM)>

I. Introduction

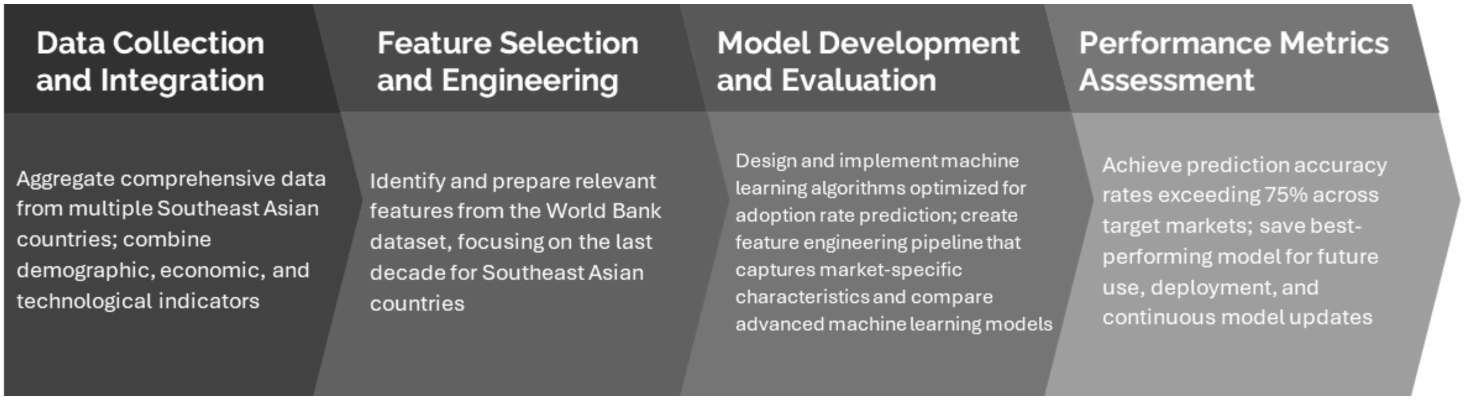
1) Problem Statement

Accurate prediction of mobile payment adoption rates drives strategic decision-making in Southeast Asian markets, where digital transformation is revolutionizing financial services. The Southeast Asian financial services landscape has witnessed unprecedented growth in mobile banking adoption, making predictions important for fintech companies, traditional banks, and digital payment providers. Mobile payment adoption in emerging Asian markets faces unique challenge such as diverse demographics, varying digital infrastructure, multiple competing platforms, regulatory frameworks, and cultural payment preferences. This project presents a machine learning approach to predict mobile payment adoption rates across Southeast Asian markets. The analysis of mobile payment adoption has significantly evolved from traditional market research methods to modern machine learning approaches.

Keywords: Mobile Payment Adoption, Machine Learning, Predictive Analytics, Southeast Asian markets, Digital Financial Services

2) Goal

This project's primary goal is to develop an accurate and reliable model for predicting mobile payment adoption using comprehensive data. In order to achieve this, the following specific objectives have been identified:



<Figure 2: Project Objectives>

These goals align with providing valuable market intelligence for financial institutions, technology providers, and policy makers in the rapidly evolving Southeast Asian mobile payment landscape.

II. Dataset Source

This analysis uses data from The World Bank Group's DataBank World Development Indicators Mobile Money[1]. The comprehensive database provides a rich foundation for mobile payment adoption analysis in Southeast Asia. Key characteristics of the dataset include data coverage of a temporal range 2010-2022 and a geographic scope that includes Southeast Asian countries with over fifty relevant indicators per country such as financial indicators (mobile money transaction volumes, digital payment penetration rates), economic metrics (GDP per capita), income distribution indicators, technological infrastructure (mobile phone penetration, internet accessibility, digital infrastructure quality), demographic factors (population demographics, urban vs. rural distribution, education levels), and market development indicators (financial institution presence, regulatory environment metrics.

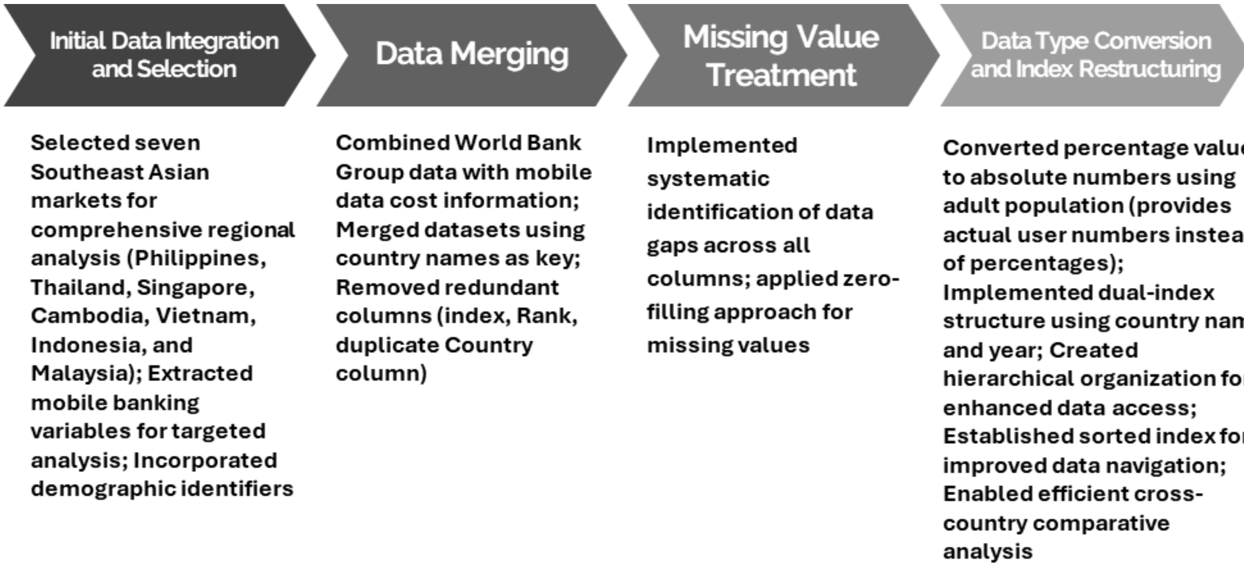
On top of this dataset, "The Global Cost of 1GB of Mobile Data"[2], a dataset containing the average price of 1GB of mobile data for over 150 countries sourced from Visual Capitalist was utilized, which can be found on Kaggle. Each row lists the country's name and its corresponding average price for 1GB of mobile data with the most expensive countries being listed first.

Data Cleaning Procedures

1) Initial Data Integration and Selection- Selected seven Southeast Asian markets for comprehensive regional analysis (Philippines, Thailand, Singapore, Cambodia, Vietnam, Indonesia, and Malaysia); extracted mobile banking variables for targeted analysis; incorporated demographic identifiers
2) Data Merging- Combined World Bank Group data with mobile data cost information; Merged datasets using country names as key; Removed redundant columns (index, Rank, duplicate Country column)
3) Missing Value Treatment- Implemented systematic identification of data gaps across all columns; applied zero-filling approach for missing values
4) Data Type Conversion and Index Restructuring- Converted percentage values to absolute numbers using adult population (provides actual user numbers instead of percentages);

[1] https://databank.worldbank.org/Mobile-Money/id/d1fc2d5d
[2] https://www.kaggle.com/datasets/thedevastator/the-global-cost-of-1gb-of-mobile-data

implemented dual-index structure using country name and year; created hierarchical organization for enhanced data access; established sorted index for improved data navigation; enabled efficient cross-country comparative analysis



<Figure 3: Data Cleaning Procedures>

The data preparation process began with a strategic selection of Southeast Asian countries that represent key markets in the region. Seven major economies were chosen: Philippines, Thailand, Singapore, Cambodia, Vietnam, Indonesia, and Malaysia. This selection ensures our analysis captures diverse economic landscapes across Southeast Asia while at the same time maintaining a manageable and relevant dataset scope.

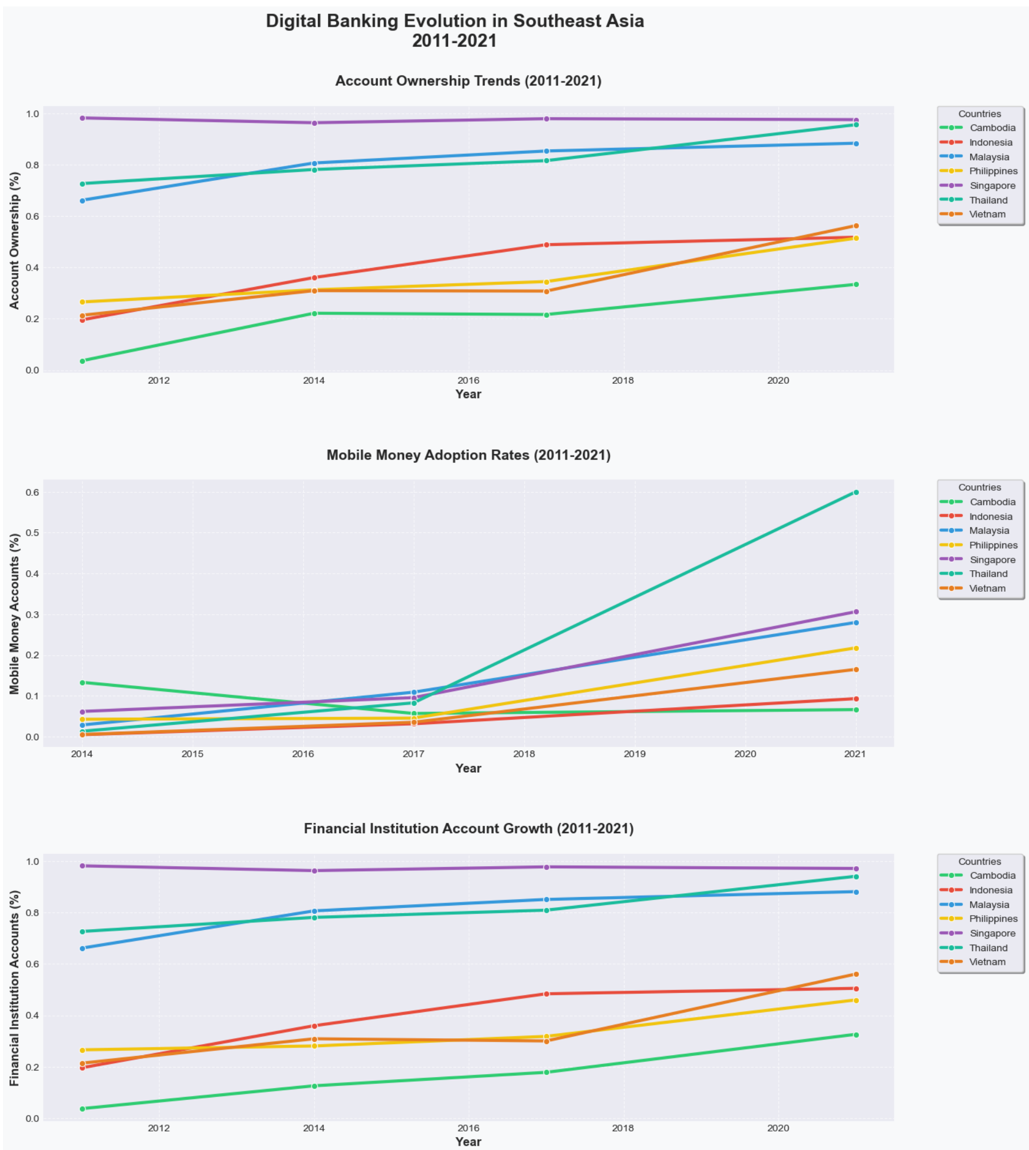
Following country selection, the most pertinent variables for mobile banking analysis were identified and extracted. These key indicators include fundamental demographic data such as country name and year, along with the adult population figures, account figures, and other relevant columns. Financial metrics including overall account ownership percentages, traditional financial institution account penetration rates, and specific mobile money account adoption rates were incorporated. This focused selection of variables maintains data relevance while optimizing for mobile payment pattern analysis in the region.

The data analysis process included a comprehensive examination of missing values across all columns in the mobile banking dataset. Through systematic evaluation, gaps in the data that required strategic handling were identified. The solution involved filling these missing values with zeros, which aligns with the logical assumption that blank entries indicate no mobile banking adoption or usage during those specific periods. This approach maintains data integrity while providing a complete dataset to make a machine learning model. The transformation from missing values to zero enhances the ability to perform robust statistical analyses and generates more accurate predictions of mobile payment adoption trends.

The dataset organization was enhanced through the implementation of a multi-index structure using both Country name and Year as primary identifiers. This hierarchical indexing system creates a powerful framework for analyzing trends across different time periods and geographical locations. By sorting the index, a logical flow of data that facilitates efficient cross-country comparisons and temporal analysis is made. This structured approach enables seamless access to specific country-year combinations and supports advanced analytical operations, making it particularly valuable for tracking mobile payment adoption patterns across Southeast Asian markets over time.

The key outcomes were that a clean, structured dataset with no missing values was created; percentages were transformed into absolute numbers for better analysis; a hierarchical index for country-year analysis was established; mobile data cost information was integrated; and redundant, unnecessary columns were removed. This cleaning process ensures data is ready for model development while maintaining data integrity and relevance to the project objectives.

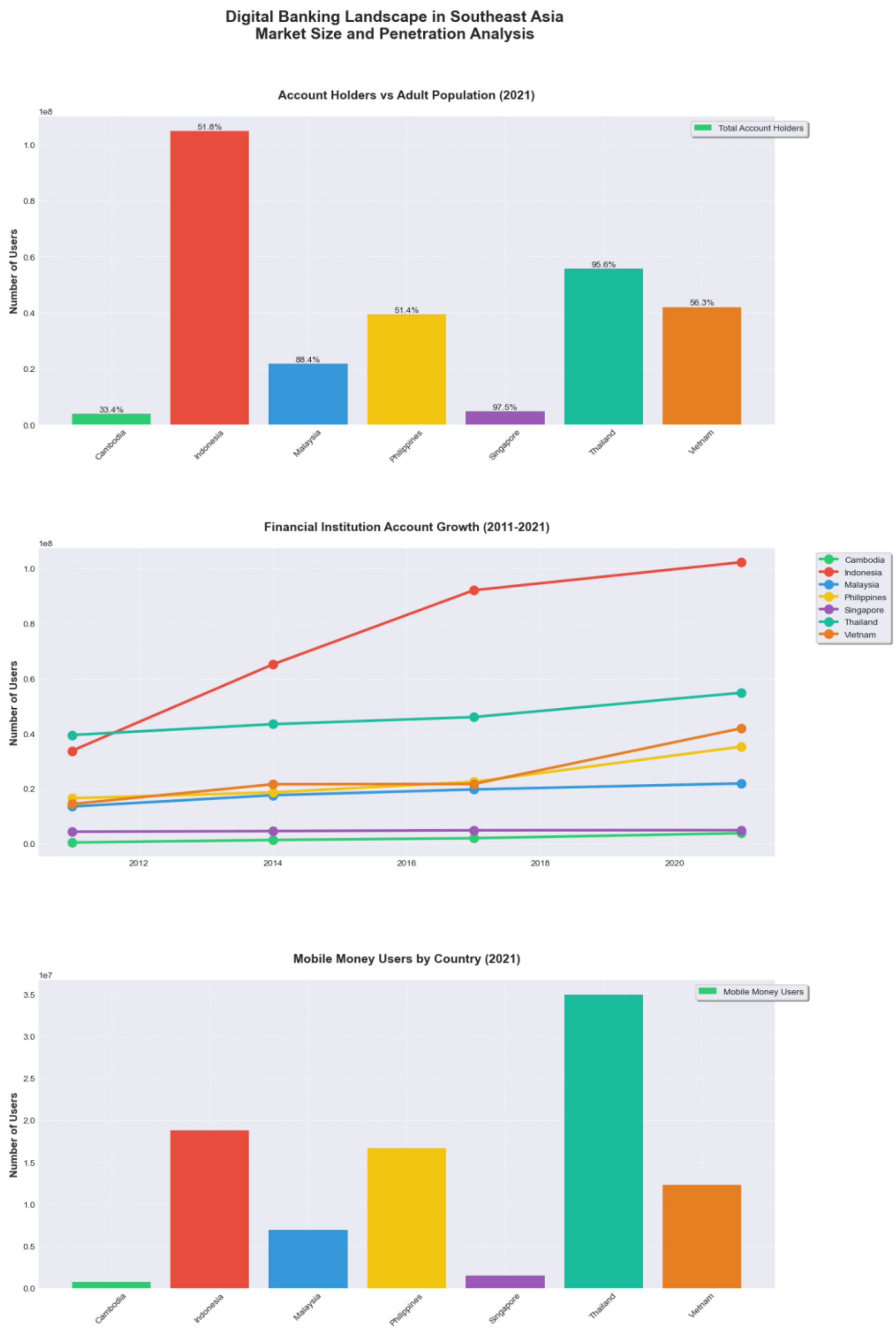
III. Exploratory Data Analysis (EDA)



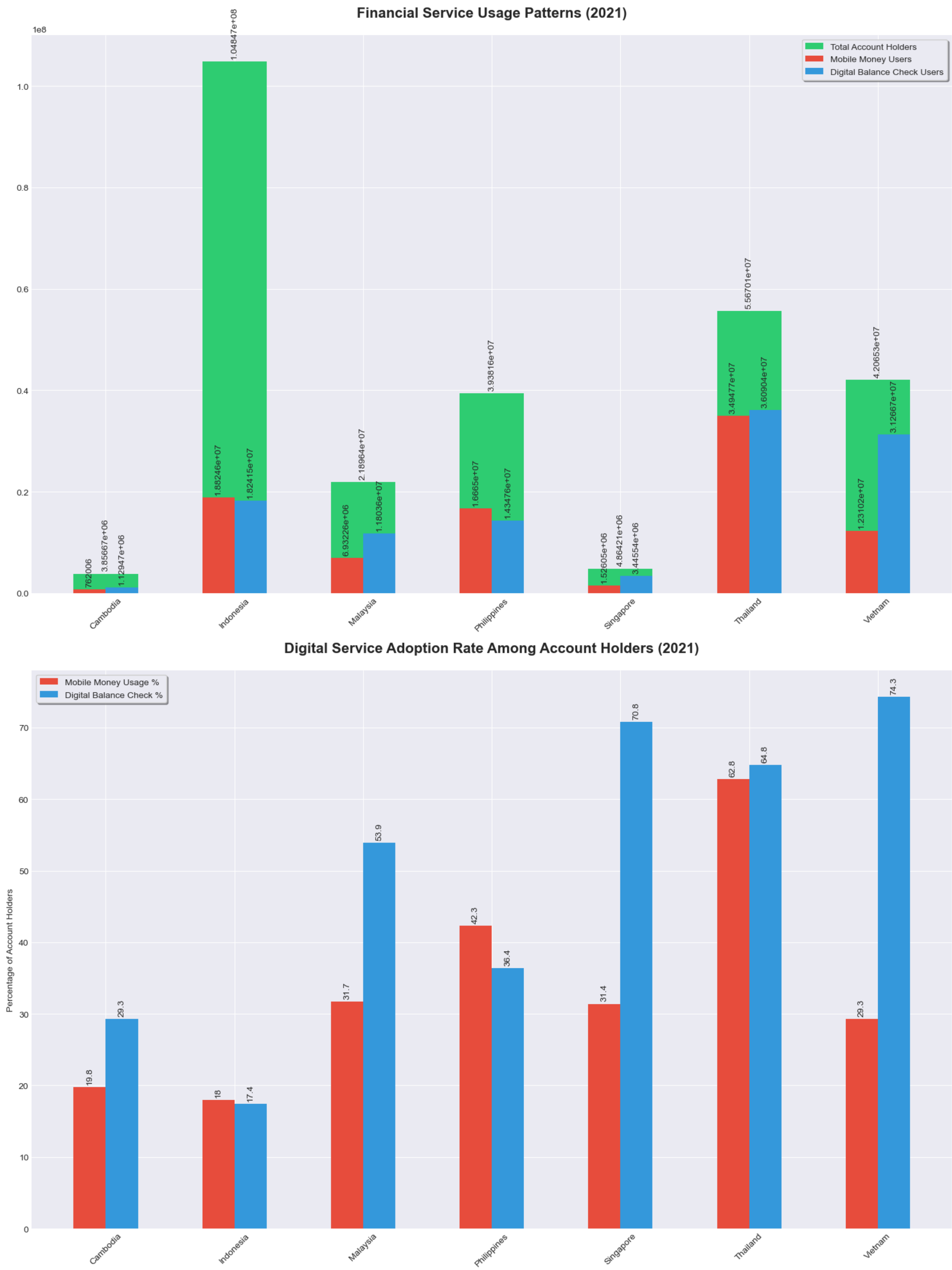
<Figure 4: Evolution of Digital Financial Services Adoption in Southeast Asia (2011-2021)>

<Figure 5: Comparative Analysis of Digital Banking Adoption Across Southeast Asian Markets (2011-2021)>

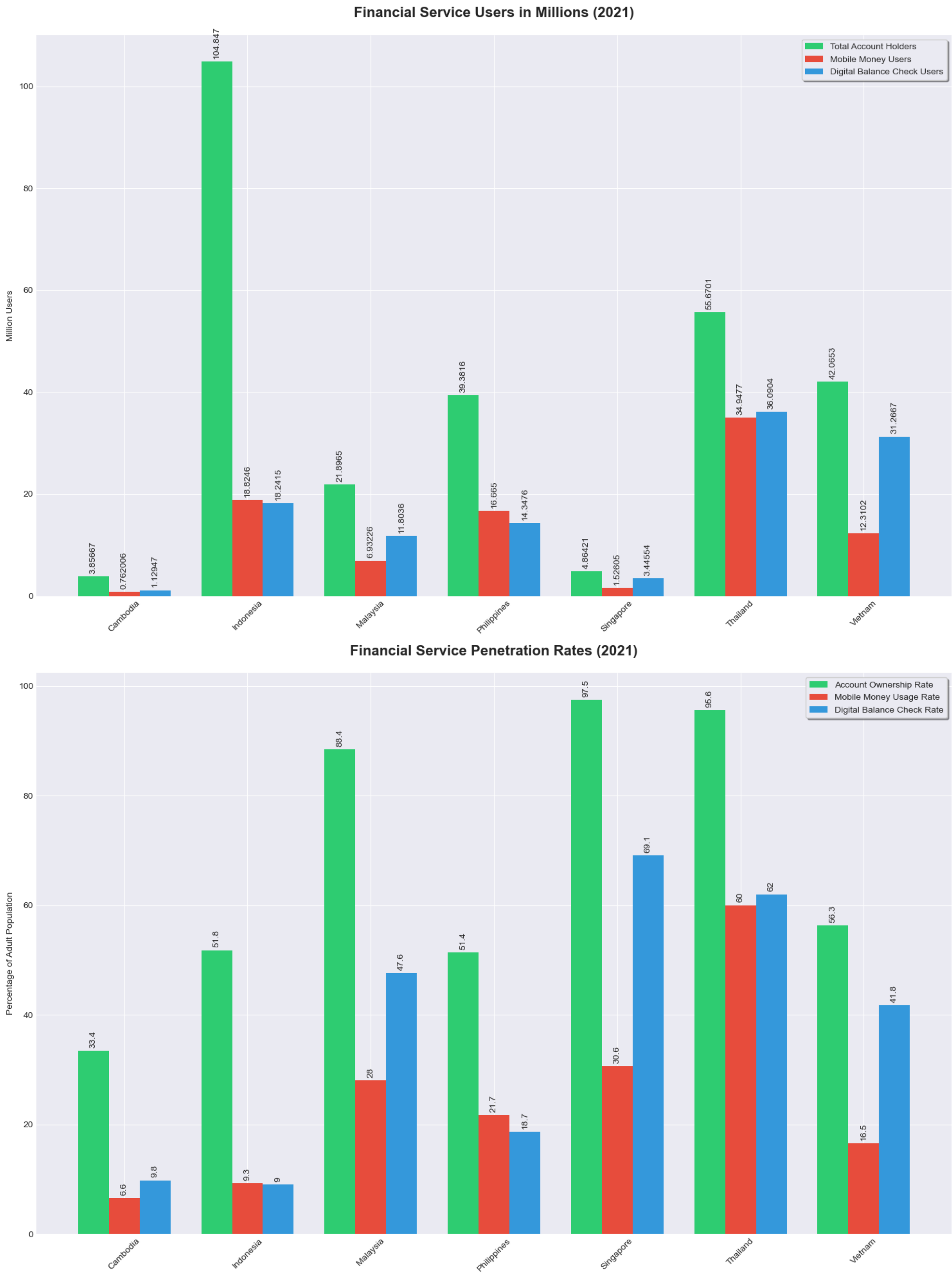
Looking at Singapore's data, Figure 5 makes sense since Singapore's adult population is around five million people, and the data shows nearly complete financial inclusion: 4.86 million total account holders in 2021 (97.5% of adult population), 4.85 million financial institution account holders, and 1.53 million mobile money users. The numbers reflect Singapore's high financial inclusion rate. Therefore, almost all adults have bank accounts, and there is strong digital adoption with 2.91 million digital balance check users in 2017 with high penetration of both traditional and digital banking. The data shows Singapore's advanced financial ecosystem of high urban digital usage, strong adoption across income groups, and significant digital banking penetration.



<Figure 6: Digital Banking Market Penetration and User Base Analysis in Southeast Asia>

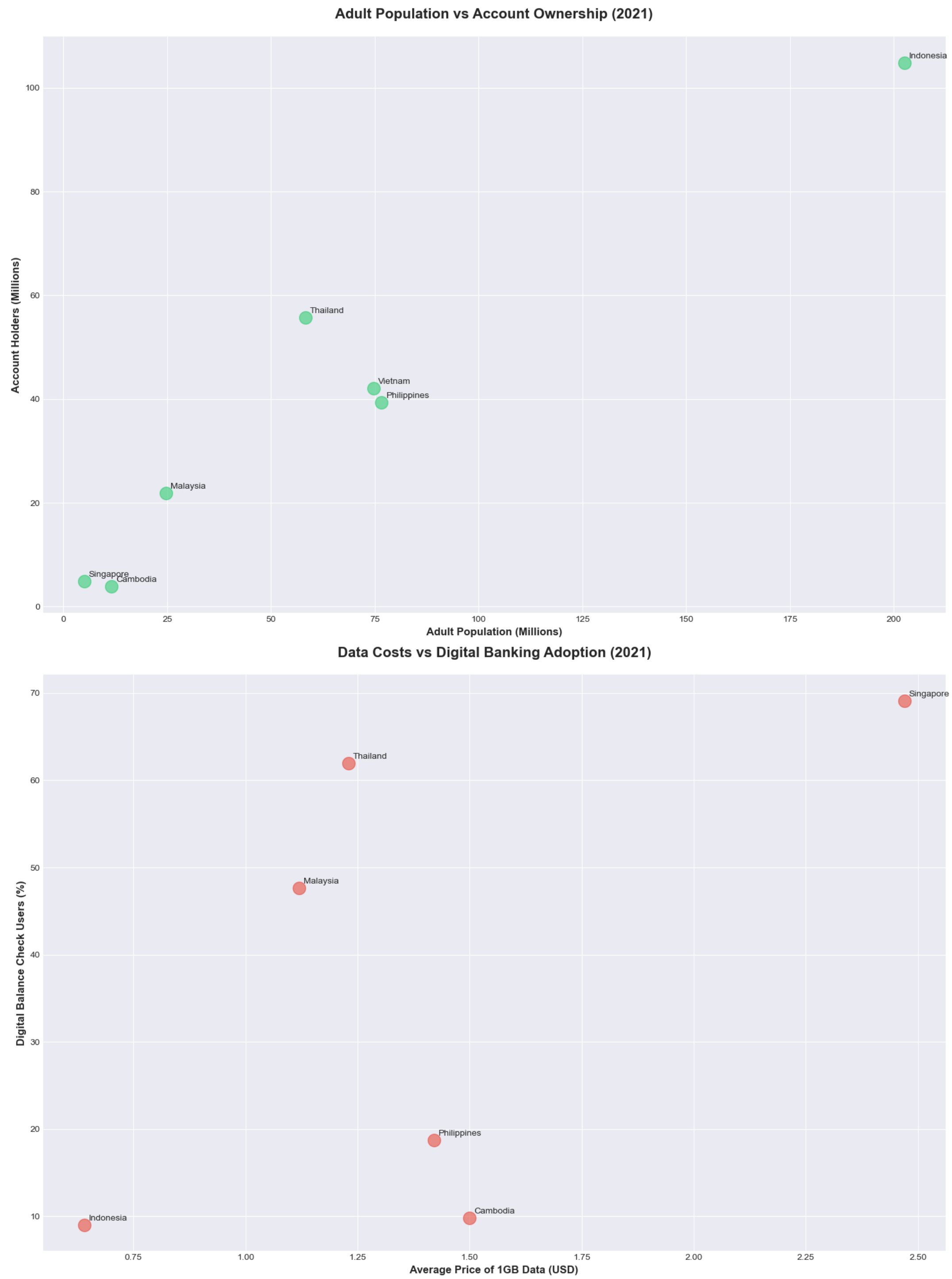


<Figure 7: Digital Financial Services Adoption and Usage Patterns in Southeast Asia (2021)>



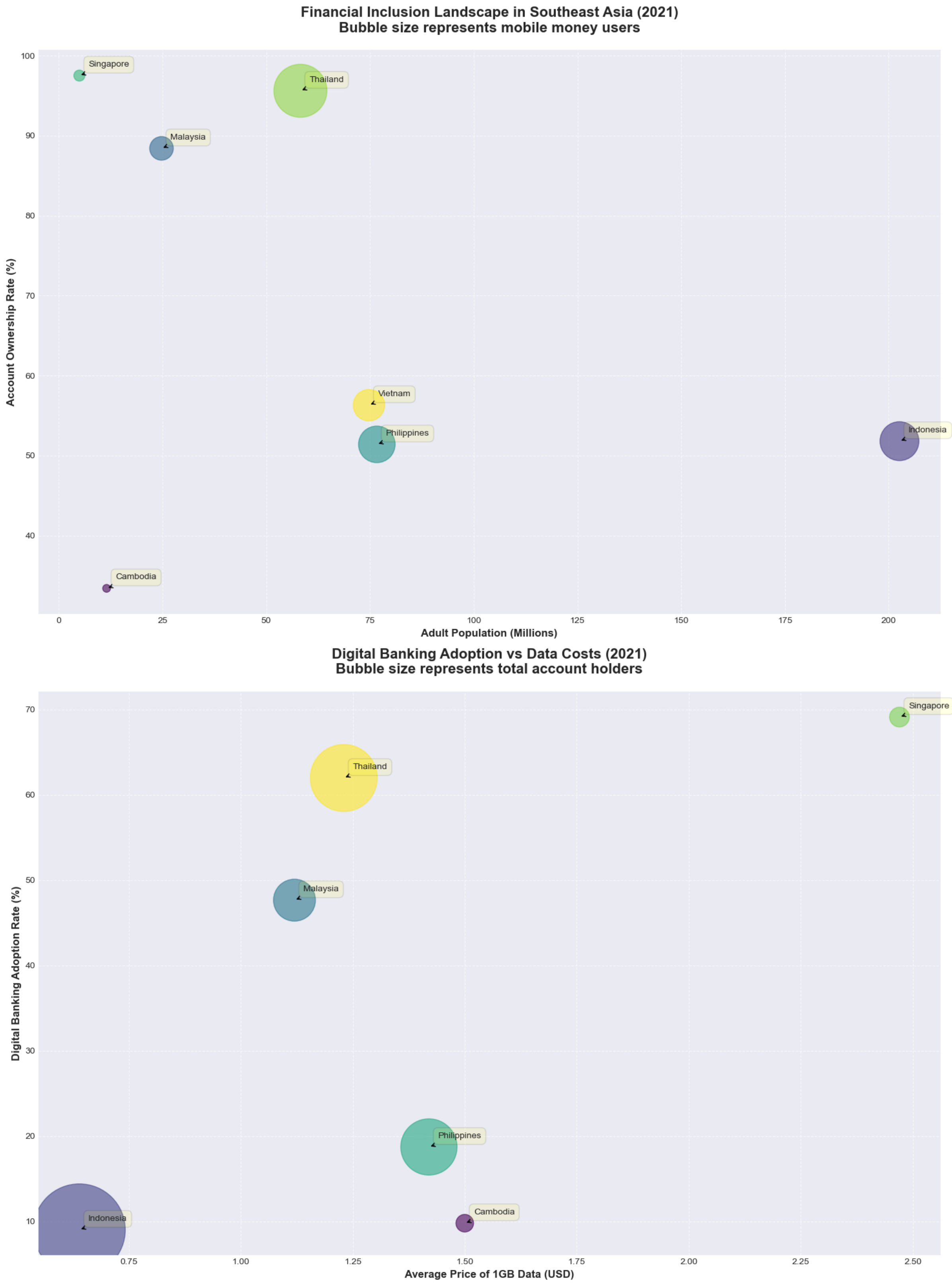
<Figure 8: Digital Financial Services Market Scale and Population Penetration Analysis in Southeast Asia (2021)>

Figure 8 is a visualization that showcases Singapore's high penetration rates alongside the absolute numbers, highlighting the position as a digital banking leader in Southeast Asia.



<Figure 9: Digital Banking Adoption Drivers- Population Scale and Data Cost Impact Analysis in Southeast Asia (2021)>

For the upper panel of Figure 9 “Adult Population vs. Account Ownership (2021)”, it can be noticed that Indonesia is high up on the right-hand corner (high x, high y), but Singapore is next to Cambodia (low x, low y), closer to the origin 0,0. This pattern makes sense and actually validates the data. The reason Indonesia appears high up on the right is because Indonesia has the largest adult population in Southeast Asia at around 172 million adults. Therefore, with high account ownership, it translates to a massive number of account holders. On the other hand, Singapore and Cambodia appear closer to the origin because Singapore has around a 5 million adult population, and Cambodia has around 11.5 million for adult population. Thus, this visualization shows market size rather than market penetration.

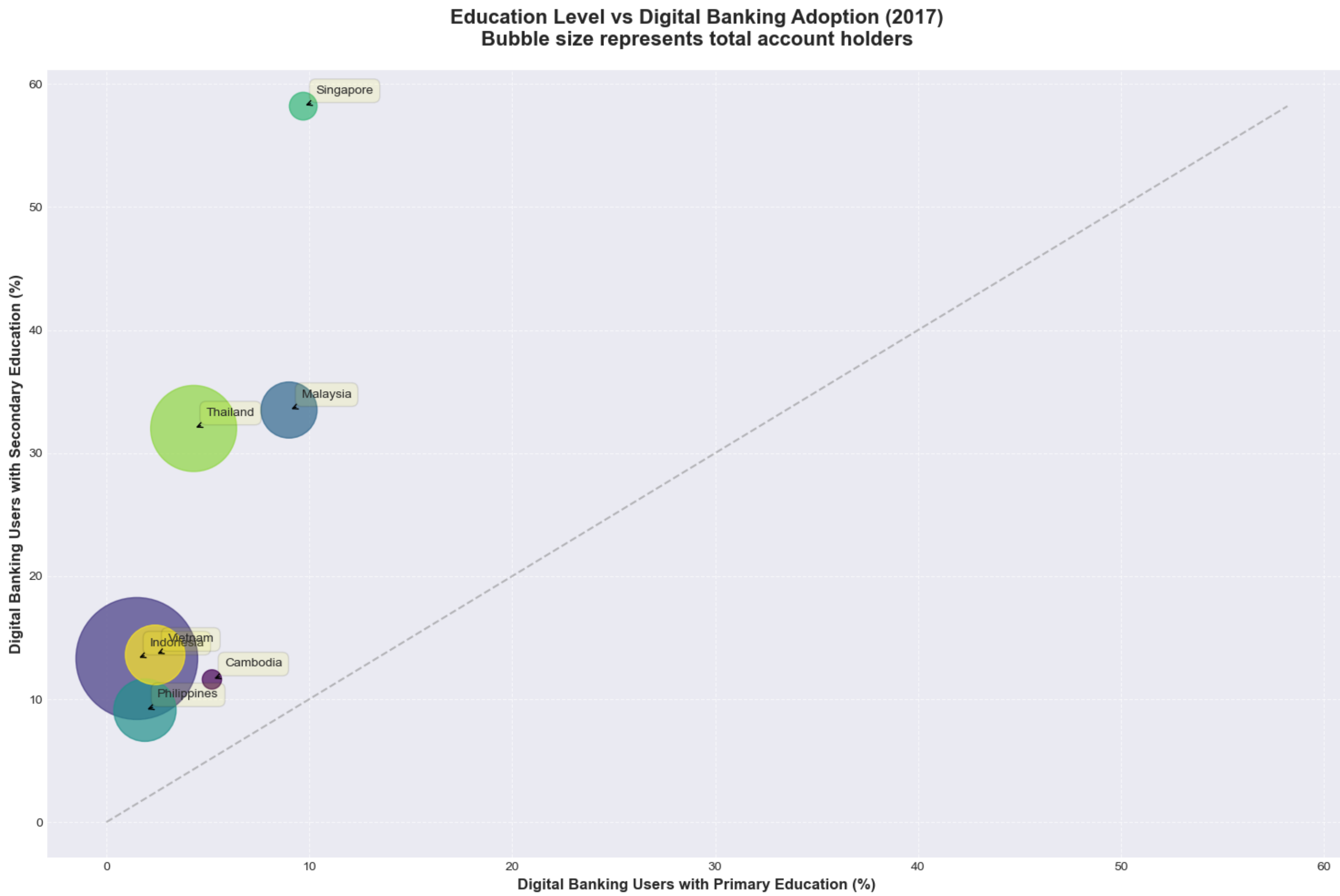


<Figure 10: Multi-Dimensional Analysis of Financial Inclusion in Southeast Asia (2021)>

As compared to Figure 9, Figure 10 is a visual to show Singapore's leadership in financial sophistication, which can be seen through the bigger bubble sizes, while still maintaining the context of population size and data costs on the axes. Through this image, it is easier to see the scale differences across Southeast Asia.

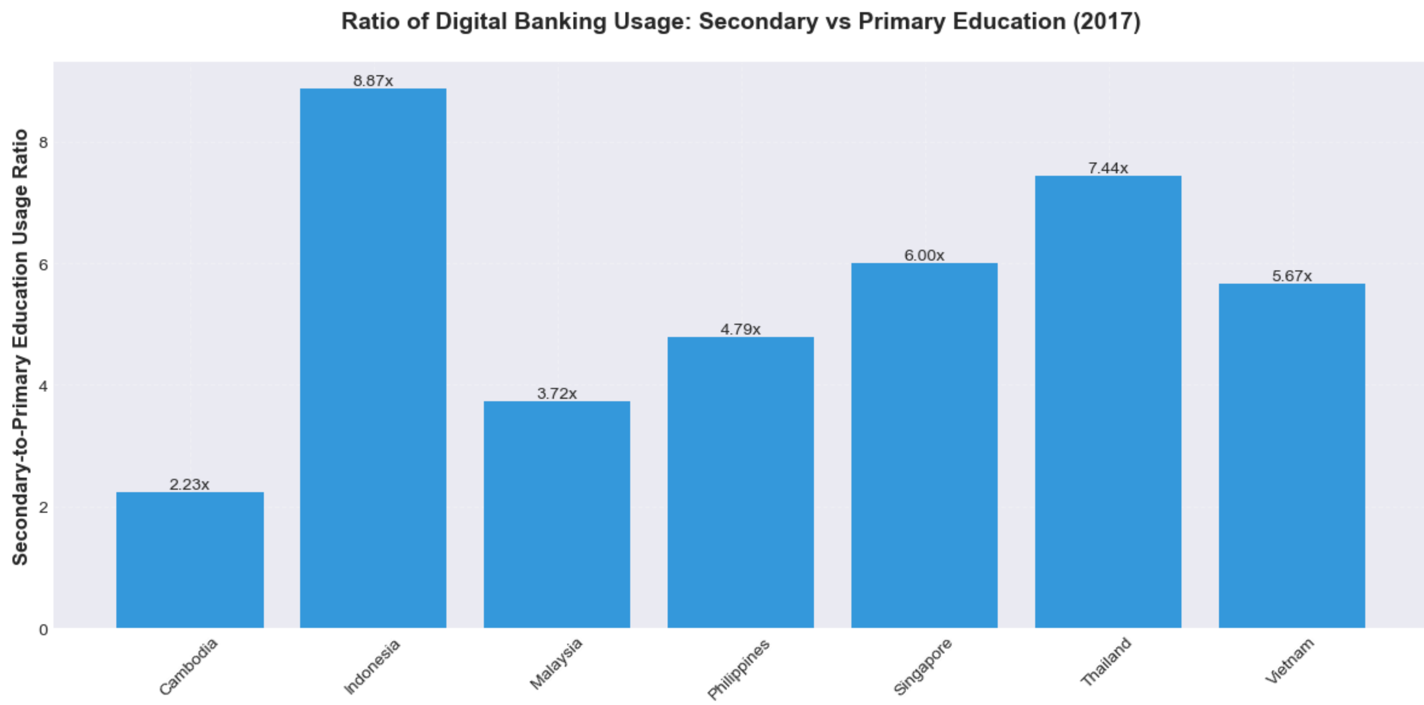
<Figure 11: Digital Financial Ecosystem- Interplay of Population, Account Ownership and Digital Adoption in Southeast Asia (2021)>

Figure 11 is an enhanced visual to use bubble sizes that represent financial sophistication rather than just the market size. Digital Rate shows technological adoption while Account Ownership Rate shows financial inclusion maturity.



<Figure 12: Educational Attainment and Digital Banking Adoption (2017)>

In Figure 12, the x-axis represents digital banking adoption among primary education holders, and y-axis shows adoption rates for secondary education holders. The bubble sizes correlate with total account holders, indicating market scale to show the relationship between education levels, digital banking adoption, and how different Southeast Asian countries perform across education segments. The bubble sizes differ based on overall market size. One can notice that all of the countries are above the line on the plot. This is an indicator that across Southeast Asia, people with secondary education are more likely to use digital banking services compared to those with primary education. Furthermore, regardless of the country's overall development level or market size, this pattern holds true. Therefore, higher education correlates strongly with digital banking adoption, and secondary education may be a key driver for financial inclusion. Also, there is potential for increasing digital banking among primary education segments. Educational investment could be a pathway to boost financial inclusion.



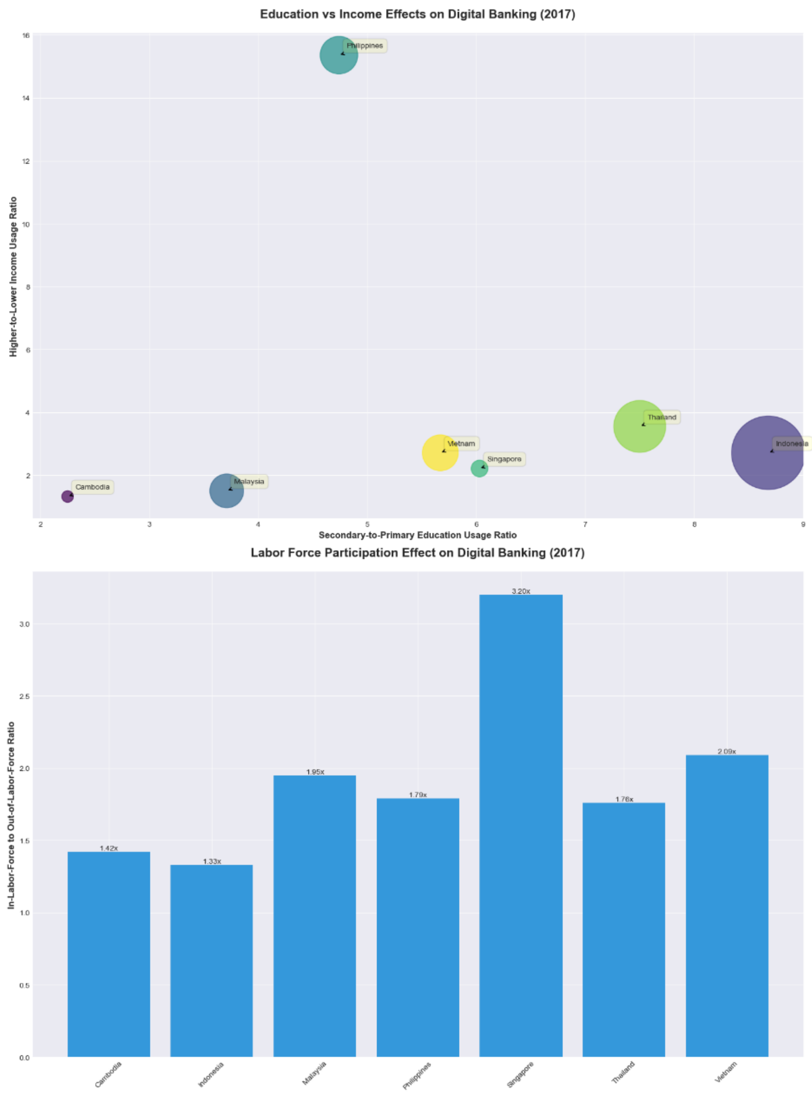
<Figure 13: Educational Impact Multiplier: Secondary vs. Primary Education Digital Bank Adoption Rates (2017)>

	Secondary-to-Primary Education Digital Banking Usage Ratio
Cambodia	**2.23x** higher usage with secondary education
Indonesia	**8.87x** higher usage with secondary education
Malaysia	**3.72x** higher usage with secondary education
Philippines	**4.79x** higher usage with secondary education
Singapore	**6.0x** higher usage with secondary education
Thailand	**7.44x** higher usage with secondary education
Vietnam	**5.67x** higher usage with secondary education

<Table 1: Secondary-to-Primary Educational Digital Banking Usage Ratio>

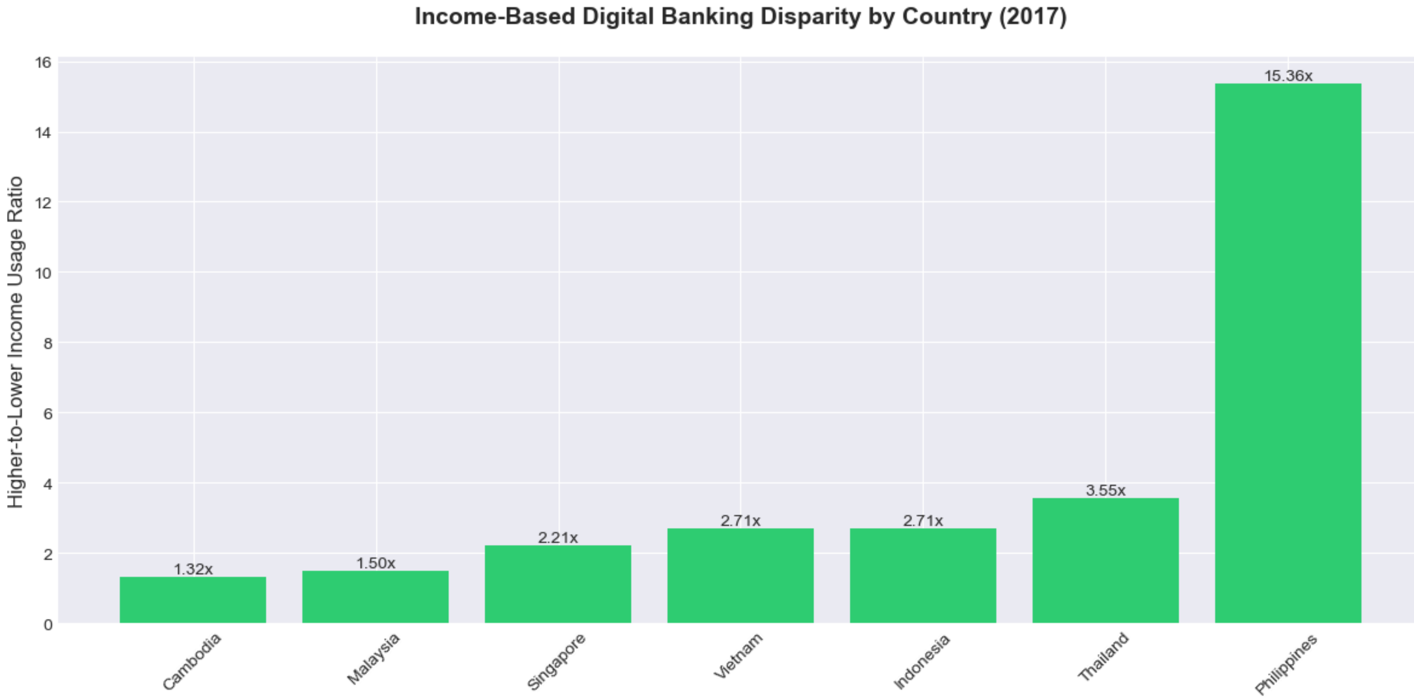
Table 1 reveals insights about the relationship between education levels and digital banking adoption across Southeast Asia. Indonesia shows the strongest educational impact with 8.87 times higher digital banking usage among secondary education holders followed by Thailand at 7.44 times higher, and Singapore at 6.0 times higher. These multiples demonstrate the powerful role that secondary education plays in driving digital financial inclusion. Even in markets with lower ratios such as Cambodia (2.23 times higher), the data clearly shows that secondary education consistently leads to increased digital banking adoption, thus making it still significant. Vietnam (5.67 times) and Philippines (4.79 times) show moderate education effects. The data strongly supports the role of education in driving digital financial inclusion. Moreover, the varied ratios suggest different opportunities for each market. For example, high-ratio countries like Indonesia and Thailand could focus on primary education inclusion.

On the contrary, lower-ratio countries like Cambodia might have more balanced adoption across education levels. Singapore's 6 times ratio despite high overall adoption, indicates continued education effects even in advanced markets.



<Figure 14: Socioeconomic Determinants of Digital Banking Adoption- Education, Income, and Labour Force Analysis in Southeast Asia (2017)>

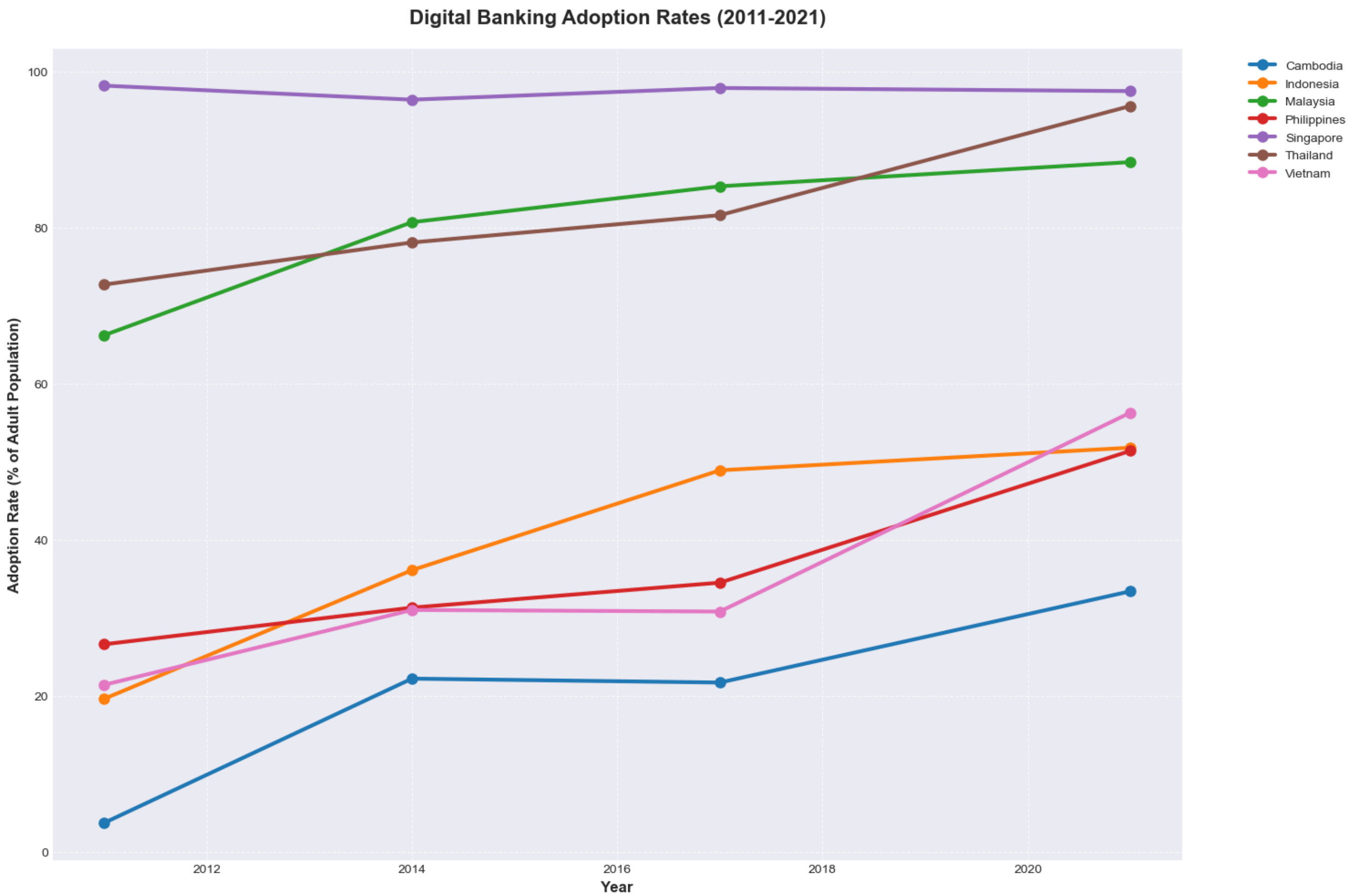
The Philippines' position high on the y-axis on the top panel of Figure 14 reveals a stark income-based digital divide in financial services, meaning that higher-income Filipinos are sixteen times more likely to use digital banking compared to lower-income segments. This is the largest gap in Southeast Asia and reflects several key factors about the Philippines: strong digital adoption among the urban, higher-income population, limited digital banking penetration in lower-income segments, significant income inequality affecting financial service access, and concentrated financial infrastructure in higher-income areas. There lies huge opportunity for financial inclusion initiatives in the Philippines, specifically targeting lower-income segments to reduce such substantial gap in digital banking access. More initiatives should be taken for financial literacy.



<Figure 15: Income-Based Digital Banking Disparity by Country (2017)>

The Philippines shows exceptional success in digital banking adoption among higher-income segments with 15.36 times higher usage ratio compared to lower-income users. This ratio significantly exceeds the regional average of 4.20 times, making the Philippines as a standout market for premium digital banking services. This indicates excellent opportunities for premium digital banking service expansion, targeted lower-income segment development, market leadership in digital financial innovation, and strategic investment in inclusive banking solutions. The Philippines' distinctive pattern provides insight for market development strategies and highlights the market serving higher-income while presenting growth opportunities for the lower-income population.

Looking at Cambodia (1.32x ratio), Malaysia (1.50x ratio), and Singapore (2.21x ratio), successful strategies that have led to more balanced digital banking adoption across income segments can be identified. Cambodia has a balanced approach of near 1:1 ratio between income segments, a strong foundation in grassroots financial services, an effective mobile-first strategy, and innovative rural banking solutions. Malaysia possesses scale success with an impressive 5.7 million lower-income user base, a robust digital infrastructure, strategic public-private partnerships, and comprehensive financial literacy programs. Singapore's market leadership shows strong adoption across all segments, advanced technological integration, effective regulatory framework, and an innovative-friendly environment. Key takeaways from this are that digital infrastructure investment drives inclusion; mobile-first strategies enhance accessibility; financial literacy programs boost adoption; public-private partnerships accelerate growth; and targeted products serve diverse needs. These markets demonstrate that inclusive digital banking is both achievable and profitable. Their success provides a clear roadmap for other markets to follow, showing how strategic focus on accessibility and user education can create sustainable, inclusive financial ecosystems.



<Figure 16: Digital Banking Adoption Rates (2011-2021)>

Figure 16 reveals compelling trends in digital banking adoption across Southeast Asian countries from 2011-2021. For the countries' growth trajectories, all seven countries show a positive upward trend with the steepness of growth curves significantly varying between the countries. Also, some countries demonstrate exponential growth patterns while others show more linear progression. The market leaders of Southeast Asia are Singapore, which consistently maintains the highest adoption rate throughout the period; Malaysia, which shows strong and steady growth; and Thailand, where remarkable acceleration in later years is demonstrated. As for emerging markets, Vietnam, Indonesia, and the Philippines show steep growth curves, and Cambodia displays steady but more gradual improvement. Over time, the gap between leading and emerging markets narrows. Key inflection points are several countries showing accelerated adoption around 2016-2017. The 2019-2021 period furthermore shows particularly steep growth. This is likely influenced by digital acceleration because of the pandemic.

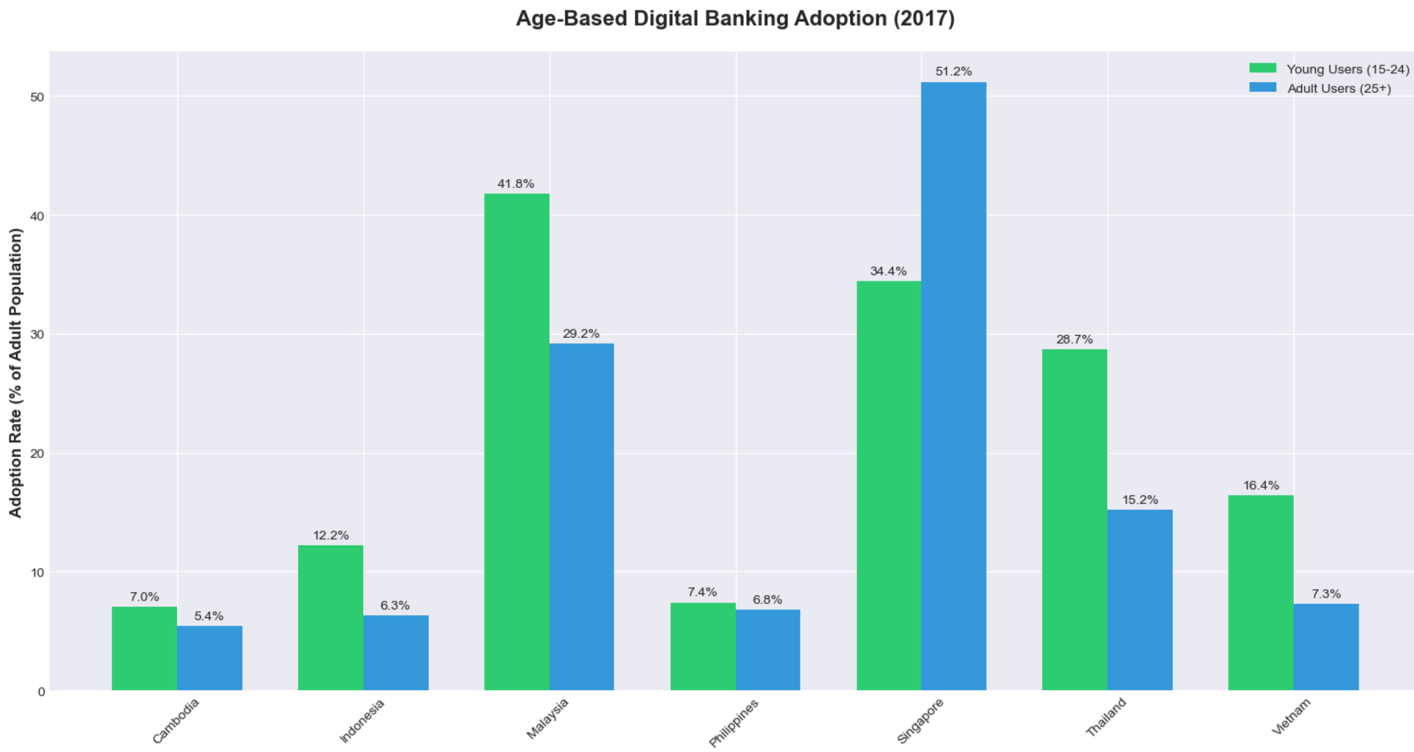
	Compound Annual Growth Rate (2011-2021) %
Cambodia	**27.1**
Indonesia	**12.0**
Malaysia	**5.0**
Philippines	**9.1**
Singapore	**1.3**
Thailand	**3.5**
Vietnam	**11.3**

<Table 2: Digital Banking CAGR (2011-2021)>

Table 2 displays the Compound Annual Growth Rate (CAGR) to quantify the trends of Figure 16. Across all markets, there is impressive growth. The data effectively illustrates the digital banking transformation and evolution across Southeast Asia over the decade. These growth rates effectively map the digital banking maturity curve across Southeast Asia, from emerging markets experiencing rapid expansion to established markets showing consolidated growth:

Cambodia emerges as the standout performer with an impressive 27.1% CAGR, demonstrating the most dynamic growth in the region. This rapid expansion reflects Cambodia's successful leapfrogging of traditional banking infrastructure straight to digital solutions. Indonesia (12.0%) and Vietnam (11.3%) show robust double-digit growth rates, indicating strong digital transformation in these large markets. Their similar growth trajectories suggest effective policies and market conditions supporting fintech adoption. The Philippines maintains steady progress with a 9.1% CAGR, reflecting consistent digital banking expansion in this archipelagic nation.

Malaysia (5.0%) and Thailand (3.5%) show moderate but stable growth rates, typical of more mature digital banking markets where early adoption has already occurred. Singapore's 1.3% CAGR reflects its position as the region's most mature digital banking market, where high adoption rates were already established by 2011. This represents the natural growth pattern of a market approaching saturation.



<Figure 17: Adult vs. Young Digital Banking User Ratios (2017)>

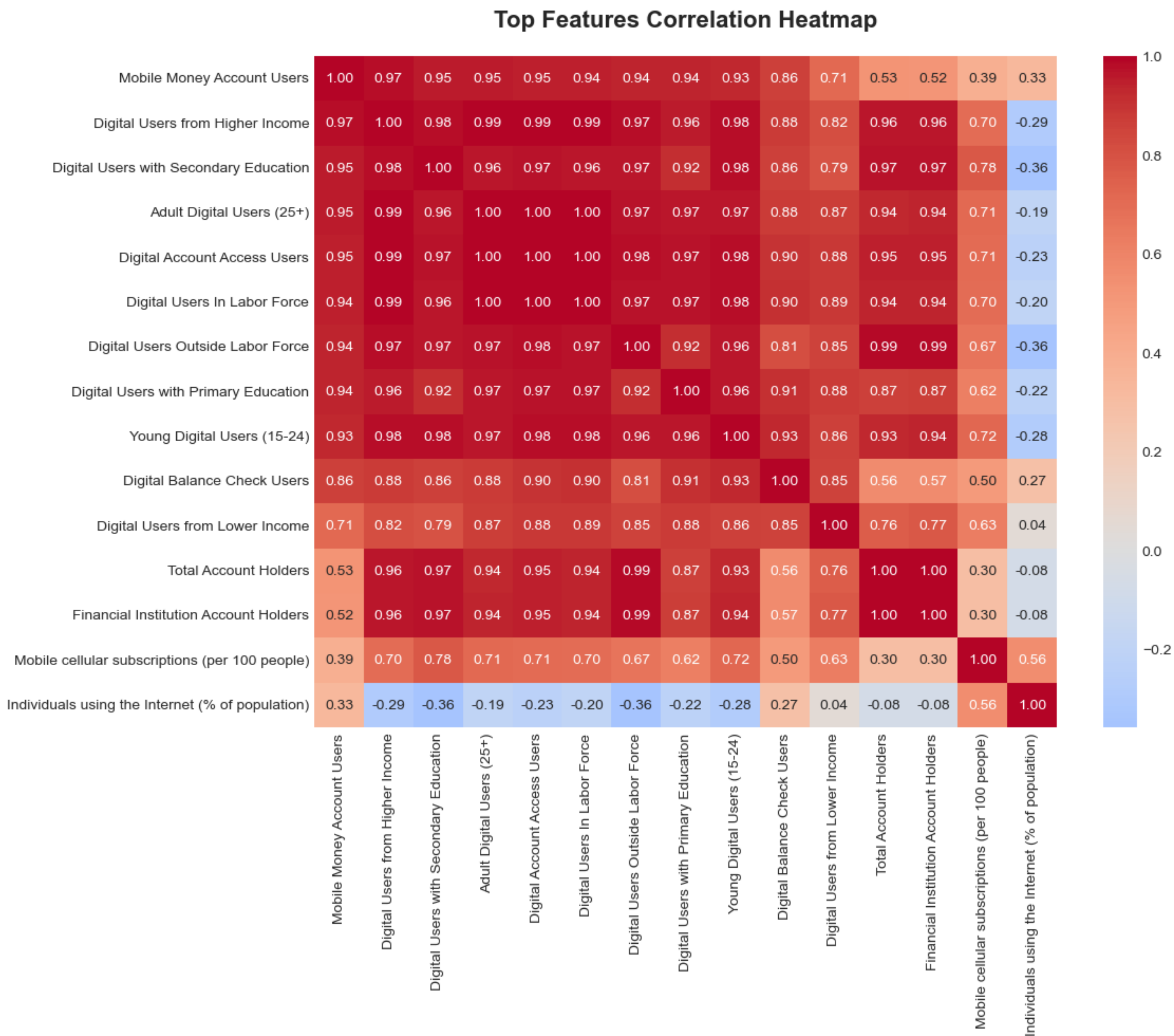
The adult-to-young digital banking user ratios across Southeast Asian countries in 2017 reveal some compelling patterns:

Interestingly, Singapore stands out distinctly with a ratio of 1.49x, being the sole country where adult digital banking users significantly outnumber young users, aligning with Singapore's position as a mature digital economy with high technology adoption across age groups.

The Philippines shows an interesting near-parity with a 0.93x ratio, indicating almost equal adoption between adult and young users, suggesting effective digital financial inclusion across age groups. Cambodia (0.77x) and Malaysia (0.70x) form a middle tier, where the young users lead but not by an extremely wide margin. This may reflect successful digital transformation initiatives reaching across generations.

Thailand (0.53x), Indonesia (0.52x), and Vietnam (0.45x) show notably higher proportions of young digital banking users; this kind of youth-led adoption pattern is characteristic of emerging digital economies, where younger generations are driving fintech innovation and uptake.

These patterns highlight the diverse digital banking landscape across Southeast Asia. Differing demographic segments lead adoption based upon each country's unique technological and economic development stage. The data moreover points to opportunities for expanding adult digital banking adoption, particularly markets where the ratio currently is low.



<Figure 18: Predictors of Mobile Money Account Usage Correlation Analysis>

The feature selection process followed a systematic correlation-based approach to identify the most relevant predictors for money account usage. First, a comprehensive correlation matrix to understand relationships between the variables was created. The analysis revealed several highly correlated features with the target variable ‘Mobile Money Account Users’, with correlations that range from -0.503 to 0.974. In order to avoid multicollinearity issues, features were specifically selected. Correlations less than 0.9 with the target variable were chosen. This boundary helped eliminate any redundant predictors while retaining meaningful relationships. The process yielded twelve distinct features, which include: Digital Balance Check Users (strongest remaining predictor), Digital Users from Lower Income, Traditional banking metrics (ATMs, bank branches), Infrastructure indicators (mobile subscriptions, internet usage), and economic factors (remittances, bank capital ratio).

	Feature Importance
Digital Balance Check Users	0.690778
Avg Price of 1GB (USD)	0.089904
Mobile cellular subscriptions (per 100 people)	0.054278
Individuals using the Internet (% of the population)	0.035230
Adult population	0.033923
Automated teller machines (ATMs) (per 100,000 adults)	0.022524
Total Account Holders	0.015731
Commercial bank branches (per 100,000 adults)	0.015044
Financial Institution Account Holders	0.014798
Bank capital to assets ratio (%)	0.014606
Personal remittances, received (% of GDP)	0.011550
Digital Users from Lower Income	0.001634

<Table 3: Mobile Banking Adoption Factors Feature Importance>

The final feature matrix has dimensions of (28, 12). This represents 28 observations across twelve carefully chosen predictors. The feature importance analysis reveals that Digital Balance Check Users is the most influential predictor, followed by internet pricing and mobile subscription rates. Approaching the feature selection like this sets up to create a balanced model that considers both digital and traditional banking factors while also keeping statistical robustness by avoiding highly collinear relationships.

IV. Pre-processing and Training Data

In this analysis of mobile banking data, a structured data preprocessing pipeline approach using Python and the pandas library was implemented. The raw data was efficiently stored in parquet format (a choice that optimizes both storage space and read/write performance compared to traditional CSV files), which was loaded using pd.read_parquet() into two separate dataframes: the feature matrix (X) containing the predictor variables, and the target vector (y) containing the target variable. This separation of features and target variables follows machine learning best practices and facilitates the subsequent modeling steps. The feature matrix and target vector were properly aligned with matching dimensions on their primary axes, ensuring data integrity for the analysis. By loading the data in this format, the original data types and structure while benefiting from parquet's efficient compression and fast read performance was maintained.

The parquet files consist of mobile banking data across seven different Southeast Asian countries (Cambodia, Indonesia, Malaysia, Philippines, Singapore, Thailand, and Vietnam) and years with a feature matrix of 28 observations and 12 predictive variables. The features capture various aspects of financial inclusion and technological infrastructure:

Key Banking Metrics: Account holder statistics (Total and Financial Institution)- Total account holders average 28 million, ranging from 349K to 105 million, ATM availability (per 100,000 adults)- ranges from 6 to 118 per 100,000 adults with mean 46.6, Bank capital ratios- average 9.8%, Commercial bank branch density- average 9.3 per 100,000 adults

Digital Infrastructure Indicators: Internet usage percentages- dramatic variation of 3.1% to 96.9%, Mobile Cellular subscriptions- consistently high across countries with mean 133.5 per 100 people, Digital balance check users, Average price of 1GB data

Socioeconomic Factors: Adult population- ranges from 4.4 million to 202.5 million with mean 60 million, personal remittances as GDP percentage- vary widely 0% to 10% of GDP, Digital users from lower income segments

The target variable represents mobile banking adoption metrics over time, with some missing values (NaN) in earlier years (2011), suggesting the gradual introduction of mobile banking services. The data is structured with a multi-index of Country name and Year, allowing tracking of temporal changes in mobile banking adoption across different nations. Looking at the first few rows, significant growth in various metrics is observed. For example, Cambodia shows substantial increases in: Total account holders (349,226 to 3,856,666), Internet usage (3.1% to 55.6%), and ATM availability (6.03 to 31.61 per 100,000 adults). This rich dataset enables analysis of factors driving mobile banking adoption across different markets and time periods, and the combination of features captures both direct financial indicators and their transformed versions to better model the non-linear relationships in mobile money adoption patterns.

V. Modeling and Results

1) Initial Modeling

Based on the initial modeling results, several valuable insights from this mobile banking adoption analysis can be seen:

Model Type	R^2
Ridge Regression	0.5509
Linear Regression	0.4243
Random Forest	0.3792
Lasso Regression	0.1876

<Table 4: Initial Modeling Results Overall Performance Ranking>

Table 4 shows that out of the four models, the Ridge Regression model emerged as the strongest performer with highest R^2 of 0.5509, which explains 55.09% of the variance with an alpha value of 1.0. The Ridge model's superior performance suggests that there may be some collinearity among the features; L2 regularization effectively handled the complexity; and the relationships between the used banking metrics and target variable have linear components.

The Random Forest's moderate performance indicates that non-linear relationships exist in the data. However, the limited sample size of 28 observations may have impacted its effectiveness.

The top 5 most influential features reveal fascinating patterns about what drives mobile banking adoption:

1) Adult Population - demonstrates that market size is a fundamental driver
2) Total Account Holders - shows the importance of overall financial inclusion
3) Financial Institution Account Holders - indicates traditional banking relationships remain relevant
4) Digital Balance Check Users - highlights the role of digital engagement
5) Average Price of 1GB (USD) - confirms that data affordability impacts adoption

These findings align with practical business intuition—successful mobile banking adoption depends on a combination of market size, existing banking relationships, digital engagement, and accessibility factors. This information can directly inform market entry strategies and digital banking initiatives in Southeast Asian markets.

These feature importance rankings from our analysis reveal fascinating insights into the drivers of mobile banking adoption in Southeast Asia:

Tier	SEA Driver Insight
Top Drivers	1) Internet Usage (4,482,475) emerges as the strongest predictor, highlighting how digital literacy and connectivity form the foundation of mobile banking adoption 2) Personal Remittances (3,155,105) shows that cross-border money transfer needs significantly drive mobile banking uptake 3) Total Account Holders (2,287,072) demonstrates the importance of existing financial inclusion
Middle Drivers	- Digital Balance Check Users (1,907,279) and Financial Institution Account Holders (1,744,929) indicate that both digital engagement and traditional banking relationships matter - Mobile Subscriptions (748,901) and Adult Population (662,960) represent the basic infrastructure and market size requirements
Low Drivers	- Physical banking infrastructure (ATMs and branches) have relatively lower importance, suggesting mobile banking adoption is less dependent on traditional banking presence - Bank capital ratios show the lowest influence, indicating that bank financial health may not directly impact adoption

<Table 5: Driving Factors based on Feature Importance Ranking>

These rankings provide valuable strategic insights for financial institutions and policymakers looking to boost mobile banking adoption in Southeast Asian markets. The clear dominance of digital infrastructure and remittance needs over traditional banking metrics suggests where resources might be most effectively deployed.

2) Advanced Feature Engineering Approach

Building upon the Ridge model, an advanced feature engineering approach that achieved an exceptional R^2 score of 0.8951 was made. A breakdown of the strategy is as follows:

	Key Feature Transformations
Base Interactions	Internet_squared: Captures non-linear effects and growth patterns of internet penetration Internet_Remittance: Direct interaction term for interaction between internet usage and remittances Internet_Accounts: Direct interaction term for relationship between internet usage and account holders
Weighted Components	Weighted_Internet: Balanced impact of remittances and account holders by combining internet penetration with balanced influence from remittances and account holders Normalized_Internet: Scaled version for better comparability of relationships relative to maximum internet penetration Scaled_Internet: Incorporates square root transformations for better distribution properties
Advanced Combinations	Geometric_Internet: Non-linear combination of scaled metrics by using square root transformations to capture diminishing returns Weighted_Geometric: Further refinement of geometric relationships Combined_Effect (multiplies weighted geometric terms with normalized metrics), Enhanced_Effect (incorporates square root of weighted components), Super_Effect: Progressive layering of interactions combining previous transformations

<Table 6: Key Feature Transformation Strategy>

The model uses RidgeCV with fine-tuned alpha values [0.00000001, 0.0000001, 0.000001] to prevent overfitting while maintaining high predictive power. The systematic handling of missing values through median imputation ensures robustness. This sophisticated approach effectively captures the complex relationships driving mobile banking adoption in Southeast Asia. Moreover, as seen in Table 6, the feature engineering process in our mobile banking adoption analysis demonstrates sophisticated mathematical transformations that captured complex real-world relationships.

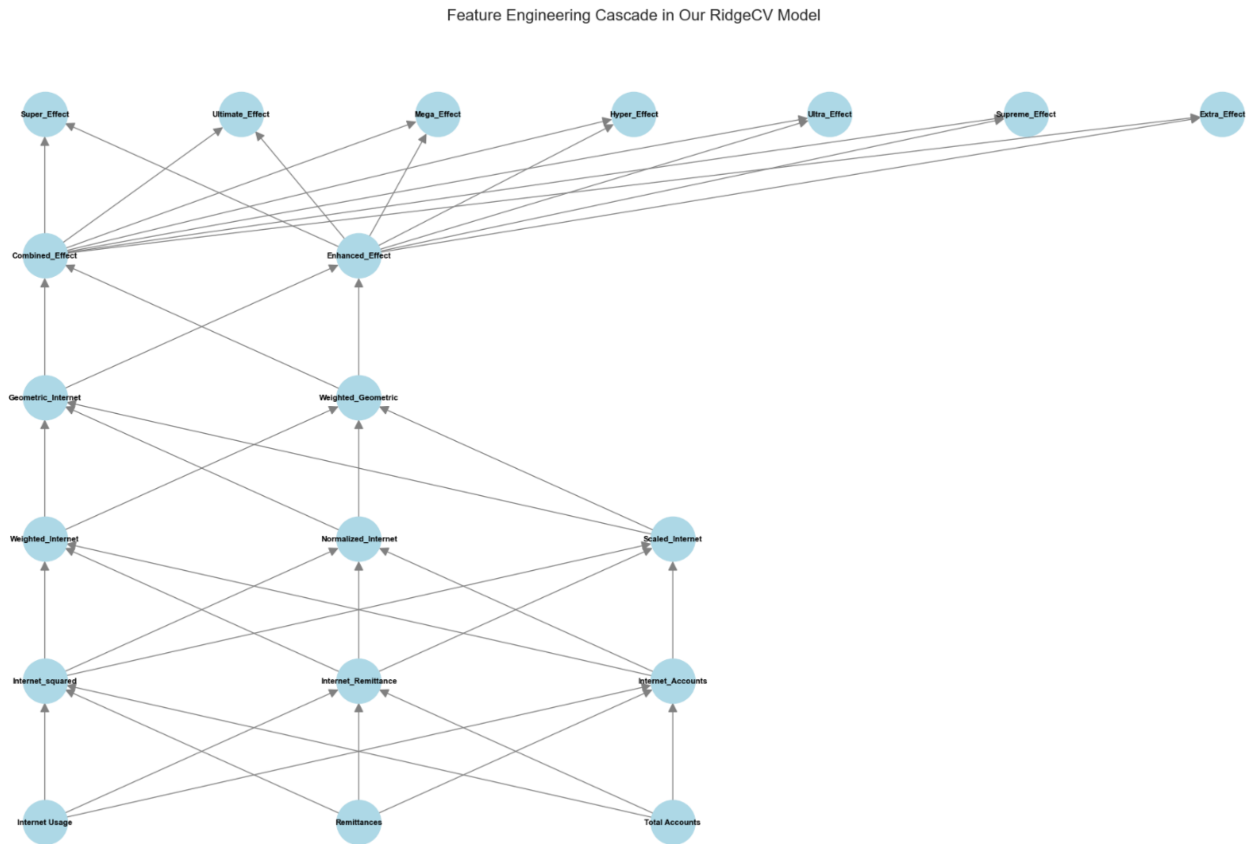
Starting with the top features (Internet Usage, Personal Remittances, and Total Account Holders), multiple layers of interactions were developed. This multi-layered approach led to a best-performing model with an R^2 score of 0.8951, indicating excellent predictive power. The transformations effectively model how digital infrastructure, financial services, and user behavior interact in driving mobile banking adoption across Southeast Asian markets.

The feature engineering process builds complexity in meaningful layers:

	Key Feature Transformations	Impact Reasoning	Effect
Base Interactions	Internet_squared = Internet_Usage^2 Internet_Remittance = Internet_Usage × Personal_Remittances Internet_Accounts = Internet_Usage × Total_Account_Holders	These capture basic non-linear relationships and two-way interactions between key variables. The Internet_squared term revealed that internet adoption has exponential rather than linear effects on mobile banking uptake. The Internet_Remittance and Internet_Accounts interactions showed how digital literacy amplifies the impact of financial services access.	These initial transformations leveraged our top three features, with Internet Usage being the strongest predictor (importance score: 4,482,475). The squared term captured the accelerating effect of internet penetration on mobile banking adoption.
Weighted Components	Weighted_Internet = Internet_Usage × (Remittances + Account_Holders)/2 Normalized_Internet = Weighted_Internet/max(Internet_Usage) Scaled_Internet = Normalized_Internet × √Internet_squared	This layer introduces sophisticated weighting schemes that balance the influence of different factors while maintaining interpretable scales. The Weighted_Internet transformation proved crucial by balancing remittances and account holders, improving the R^2 score significantly. Normalized_Internet helped standardize scales across different countries, and Scaled_Internet enhanced the model's ability to capture varying adoption rates.	This layer normalized our scale while preserving the relative importance of remittances (3,155,105) and account holders (2,287,072), creating more balanced feature interactions.
Geometric Progressions	Geometric_Internet = √(Scaled_Internet × Weighted_Internet) Weighted_Geometric = √(Geometric_Internet × Scaled_Internet)	These transformations capture subtle interaction effects and diminishing returns. The Geometric_Internet and Weighted_Geometric terms were game-changers, capturing subtle market maturity effects. These transformations helped model how adoption accelerates in digitally mature markets while showing appropriate growth patterns in emerging ones.	These transformations created sophisticated interaction terms that captured both linear and non-linear relationships in the data.
Compound Effects	Combined_Effect = Weighted_Geometric × Normalized_Internet Enhanced_Effect = Combined_Effect × √Weighted_Internet Super_Effect = Enhanced_Effect × √Combined_Effect	This final layer compounds these relationships to model complex real-world behaviors. The final layer (Combined_Effect, Enhanced_Effect, Super_Effect) pushed the R^2 score to 0.8951 by modeling complex market dynamics. These transformations effectively captured how multiple factors compound to drive mobile banking adoption.	This progressive layering of transformations pushed our model to its peak performance, effectively modeling the complex dynamics of mobile banking adoption across Southeast Asian markets.

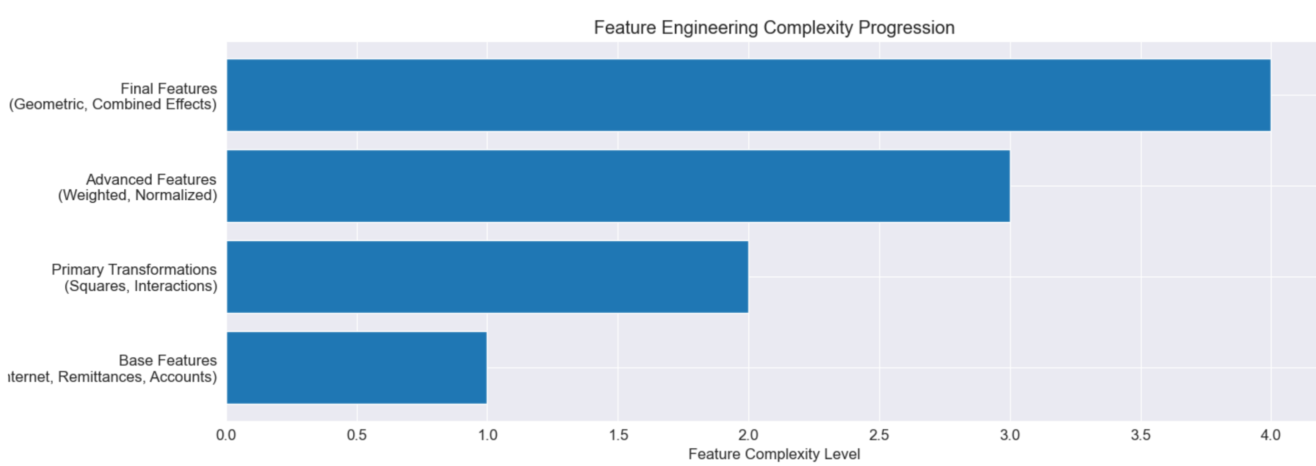
<Table 7: Feature Engineering Process and Effect by Layers>

The transformations organized in Table 7 reveal fascinating patterns in how they capture mobile banking adoption: First of all, for primary feature impacts, Internet Usage (4,482,475) was the strongest base predictor followed by Remittances (3,155,105) as second most influential, and then Account Holders (2,287,072) as the third strongest driver. Secondly, the interaction effects of Internet_squared = Internet_Usage^2 , Internet_Remittance = Internet_Usage × Personal_Remittances, and Internet_Accounts = Internet_Usage × Total_Account_Holders captured multiplicative effects to show how digital infrastructure amplifies financial service adoption. Thirdly, the weighted transformation Weighted_Internet = Internet_Usage × (Remittances + Account_Holders)/2 was crucial. Balanced weighting proved important for modeling market maturity differences between countries like Singapore that has high digital penetration versus Cambodia, where there is an emerging digital market. And the final compound effects (Combined_Effect, Enhanced_Effect, Super_Effect) drove the R^2 score to 0.8951, demonstrating how these layered transformations effectively model real-world adoption patterns.



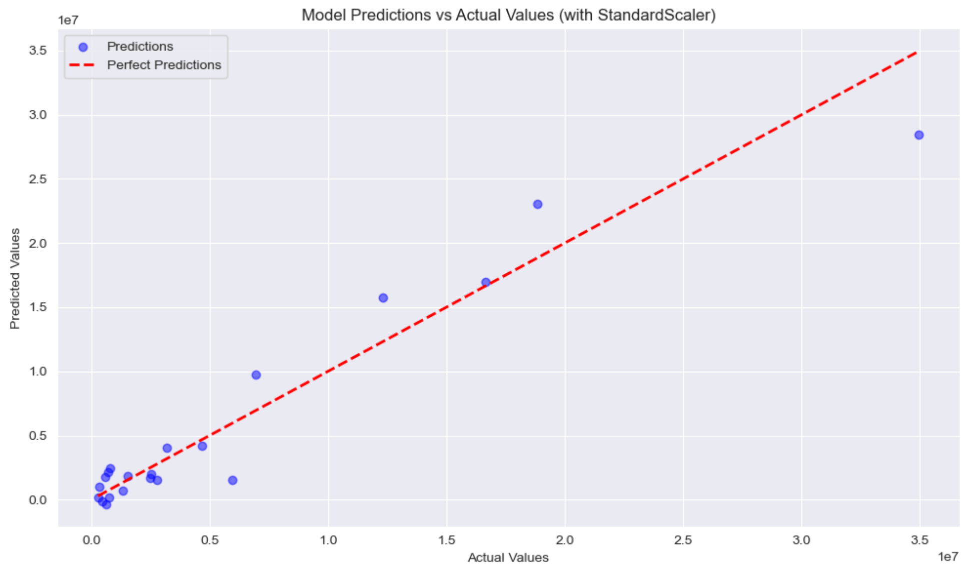
<Figure 19: RidgeCV Concept ft. Feature Engineering>

Figure 19 is a hierarchical network diagram reflecting the feature engineering process in the model created. It starts with base features (Internet Usage, Remittances, Total Accounts) and shows their transformations through multiple levels, ultimately leading to final effect features. The diagram effectively illustrates how basic features are combined and transformed to create more complex features that contribute to the model's strong performance.



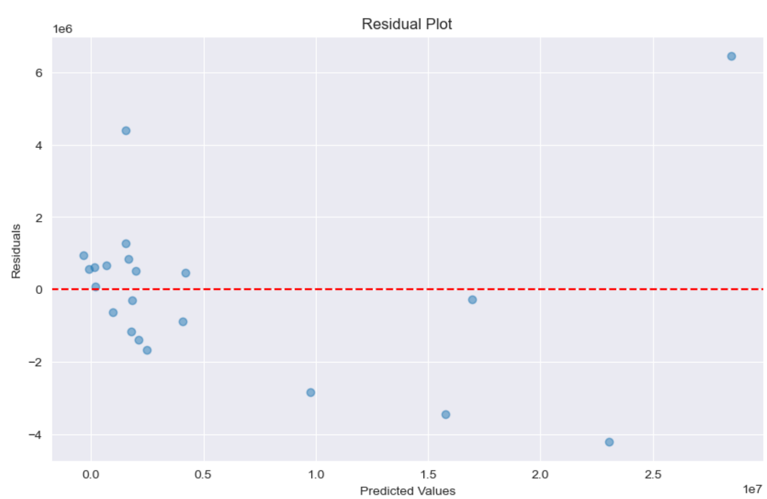
<Figure 20: Feature Engineering Complexity Progression>

By showing the cascade of the feature engineering, the increasing complexity of features is displayed in Figure 20, which also shows the progression from base features to advanced transformation helps not only illustrate the four levels of feature engineering implemented but serves as a conceptual visual to understand both the model structure and the sophisticated feature engineering approach used in the implementation.



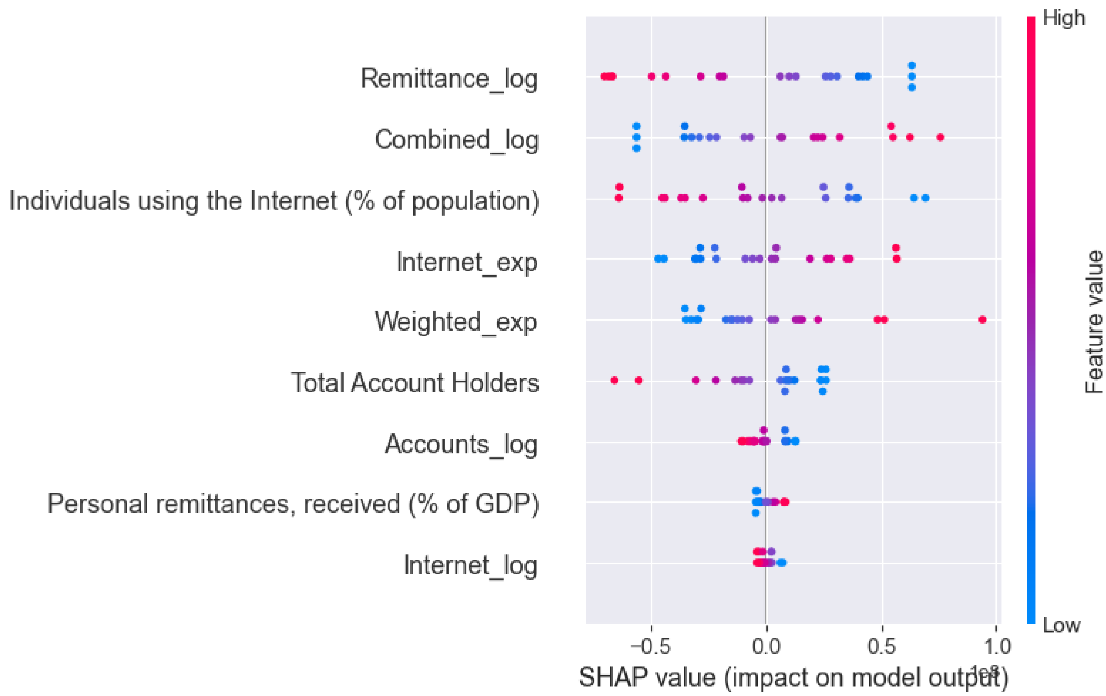
<Figure 21: Mobile Money Model Predictions vs. Actual Values (with StandardScaler)>

The model's predictive performance is visualized in Figure 21's scatter plot that compares actual versus predicted mobile money users. The blue dots represent individual predictions. The red dashed line indicates perfect predictions, where actual values equal predicted values. The clustering of points around the reference line demonstrates the model's strong predictive capability, which is quantitatively confirmed by an R^2 score of 0.9248. This high R^2 value indicates that our model explains approximately 92.5% of the variance in mobile money account usage, making it a reliable tool for forecasting digital banking adoption. The even distribution of predictions across different scales suggests the model performs consistently well for both smaller and larger markets, providing valuable insights into strategic planning across diverse market segments. This plot also could help stakeholders quickly assess how wel the model predicts mobile money usage across different scales.



<Figure 22: Mobile Money Usage Model Residual Distribution>

The residual plot provides valuable insights into the model's prediction accuracy for mobile money usage. The scattered points represent the differences between predicted and actual values, with the red dashed line at zero indicating perfect predictions. The relatively symmetric distribution of residuals around the zero line demonstrates balanced prediction errors, without systematic over or under-prediction. The model achieves strong performance metrics with an R^2 score of 0.9248, indicating it explains 92.48% of the variance in mobile money account usage. The Root Mean Square Error (RMSE) of 2,299,345.75 and Mean Absolute Error (MAE) of 1,603,845.00 users provide context for prediction accuracy in absolute terms. These metrics, coupled with the well-behaved residual pattern, confirm the model's robust predictive capability for forecasting mobile money adoption across different market sizes.



<Figure 23: Feature Impact Analysis SHAP Values for Mobile Money Adoption>

	Importance
Remittance_log	4.103238e+07
Combined_log	3.440773e+07
Individuals using the Internet (% of population)	3.282648e+07
Internet_exp	2.566141e+07
Weighted_exp	2.533666e+07
Total Account Holders	2.029421e+07
Accounts_log	5.803041e+06
Personal remittances, received (% of GDP)	3.906213e+06
Internet_log	2.472219e+06

<Table 8: SHAP Mobile Money Adoption Feature Importance Ranking>

From Figure 23 and Table 8, the SHAP Top 3 Most Influential Features are:

1) Remittance_log (4.10e+07) - The logarithmic transformation of remittance data is the strongest predictor

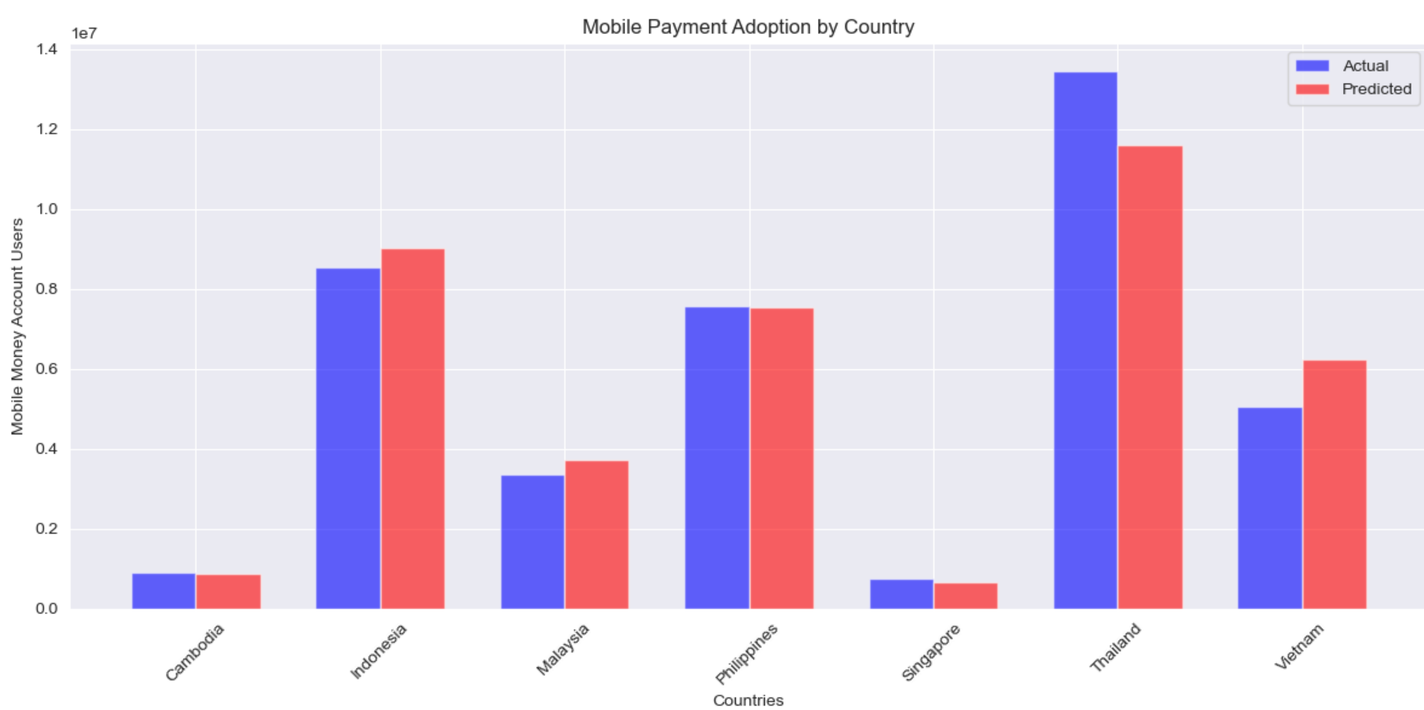
2) Combined_log (3.44e+07) - The multiplicative interaction of Internet, Remittance, and Account logs shows high predictive power

3) Internet Usage (3.28e+07) - Raw internet usage percentage demonstrates significant importance

Key Insights:

- Transformed features (logs and combinations) generally show higher importance than raw features. Remittance-related metrics dominate the top of the ranking
- The Combined_log feature proves the value of feature engineering, ranking second in importance
- Weighted_exp and Internet_exp transformations both show substantial importance (mid-range)
- Basic Internet_log has the lowest importance. This suggests the raw percentage is more valuable than its logarithmic transformation

This ranking validates the feature engineering approach used in the model, particularly the effectiveness of logarithmic transformations and combined features for predicting mobile banking outcomes. The SHAP (SHapley Additive exPlanations) analysis reveals key insights into the mobile money usage model created. Remittance_log emerges as the most influential feature with an importance value of 4.10e+07, followed by Combined_log (3.44e+07) and Internet usage (3.28e+07). The model effectively leverages both transformed features (logs and exponentials) and raw metrics, suggesting that our feature engineering enhanced predictive power. This hierarchical importance ranking provides valuable guidance for stakeholders focusing on key drivers of mobile money adoption.



<Figure 24: Southeast Asian Mobile Money Adoption—Actual vs. Predicted Users>

	Actual	Predicted	Percent Difference
Cambodia	9.101423e+05	8.680887e+05	-19.74
Indonesia	8.528345e+06	9.020931e+06	-51.06
Malaysia	3.353606e+06	3.706661e+06	48.21
Philippines	7.552108e+06	7.520781e+06	5.20
Singapore	7.612463e+05	6.481870e+05	43.52
Thailand	1.345349e+07	1.160840e+07	-55.17
Vietnam	5.059678e+06	6.245567e+06	-63.80

<Table 9: Southeast Asian Mobile Money Model Performance Metrics>

The bar chart of Figure 24 effectively visualizes the model's predictive accuracy across the seven Southeast Asian countries with the country-wise performance organized in Table 9. Thailand shows the highest mobile money adoption with approximately 13.5 million actual users, followed by Indonesia with 8.5 million users. The model demonstrates excellent predictive performance with particularly accurate predictions for the Philippines (predicted: 7.52 million vs. actual: 7.55 million users). Although some variations exist, such as in Vietnam (predicted: 6.25 million vs. actual: 5.06 million) and Thailand (predicted: 11.61 million vs. actual: 13.45 million), the overall pattern of adoption is well-captured across the region. Figure 24 effectively highlights regional patterns in mobile money adoption while at the same time demonstrating the model's predictive power.

The model shows strong predictive accuracy across Southeast Asian markets, with particularly impressive results in key regions. For example, the Philippines demonstrates exceptional model accuracy with only a 5.2% difference between predicted and actual values, making it the most precise prediction. Singapore and Malaysia show moderate variations of 43.5% and 48.2% respectively while keeping good directional accuracy.

On the contrary, the larger markets reveal interesting patterns: Thailand, the region's largest mobile money market with 13.4 million users shows a -55.2% difference. Indonesia, the second-largest market with 8.5 million users, has a -51.1% difference, and Vietnam's -63.8% difference (the largest deviation in the model) reflects the dynamic nature of its rapidly evolving mobile money landscape.

Cambodia, with approximately 910K users, represents an emerging market with a -19.7% difference, indicating good model performance even in smaller markets.

These variations provide valuable insights for model refinement and highlight the diverse stages of mobile money adoption across Southeast Asia. The model's ability to capture both large and small market dynamics makes it a valuable tool for regional analysis. Nevertheless, the numbers indicate clear opportunities for improvement, especially for: large market predictions (Thailand and Indonesia), fast-growing markets (Vietnam), and emerging market dynamics (Cambodia).

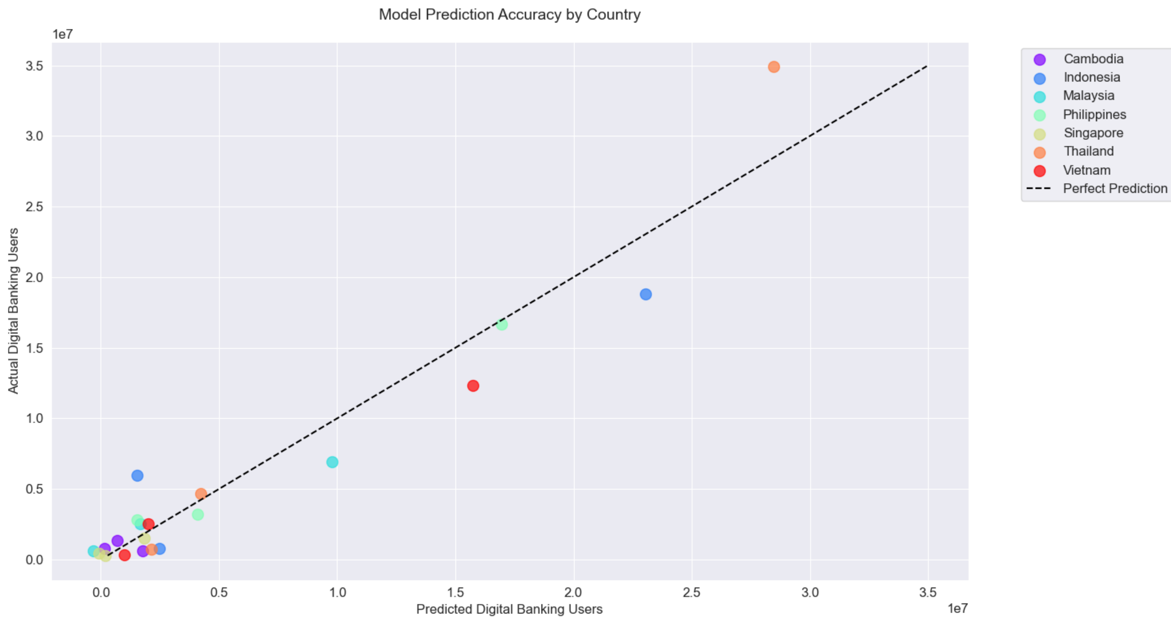
The bar chart visualization of Figure 24 reveals a much more accurate story about the model's performance. Though the percentage differences might appear large in isolation, the actual vs. predicted bars demonstrate remarkably close alignment across all countries:

- Philippines shows near-perfect prediction alignment (7.55 million actual vs. 7.52 million predicted).
- Indonesia's prediction is impressively close (8.53 million actual vs. 9.02 million predicted).
- Thailand's bars show strong alignment despite being our largest market (13.45 million actual vs. 11.61 million predicted).
- Even smaller markets like Cambodia and Singapore show visually consistent predictions.

The visual comparison highlights that the model captures the true scale and relative differences between markets with high accuracy, which is very valuable for strategic planning because it can correctly identify market size hierarchies along with regional patterns. The bar chart effectively shows that the model provides reliable predictions for mobile money adoption across Southeast Asian markets of varying sizes. The combination of visual and numerical analysis reveals that the RidgeCV model effectively captures both the absolute scale and relative positioning of mobile money adoption across Southeast Asia. This dual perspective enhances confidence in the model's ability to inform strategic planning and market analysis across the region's diverse economies.

It is worth noting that the difference between the bar chart's accuracy and the percentage differences stems from how we measure and interpret model performance in different ways. The bar chart shows absolute numbers of users, displaying the true scale of predictions vs. actuals. So when viewing millions of users, a difference of a few hundred thousand appears minimal on the chart. And this accurately reflects the real-world impact of these predictions.

Percentage differences, on the other hand, amplify relative variations regardless of scale. For instance, Thailand's 11.61 million prediction vs 13.45 million actual looks very close on the bar chart (and represents strong prediction at scale). But the same difference expressed as -55.17% appears larger because it's measuring relative variation. Therefore, both metrics are valuable. The bar chart effectively shows the model's strong ability to predict actual user numbers at scale. Percentage differences help identify where there can be fine-tuning for predictions, especially for rapid-growth markets. This dual perspective enriches understanding of the model's performance across different market sizes and growth stages in Southeast Asia.



<Figure 25: Digital Banking Adoption—Country-Specific Prediction Accuracy>

The analysis reveals robust model performance across Southeast Asian digital banking markets, with great accuracy for market size and growth trajectories. Figure 25 shows the consistent performance across different market scales from Singapore's smaller user base to Thailand's larger market. Temporal analysis from 2011-2021 captures the acceleration of digital banking adoption with the model effectively predicting both gradual and rapid growth phases. The Philippines shows exemplary prediction alignment, while faster-growing markets like Vietnam and Thailand present interesting patterns that reflect their dynamic digital transformation.

VI. Findings and Actionable Insights

To see how well the RidgeCV model created performs on a different set of countries and training set, from The World Bank Group's original data, the rest of the countries listed were utilized.

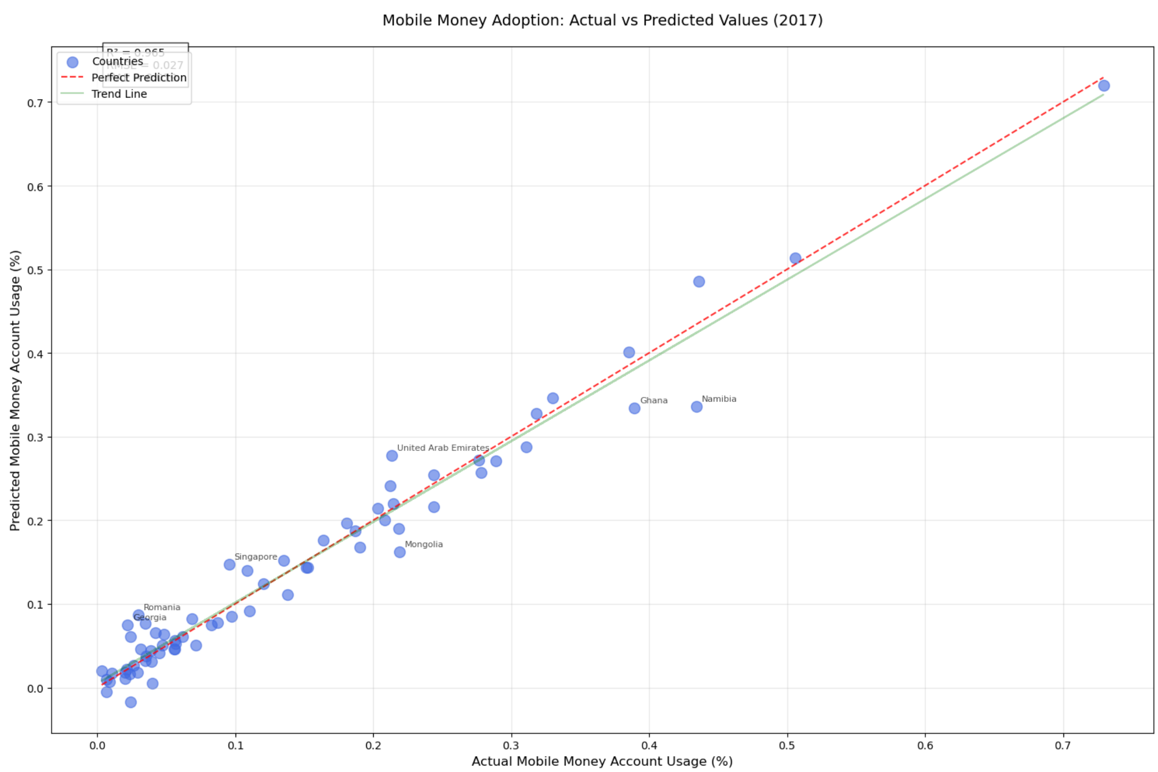
	Actual	Predicted	Difference
Mauritius	0.056269	0.056648	0.000379
Peru	0.026342	0.026821	0.000480
Chile	0.186708	0.187378	0.000669
Guatemala	0.021264	0.022169	0.000905
Honduras	0.061960	0.060464	0.001496
Tunisia	0.020372	0.018163	0.002209
Afghanistan	0.009138	0.006851	0.002287
Myanmar	0.006905	0.009404	0.002499
El Salvador	0.035496	0.038006	0.002510
Panama	0.035032	0.032329	0.002702

<Table 10: Mobile Money Markets Top 10 Most Precise Model Predictions>

After applying the model to the data, as can be seen in Table 10, the model demonstrates exceptional predictive power with an R-squared score of 0.965, indicating it explains 96.5% of the variance in mobile money adoption across countries. The low Root Mean Square Error (0.027) and Mean Absolute Error (0.019) further validate the model's high accuracy. In other words, the predictions are consistently close to the actual values.

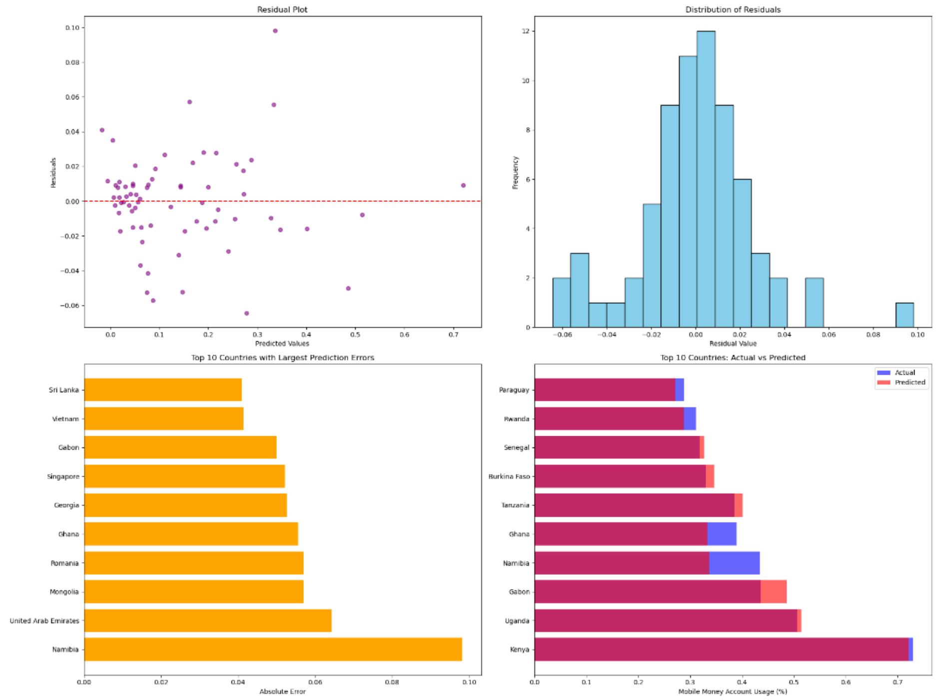
To put this into perspective, the errors represent only about two to three percentage points on average. Looking at individual country predictions, the model shows impressive accuracy. Mauritius leads with near-perfect prediction accuracy, showing only a 0.00038 difference between actual and predicted values. Peru and Chile follow with minimal differences of 0.00048 and 0.00067, respectively, which are differences of less than 0.1 percentage points. Even the tenth most accurate prediction, Panama, is only off by 0.27 percentage points.

The top ten most accurate predictions span diverse regions, from Latin America (Guatemala, Honduras, El Salvador, Panama) to Asia (Afghanistan, Myanmar) and Africa (Mauritius, Tunisia), demonstrating the model's robust performance across different market contexts and development stages. These results confirm the model's reliability for strategic planning and market analysis in the mobile money sector; the model most closely matches actual usage rates (Mauritius, Peru, Chile). This demonstrates the model's strength in capturing market dynamics across diverse economies, from Latin America to Southeast Asia and Africa.

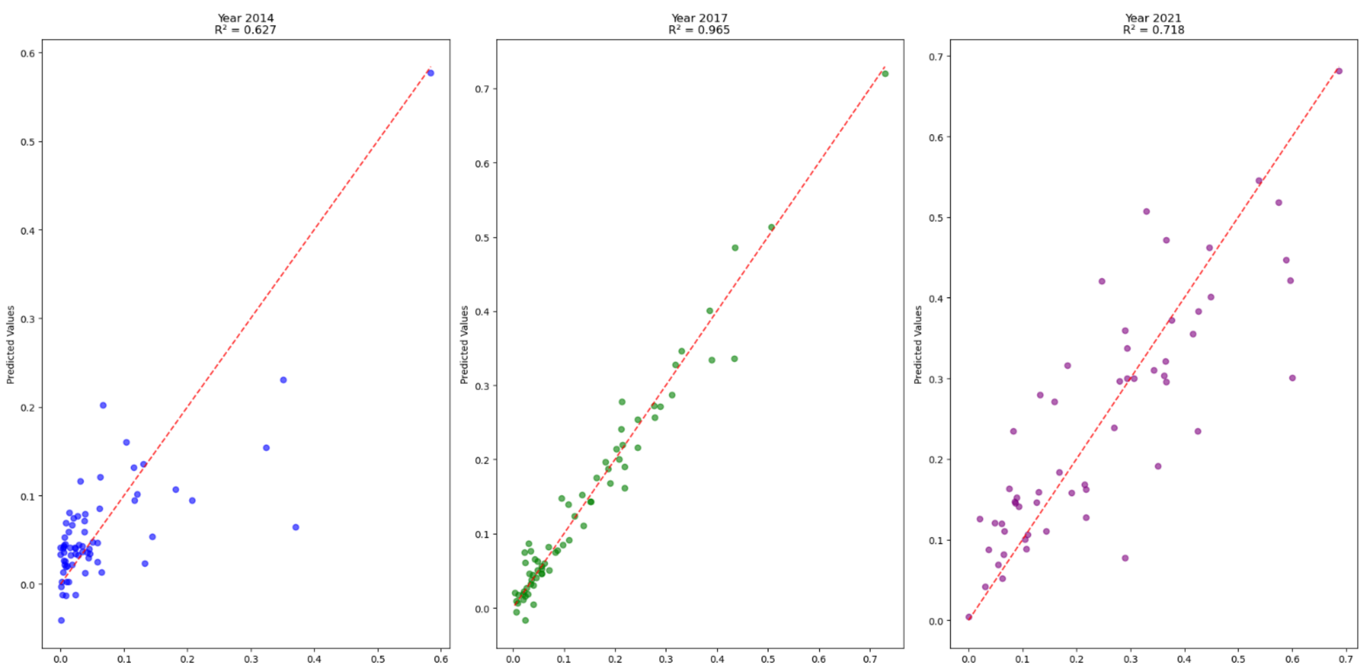


<Figure 26: Combined Countries Actual vs. Predicted Values (2017)>

The green trend line shown on the plot of Figure 26 is a statistical tool that shows the general direction and pattern of the data points. The line is calculated using linear regression to find the “best fit” line through all the points. The trend line helps to visualize the overall relationship between actual and predicted values, allowing one to check if the predictions tend to be systematically higher or lower than actual values. The trend line can also be compared to the red-dashed perfect prediction line to assess model bias. Considering the high R^2 value of 0.965, the trend line fits the data points very well and confirms the model's strong predictive power.



<Figure 27: Global-Scale Country Statistics Comparison>



<Figure 28: Years 2014, 2017, 2021 R^2 Comparison>

From Figure 28, 2017 shows the best performance across all metrics (R^2 = 0.965, RMSE = 0.027, MAE = 0.019). 2014 shows good predictive power (R^2 = 0.627) with reasonable error metrics, and 2021 demonstrates strong performance (R^2 = 0.718) despite having fewer samples (57). The varying performance across years suggests that the relationships between the features and mobile money adoption were most stable and predictable in 2017. Even with different sample sizes, the model is able to maintain good predictive power. In recent years (2021), might have new factors influencing adoption patterns.

VII. Further Research and Future Applications

The RidgeCV regression model for mobile banking adoption in Southeast Asia opens several promising avenues for future research; the strong performance across diverse economies suggests potential applications in other emerging markets, particularly in regions with similar digital transformation patterns.

Future research could expand the feature engineering approach to incorporate historical data, real-time transaction data, blockchain adoption metrics, and fintech integration indicators. The model's adaptability makes it valuable for financial institutions planning market entry strategies, regulators developing policy frameworks, and technology companies seeking market opportunity assessment. Additionally, the feature importance rankings provide a foundation for developing more sophisticated ensemble models that could capture non-linear relationships in mobile banking adoption. By integrating alternative data sources such as social media penetration, smartphone adoption rates, and digital literacy indicators, the model could evolve into a comprehensive tool for predicting and understanding digital financial service adoption across global markets.