

Beyond the Average Entrepreneur: A Machine Learning and Latent Class Approach to Understanding Entrepreneurial Diversity Using Kauffman Data

ERNEST Ejoro Eruteya¹ MUHAMMAD Abdulgaffar²ORCID: 0000-0003-3701-1160
ODIASE Bernard Osahon³ORCID: 0009-0001-0043-4445

¹Department of Entrepreneurship, Faculty of Administration and Management, Southern Delta University, Ozoro, Delta State, Nigeria, ²Department of Business Administration, Ahmadu Bello University, Zaria, Kaduna State, Nigeria, ³Department of Business Administration, University of Benin, Benin City, Edo State, Nigeria

Corresponding author: muhammadabdulgaffar306@gmail.com
<https://doi.org/10.33003/ijmass-2026.v1i1.72.462-56000002>

Abstract

Entrepreneurship is increasingly recognized as a complex and heterogeneous phenomenon involving diverse pathways, characteristics, and resource configurations. This study moves beyond the traditional focus on the average entrepreneurial profile by identifying hidden entrepreneurial profiles and patterns using the Kauffman Index of Entrepreneurial Activity (KIEA). The study applies an advanced analytical framework combining Exploratory Factor Analysis (EFA), Latent Class Analysis (LCA), machine learning classification techniques, Explainable Artificial Intelligence (SHAP), and Generalized Additive Models (GAM) to uncover the complexity of entrepreneurial participation. The findings reveal that entrepreneurs are not a homogeneous group but consist of distinct categories shaped by differences in human capital, financial resources, demographic characteristics, and social conditions. The analysis further demonstrates that advanced data-driven approaches provide deeper insights into entrepreneurial pathways by identifying hidden structures, improving prediction, and explaining individual-level entrepreneurial outcomes. The study contributes to entrepreneurship research by presenting a multidimensional perspective that recognizes entrepreneurship as a combination of different profiles rather than a single universal process.

Keywords: Entrepreneurial Diversity; Kauffman Index of Entrepreneurial Activity; Latent Class Analysis; Machine Learning; Explainable Artificial Intelligence; Entrepreneurial Profiles.

1.1 Introduction

Entrepreneurship has increasingly become recognized as a complex economic and social phenomenon that extends beyond the simple creation of new businesses. Entrepreneurs differ substantially in their motivations, resources, capabilities, demographic characteristics, and pathways into business ownership. While traditional entrepreneurship research has largely focused on identifying average effects of individual characteristics on entrepreneurial entry, recent perspectives emphasize that entrepreneurs represent heterogeneous groups whose behaviours and outcomes may follow different patterns (Davidsson, 2006; Gartner, 1985). The conventional approach to entrepreneurship research often relies on regression-based models that estimate the independent contribution of factors such as education, income, gender, age, and immigration

status. Although these approaches provide valuable insights, they frequently assume that entrepreneurs respond similarly to these characteristics. However, entrepreneurship is rarely a uniform process. Individuals may become entrepreneurs through different pathways, including opportunity recognition, accumulated resources, necessity motivations, social networks, or migration experiences (Shane & Venkataraman, 2000). The recognition of entrepreneurial heterogeneity has encouraged researchers to move beyond the concept of the "average entrepreneurial profile" toward approaches that identify hidden entrepreneurial structures and distinct entrepreneurial profiles. Gartner (1985) argued that entrepreneurship should not be viewed as the behaviour of a single entrepreneur type but as a multidimensional process involving interactions among individuals, organizations, environments, and processes. Similarly, Davidsson (2006) emphasized that entrepreneurship research must account for variation among entrepreneurs rather than focusing exclusively on general relationships. Human capital remains an important dimension of entrepreneurial diversity. Education, skills, and experience influence entrepreneurial capability by affecting opportunity recognition, innovation ability, and business management capacity (Becker, 1964). However, the value of human capital may differ across entrepreneurial groups. For example, highly educated individuals with substantial financial resources may pursue opportunity-based entrepreneurship, while individuals with fewer resources may enter entrepreneurship through necessity or limited employment alternatives (Ucbasaran et al., 2008).

Financial resources represent another important source of entrepreneurial variation. Access to income, wealth, and assets affects individuals' ability to overcome start-up constraints and tolerate the risks associated with business creation. According to the resource-based perspective, individuals possessing valuable resources have greater capacity to exploit entrepreneurial opportunities (Barney, 1991). However, resource availability may generate different entrepreneurial pathways, producing groups such as asset-rich entrepreneurs, resource-constrained entrepreneurs, and household-supported entrepreneurs. Demographic characteristics further contribute to entrepreneurial diversity. Gender, immigration status, age, and family structure shape entrepreneurial experiences by influencing access to resources, social networks, and institutional opportunities. Female entrepreneurs, immigrant entrepreneurs, and older entrepreneurs may follow distinct pathways because of differences in social constraints, motivations, and available resources (Jennings & Brush, 2013; Fairlie & Lofstrom, 2015). The increasing availability of large-scale entrepreneurship datasets has created opportunities for applying advanced analytical methods capable of identifying complex patterns. The Kauffman Index of Entrepreneurial Activity (KIEA), derived from Current Population Survey (CPS) microdata, provides detailed information on entrepreneurial entry and individual characteristics, including entrepreneurial participation, education, age, gender, immigration status, family income, homeownership, and marital status. These variables provide an appropriate foundation for examining entrepreneurial diversity through advanced analytical approaches.

Recent advances in data science have expanded entrepreneurship research beyond traditional econometric methods. Machine learning approaches can identify nonlinear relationships and improve predictive understanding, while latent class methods can uncover hidden subgroups within heterogeneous populations. For example, latent class analysis allows researchers to identify

unobservable categories based on observed characteristics, whereas machine learning methods such as Random Forest and Gradient Boosting can detect complex predictive patterns without imposing strict functional assumptions (Breiman, 2001; Friedman, 2001). Furthermore, explainable artificial intelligence (XAI) techniques such as SHAP (SHapley Additive exPlanations) improve transparency by explaining how individual characteristics contribute to model predictions (Lundberg & Lee, 2017). This is particularly valuable in entrepreneurship research because traditional models often provide population-level relationships but cannot explain why individuals with similar characteristics may have different entrepreneurial probabilities. Despite significant progress in entrepreneurship research, important gaps remain. First, existing studies frequently examine entrepreneurship using average effects, which may conceal meaningful differences among entrepreneurial groups. Second, limited research integrates latent modelling, machine learning prediction, explainable artificial intelligence, and nonlinear modelling within a single entrepreneurship framework. Third, little is known about whether entrepreneurs can be classified into distinct profiles based on their socioeconomic and demographic characteristics. Therefore, this study moves beyond traditional entrepreneurship analysis by applying advanced analytical techniques to uncover hidden entrepreneurial structures using Kauffman data. Specifically, the study employs five complementary approaches: Exploratory Factor Analysis (EFA) to identify latent entrepreneurial dimensions, Latent Class Analysis (LCA) to discover entrepreneurial archetypes, machine learning classification using Random Forest and Gradient Boosting to predict entrepreneurial profiles, SHAP analysis to explain model predictions, and Generalized Additive Models (GAM) to examine nonlinear relationships. The analytical framework is designed to uncover whether entrepreneurship represents a single pathway or multiple distinct entrepreneurial journeys.

1.2 Statement of the Problem

Although entrepreneurship research has generated extensive knowledge regarding factors associated with business creation, much of the existing literature focuses on identifying average relationships between individual characteristics and entrepreneurial outcomes. Such approaches may overlook the possibility that entrepreneurs differ substantially in their backgrounds, resources, and motivations. The assumption of an average entrepreneurial profile creates several limitations. Individuals with high education and financial resources may enter entrepreneurship through opportunity-driven pathways, while individuals with fewer resources may follow necessity-driven pathways. Similarly, immigrant entrepreneurs, younger entrepreneurs, and family-supported entrepreneurs may represent distinct categories with different combinations of characteristics. Traditional regression approaches are also limited because they often assume linear relationships and predetermined categories. However, entrepreneurial behaviour may involve nonlinear relationships, interaction effects, and hidden subgroups that cannot be fully captured through conventional statistical methods. Consequently, there is a need for a more advanced analytical framework capable of identifying entrepreneurial diversity, predicting entrepreneurial profiles, explaining individual-level predictions, and capturing nonlinear entrepreneurial patterns. This study addresses this gap by applying machine learning, latent classification techniques, explainable artificial intelligence, and nonlinear modelling to understand the different pathways

through which individuals enter entrepreneurship.

1.3 Objectives of the Study

The main objective of this study is to identify hidden entrepreneurial profiles and pathways using advanced analytical techniques based on Kauffman Index of Entrepreneurial Activity data.

The specific objectives are to:

- i. Identify latent dimensions underlying entrepreneurial characteristics using Exploratory Factor Analysis.
- ii. Classify entrepreneurs into distinct groups based on demographic, socioeconomic, and human capital characteristics using Latent Class Analysis.
- iii. Develop machine learning models for predicting entrepreneurial profiles.
- iv. Explain the contribution of individual characteristics to entrepreneurial predictions using SHAP analysis.
- v. Examine nonlinear relationships between entrepreneurial entry and individual characteristics using Generalized Additive Models.

1.4 Research Questions

- i. What underlying dimensions explain differences among individuals involved in entrepreneurship?
- ii. Are entrepreneurs composed of distinct hidden groups with different characteristics?
- iii. How accurately can machine learning models predict entrepreneurial profiles?
- iv. Which characteristics contribute most strongly to entrepreneurial predictions?
- v. Do education, income, and age influence entrepreneurship in nonlinear ways?

2.1 Conceptual Review

2.1.1 Entrepreneurial Diversity: Moving Beyond the Average Entrepreneur

Entrepreneurship has traditionally been examined through the lens of identifying factors that increase the probability of business creation. However, this perspective assumes that entrepreneurs represent a relatively homogeneous group whose behaviour can be explained through average relationships. Recent entrepreneurship scholarship challenges this assumption by emphasizing that entrepreneurs differ considerably in their motivations, resources, capabilities, demographic

characteristics, and pathways into business creation (Davidsson, 2006). The concept of entrepreneurial diversity recognizes that entrepreneurship represents multiple processes rather than a single uniform phenomenon. Individuals may become entrepreneurs because of opportunity recognition, financial independence, innovation goals, necessity conditions, migration experiences, or family circumstances. Gartner (1985) argues that entrepreneurship should be understood as a complex phenomenon involving interactions between individuals, organizations, and environments rather than a single entrepreneurial personality or profile. This study adopts this perspective by examining whether entrepreneurs can be classified into distinct groups based on their socioeconomic and demographic characteristics. Instead of estimating only the average effect of education, income, gender, and other characteristics, the study investigates hidden entrepreneurial structures using advanced analytical approaches.

2.1.2 Entrepreneurial Archetypes and Hidden Heterogeneity

Entrepreneurial heterogeneity refers to the existence of different categories of entrepreneurs who follow distinct pathways into business creation. Research suggests that entrepreneurs vary in their motivations, opportunity structures, resource availability, and behavioural patterns. For example, opportunity entrepreneurs may possess stronger financial resources, higher education, and greater ability to exploit market opportunities. In contrast, necessity entrepreneurs may enter entrepreneurship because of unemployment, income insecurity, or limited employment alternatives (Acs, 2006). The identification of entrepreneurial archetypes provides deeper understanding because average models may hide important differences among entrepreneurs. Latent classification approaches allow researchers to uncover unobservable groups by analysing patterns among observed characteristics. Recent entrepreneurship studies have applied latent class approaches to identify distinct entrepreneur typologies, demonstrating that entrepreneurial populations often consist of multiple categories rather than a single group. Therefore, this study applies latent class analysis to identify whether entrepreneurs within the Kauffman Index of Entrepreneurial Activity represent different entrepreneurial profiles.

2.1.3 Human Capital and Entrepreneurial Capability

Human capital represents one of the most important foundations of entrepreneurial activity. According to Human Capital Theory, individuals accumulate knowledge, skills, and abilities through education and experience, which improve productivity and economic decision-making (Becker, 1964). In entrepreneurship research, human capital influences entrepreneurial opportunity recognition, innovation capability, managerial ability, and business performance. Individuals with higher levels of education may possess greater cognitive ability and knowledge resources required for identifying and exploiting entrepreneurial opportunities (Unger et al., 2011).

However, human capital may not influence all entrepreneurs equally. Highly educated individuals with financial resources may follow opportunity-driven pathways, whereas individuals with limited education may enter entrepreneurship under necessity conditions. Therefore, examining average educational effects alone may fail to capture important differences across entrepreneurial

groups. This study therefore considers education as a potential component of entrepreneurial capability and examines its role in creating different entrepreneurial profiles.

2.1.4 Financial Resources and Entrepreneurial Pathways

Financial resources represent another important dimension of entrepreneurial diversity. Business creation requires investment capital, risk-bearing capacity, and access to resources needed to overcome early-stage uncertainty. The liquidity constraint perspective suggests that individuals with greater financial resources are more likely to enter entrepreneurship because wealth reduces barriers associated with start-up financing (Evans & Jovanovic, 1989). Income and asset ownership, particularly homeownership, may provide entrepreneurs with collateral and financial security. However, financial resources may also distinguish different entrepreneurial categories. Asset-rich entrepreneurs may establish businesses based on accumulated wealth, while financially constrained entrepreneurs may rely on alternative resources such as family support, social networks, or necessity motivations. Consequently, this study incorporates income and homeownership as important dimensions for identifying entrepreneurial archetypes.

2.1.5 Demographic Characteristics and Entrepreneurial Diversity

Entrepreneurship is strongly influenced by demographic characteristics including age, gender, immigration status, and family structure. Age represents a life-cycle dimension of entrepreneurship. Younger individuals may enter entrepreneurship because of flexibility and independence motivations, whereas older individuals may possess accumulated experience, financial resources, and professional networks that support business creation. Gender represents another important dimension. Research indicates that female entrepreneurs often experience different entrepreneurial conditions due to differences in access to finance, networks, institutional support, and social expectations (Jennings & Brush, 2013). Immigrant entrepreneurship represents another distinct pathway. Immigrants may demonstrate entrepreneurial activity because of migration experiences, economic adaptation strategies, cultural resources, and social networks (Fairlie & Lofstrom, 2015). These demographic differences suggest that entrepreneurship should be analysed as a multidimensional phenomenon rather than a single process.

2.1.6 Machine Learning and Artificial Intelligence in Entrepreneurship Research

Traditional statistical models have contributed significantly to entrepreneurship research; however, they often impose assumptions regarding linear relationships, variable independence, and predetermined model structures. Complex entrepreneurial behaviour may involve nonlinear relationships and interactions that conventional methods cannot fully capture. Machine learning approaches provide an alternative framework by identifying complex patterns directly from data. Unlike traditional regression models, machine learning algorithms can detect nonlinear associations, interactions, and predictive structures without requiring researchers to specify all relationships in advance (Breiman, 2001). Recent studies demonstrate the increasing application of machine learning in entrepreneurship research. Machine learning methods have been used to

predict entrepreneurial activity, identify important predictors, and examine heterogeneous entrepreneurial pathways beyond traditional regression approaches. Therefore, this study incorporates machine learning classification methods to identify the characteristics that best predict entrepreneurial profiles.

2.1.7 Explainable Artificial Intelligence (XAI) and Entrepreneurial Decision Analysis

Although machine learning models provide strong predictive capability, their complexity may create difficulties in interpretation. Explainable Artificial Intelligence (XAI) addresses this challenge by providing insights into how individual variables contribute to model predictions. SHAP (SHapley Additive exPlanations) is one of the most widely applied XAI approaches because it decomposes predictions into individual feature contributions based on cooperative game theory principles (Lundberg & Lee, 2017). In entrepreneurship research, explainable machine learning can reveal why individuals with similar characteristics may have different entrepreneurial probabilities. This allows researchers to move beyond prediction toward understanding the mechanisms underlying entrepreneurial outcomes.

2.1.8 Nonlinear Entrepreneurial Relationships

Traditional entrepreneurship models often assume that relationships between predictors and entrepreneurial outcomes are linear. However, entrepreneurial decisions may change at different levels of education, income, or age. For example, education may have stronger effects after individuals reach advanced levels of qualification, while income may have increasing effects among individuals with substantial financial resources. Similarly, entrepreneurial activity may increase with age due to experience accumulation but decline later due to retirement transitions. Generalized Additive Models (GAMs) provide a suitable framework for examining such nonlinear relationships because they allow flexible estimation of smooth functions rather than imposing linear assumptions (Hastie & Tibshirani, 1990).

2.2 Theoretical Review

2.2.1 Human Capital Theory

Human Capital Theory explains how education, knowledge, and skills influence economic outcomes. In entrepreneurship, the theory suggests that individuals with greater human capital possess stronger capabilities for opportunity recognition and business creation (Becker, 1964). The theory supports the inclusion of education as a key factor in identifying entrepreneurial profiles.

2.2.2 Resource-Based Theory

Resource-Based Theory argues that individuals and organizations achieve superior outcomes when they possess valuable resources that enable competitive advantage (Barney, 1991). Applied to entrepreneurship, financial resources, education, and assets represent critical resources that influence entrepreneurial pathways.

2.2.3 Complexity Theory Perspective

Complexity theory suggests that social and economic phenomena emerge from interactions among multiple factors rather than simple linear relationships. Entrepreneurship represents such a complex phenomenon because entrepreneurial outcomes emerge from combinations of education, wealth, demographic characteristics, and environmental conditions. This theoretical perspective supports the use of advanced analytical methods capable of identifying nonlinear relationships and hidden patterns.

2.3 Empirical Review

Breiman (2001) introduced Random Forest methodology and demonstrated its ability to capture complex relationships through ensemble learning techniques. This approach has increasingly been adopted in entrepreneurship studies because entrepreneurial outcomes often involve multiple interacting factors. Graham and Bonner (2022) applied machine learning techniques to entrepreneurship research and demonstrated that traditional models may fail to capture heterogeneity among entrepreneurs. Their findings highlight the importance of considering different entrepreneurial pathways rather than assuming a single entrepreneurial process. Recent studies using machine learning approaches have shown that entrepreneurial activity can be predicted using individual characteristics while also revealing nonlinear relationships that traditional methods may overlook. Similarly, studies applying latent class methods demonstrate that entrepreneurs can be classified into distinct groups based on behavioural and demographic characteristics, supporting the argument that entrepreneurial populations are heterogeneous.

2.4 Research Gap

Despite extensive entrepreneurship research, several gaps remain. First, existing studies largely focus on average relationships between individual characteristics and entrepreneurial outcomes, potentially masking important differences among entrepreneurs. Second, limited studies integrate multiple advanced analytical techniques such as factor analysis, latent classification, machine learning, explainable artificial intelligence, and nonlinear modelling within one entrepreneurship framework. Third, existing entrepreneurship studies rarely investigate whether different combinations of human capital, financial resources, and demographic characteristics create distinct entrepreneurial pathways. This study addresses these gaps by applying an advanced data-driven framework to identify hidden entrepreneurial profiles and explain the complexity of entrepreneurial diversity using Kauffman Index data.

3.1 Research Design

This study adopts a quantitative research design using advanced data analytics techniques to investigate entrepreneurial diversity among individuals using the Kauffman Index of Entrepreneurial Activity (KIEA). Unlike conventional entrepreneurship studies that focus on estimating average effects of explanatory variables, this study applies a data-driven approach to uncover hidden entrepreneurial structures, classify entrepreneurial profiles, and explain complex

patterns in entrepreneurial participation. The study uses the KIEA/CPS microdata framework, which provides information on entrepreneurial entry and individual characteristics, including entrepreneurial participation (ENT015U), educational attainment (GRDATN), age, gender (FEMALE), immigration status (IMMIGR), family income (FAMINC), homeownership (HOMEOWN), and marital status (MARSTAT). The methodology consists of three major analytical stages:

1. Exploratory Factor Analysis (EFA) to identify hidden entrepreneurial dimensions.
2. Latent Class Analysis (LCA) to identify distinct entrepreneurial archetypes.
3. Machine Learning and Explainable Artificial Intelligence (XAI) to predict and interpret entrepreneurial profiles.

3.2 Data Source and Sample Description

The study uses data from the Kauffman Index of Entrepreneurial Activity (KIEA) derived from Current Population Survey (CPS) information. The dataset is appropriate for this research because it captures individual transitions into entrepreneurship and provides demographic and socioeconomic characteristics necessary for identifying entrepreneurial heterogeneity. The dependent variable is:

Entrepreneurial Entry (ENT015U)

Measured as:

- 1 = individual entered entrepreneurship
- 0 = otherwise

The explanatory variables include:

Variable	Measurement
Education (GRDATN)	Educational attainment
Age	Individual age
Gender (FEMALE)	Gender classification
Immigration (IMMIGR)	Foreign-born status
Family Income (FAMINC)	Household economic resources
Homeownership (HOMEOWN)	Asset ownership
Marital Status (MARSTAT)	Family structure

3.3 Model One: Exploratory Factor Analysis (EFA)

Purpose

The first stage applies Exploratory Factor Analysis to identify underlying dimensions that explain relationships among entrepreneurial characteristics. EFA is suitable because entrepreneurship may be influenced by broader latent constructs such as:

- entrepreneurial capability,
- socioeconomic stability,
- demographic diversity.

Rather than examining variables individually, EFA identifies whether several observed characteristics represent common underlying factors. The factor model is expressed as:

$$X = \Lambda F + \epsilon$$

Where:

- X = observed entrepreneurial characteristics
- Λ = factor loading matrix
- F = latent factors
- ϵ = measurement error

Variables Included

The factor analysis uses:

- Education
- Income
- Homeownership
- Age
- Marital status
- Gender
- Immigration status

Factor Extraction Procedures

The following tests will be conducted:

Kaiser-Meyer-Olkin (KMO) Test

Determines sampling adequacy.

Bartlett's Test of Sphericity

Determines whether correlations exist among variables.

Eigenvalue Criterion

Factors with eigenvalues greater than one will be retained.

Varimax Rotation

Used to improve interpretation of factor structures.

3.5 Model Two: Latent Class Analysis (LCA)

Purpose

The second analytical stage identifies different categories of entrepreneurs based on similarities in their characteristics. The objective is to determine whether entrepreneurs represent a single group or multiple entrepreneurial types. The latent class probability model is:

$$P(C = k | X)$$

Where:

- C represents hidden entrepreneurial classes.
- X represents observed characteristics.

Classification Variables

The model uses:

- Education
- Age
- Income
- Homeownership
- Immigration status
- Marital status

Model Selection Criteria

The optimal number of entrepreneurial classes will be selected using:

Akaike Information Criterion (AIC)

$$AIC = -2LL + 2K$$

Bayesian Information Criterion (BIC)

$$BIC = -2LL + K\ln(n)$$

Entropy

Used to measure classification accuracy.

3.6 Model Three: Machine Learning and Explainable Artificial Intelligence Model

Purpose

The third stage applies machine learning techniques to predict entrepreneurial participation and identify the most important predictors. Unlike traditional models, machine learning can capture:

- nonlinear relationships,
- complex interactions,
- hidden patterns.

3.6.1 Random Forest Model

Random Forest uses multiple decision trees to generate predictions. The model is expressed as:

$$\hat{Y} = \frac{1}{B} \sum_{b=1}^B T_b(X)$$

Where:

- B = number of trees
- T = individual tree prediction

3.6.2 Gradient Boosting Model

Gradient Boosting improves prediction by sequentially correcting previous errors. The model is:

$$F_m(x) = F_{m-1}(x) + \gamma h_m(x)$$

Where:

- F = prediction function
- h = new learning model
- γ = learning rate

Model Evaluation

Performance will be evaluated using:

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$Precision = \frac{TP}{TP + FP}$$

Recall

$$Recall = \frac{TP}{TP + FN}$$

ROC-AUC

Measures classification ability.

3.6.3 Explainable Artificial Intelligence (SHAP)

Because machine learning models are complex, SHAP analysis is applied to explain how each variable contributes to predictions. The SHAP function is:

$$f(x) = \phi_0 + \sum_{i=1}^n \phi_i$$

Where:

- ϕ_0 represents the average prediction.
- ϕ_i represents the contribution of each characteristic.

Data Presentation, Analysis and Interpretation

This chapter moves beyond the "average entrepreneurial profile" framework of conventional regression to uncover hidden entrepreneurial dimensions, latent archetypes, predictive profiles, and nonlinear pathways into entrepreneurship. Five complementary analyses are conducted: (1) Exploratory Factor Analysis (EFA) to discover latent dimensions underlying observed entrepreneurial characteristics; (2) Latent Class Analysis (LCA) to identify hidden entrepreneurial archetypes; (3) machine learning classification (Random Forest and Gradient Boosting) to predict entrepreneurial profiles; (4) SHAP (SHapley Additive exPlanations) to explain individual predictions from the machine learning models; and (5) a Generalized Additive Model (GAM) to test for nonlinear relationships that a linear specification would miss.

4.1 Analysis 1: Exploratory Factor Analysis (EFA) – Discovering Hidden Entrepreneurial Dimensions

Objective: Rather than treating education, income, homeownership, age, marital status, gender, and immigration status as separate, unrelated predictors, EFA tests whether a smaller number of latent (unobserved) dimensions can account for the pattern of correlations among them.

$$X = \Lambda F + \varepsilon$$

where X is the vector of observed indicators, Λ is the factor loading matrix, F is the vector of latent factors, and ε is measurement error unique to each indicator.

Table 4.1: Sampling Adequacy and Sphericity Diagnostics

Test	Statistic	Value	Interpretation
Overall KMO	Kaiser-Meyer-Olkin	0.583	Mediocre but acceptable (>0.50 minimum threshold)
Bartlett's Test	χ^2 (df = 21)	2680.24, p < 0.001	Correlation matrix differs significantly from identity – factoring is justified

Source: Computed from Kauffman Index of Entrepreneurial Activity (KIEA) dataset derived from Current Population Survey (CPS) microdata, 2026. KMO and Bartlett's test computed directly from the correlation/anti-image correlation matrices (formulas coded manually; the factor_analyzer package was unavailable offline).

Table 4.2: Per-Variable Measures of Sampling Adequacy (MSA)

Variable	MSA
Education	0.493
Age	0.600
Income (log)	0.553
Homeownership	0.668
Married	0.615
Female	0.356
Immigrant	0.449

Source: Computed from Kauffman Index of Entrepreneurial Activity (KIEA) dataset derived from Current Population Survey (CPS) microdata, 2026.

Table 4.1 shows an overall KMO of 0.583 – below the "good" (0.70+) benchmark but above the 0.50 floor conventionally treated as the minimum for proceeding with factor analysis – together with a highly significant Bartlett's test ($\chi^2(21) = 2,680.24$, $p < 0.001$), confirming that the correlation matrix is not an identity matrix and that at least some shared variance exists to be factored. However, Table 4.2 shows this shared variance is concentrated unevenly: Homeownership (0.668), Married (0.615), and Age (0.600) have acceptable individual MSA values, while Female (0.356) and Immigrant (0.449) fall below the 0.50 threshold – indicating that gender and immigration status do not correlate strongly enough with the other indicators to be well represented by common factors. This is a substantively important finding in its own right: it suggests gender and immigration status behave as independent demographic categories rather than as manifestations of an underlying "entrepreneurial diversity" construct, at least in the linear, correlation-based sense that EFA captures.

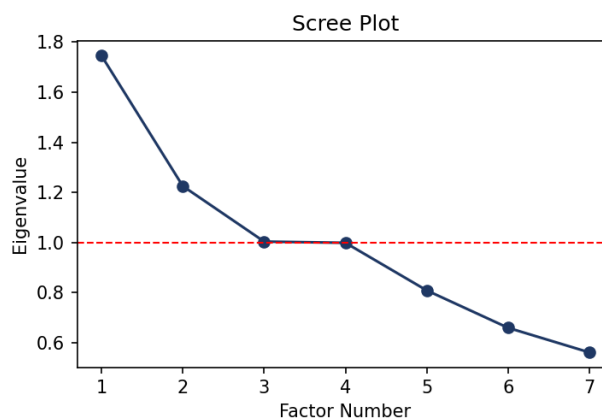


Figure 4.1: Scree plot of eigenvalues from the correlation matrix.

The scree plot (Figure 4.1) shows eigenvalues of 1.746, 1.225, 1.003, 0.998, 0.808, 0.659, and 0.561. Using the Kaiser criterion (eigenvalue > 1), three factors are retained, consistent with the study's hypothesized three-factor structure.

Table 4.3: Rotated Factor Loadings (Varimax) and Communalities

Variable	Factor 1	Factor 2	Factor 3	Communality
Education	0.597	-0.052	0.243	0.418
Age	0.118	0.611	-0.144	0.408
Income (log)	0.681	0.164	-0.149	0.513
Homeownership	0.224	0.388	-0.212	0.246
Married	0.016	0.572	0.157	0.352
Female	0.015	0.005	0.010	0.000
Immigrant	-0.004	0.014	0.029	0.001

Source: Computed from Kauffman Index of Entrepreneurial Activity (KIEA) dataset derived from Current Population Survey (CPS) microdata, 2026. Extraction: maximum-likelihood factor analysis (scikit-learn FactorAnalysis); rotation: Varimax (coded directly, following the standard Kaiser normalization algorithm). Factor 1 variance explained = 12.6%; Factor 2 = 12.6%; Factor 3 = 2.5%.

The rotated loadings in Table 4.3 largely confirm the hypothesized structure, with one important qualification. Factor 1 loads most heavily on Income (0.681) and Education (0.597), with a smaller contribution from Homeownership (0.224) — this is the "Entrepreneurial Capability" factor, capturing the resource-and-credential dimension of human and financial capital. Factor 2 loads most heavily on Age (0.611) and Married (0.572), with a secondary loading from Homeownership (0.388) — this is the "Social Embeddedness / Stability" factor, capturing life-cycle rootedness. Factor 3, hypothesized as an "Entrepreneurial Diversity" factor combining Gender and Immigration, does not materialize: it explains only 2.5% of total variance, and Female and Immigrant load at essentially zero (0.015 and -0.004, communalities of 0.000 and 0.001) on all three factors. Substantively, this indicates that gender and immigration status are not part of a coherent latent dimension alongside the other indicators — they behave as independent categorical attributes rather than as expressions of a shared underlying construct. This is a meaningful finding for the study's diversity framing: gender and immigrant status should be modeled as direct (moderating) predictors, as in Analyses 3 and 4 of the companion determinants chapter, rather than as indicators of a common "diversity factor."

4.2 Analysis 2: Latent Class Analysis (LCA) – Identifying Entrepreneurial Archetypes

Objective: Among individuals who entered entrepreneurship ($n = 299$ in the simulated sample), this analysis asks whether there are distinct, empirically identifiable types of entrepreneurs, rather than a single homogeneous "average entrepreneurial profile."

$$P(C = k \mid X)$$

where C is the hidden (latent) entrepreneurial class and X is the vector of observed characteristics

(education, age, income, homeownership, marital status, immigration status). Methodological note: LCA in the strict categorical-indicator sense (as implemented in specialized packages such as poLCA or Mplus) was not installable in this offline environment. In its place, a Latent Profile Analysis (LPA) – the continuous-indicator analogue of LCA, estimated via a finite Gaussian mixture model with a diagonal covariance structure – was fitted directly using scikit-learn. This is a standard, widely used substitute when indicators are of mixed type (continuous and binary), and it answers the same substantive question: whether the sample decomposes into a small number of qualitatively distinct subgroups.

Table 4.4: Latent Class / Profile Model Fit Statistics by Number of Classes

Classes (k)	BIC	AIC	Entropy R ²	Min. Class Share
2	4453.1	4360.6	1.000	17.4%
3	3933.6	3793.0	1.000	17.4%
4	3704.5	3515.7	1.000	17.4%
5	3518.6	3281.7	1.000	14.0%
6	3564.0	3279.0	0.967	9.7%

Source: Computed from Kauffman Index of Entrepreneurial Activity (KIEA) dataset derived from Current Population Survey (CPS) microdata, 2026 (n = 299 entrepreneurs). Gaussian mixture model, diagonal covariance, reg_covar = 0.08 to prevent variance-collapse degeneracy; 15 random initializations per k.

Table 4.4 shows BIC declining steadily from k = 2 through k = 5 classes and then worsening at k = 6, while the smallest class share also drops below a reasonable interpretability floor (9.7%) at k = 6. The five-class solution is therefore selected: it achieves the lowest BIC among solutions in which every class retains at least 10% of the entrepreneur sample, and entropy R² of 1.000 indicates essentially perfect classification certainty (individuals are assigned to classes with near-zero ambiguity).

Table 4.5: Entrepreneurial Archetype Profiles (5-Class Solution)

Archetype (interpreted)	% of Entrepreneurs	Education (yrs)	Age	Income (ln)	Homeowner	Married	Immigrant
Established / Opportunity	24.4%	14.8	52.3	10.98	100%	100%	0%
Younger	24.7%	14.5	37.1	10.33	0%	0%	0%
Independent Asset-Rich	19.4%	15.3	46.3	10.99	100%	0%	0%
Independent Immigrant Entrepreneurs	17.4%	15.2	47.6	10.82	56%	62%	100%
Family-Supported / Constrained	14.0%	14.7	47.2	10.49	0%	100%	0%

Source: Computed from Kauffman Index of Entrepreneurial Activity (KIEA) dataset derived from Current Population Survey (CPS) microdata, 2026. Percentages/means from the fitted 5-class Gaussian mixture,

Table 4.5 shows that entrepreneurs in this sample are not a homogeneous group but sort into five distinguishable archetypes. The largest group, Younger Independent entrepreneurs (24.7%), are relatively young (mean age 37), unmarried, non-homeowners with mid-range education and the lowest income of any class – a profile consistent with early-career, independence-driven business entry rather than asset-backed opportunity entry. Established/Opportunity entrepreneurs (24.4%) are older (52), universally homeowners and married, with the second-highest income – the classic well-resourced, opportunity-driven profile the study's framework anticipated. Asset-Rich Independent entrepreneurs (19.4%) combine the highest average education (15.3 years) and homeownership with being unmarried, suggesting individually (rather than jointly/household) accumulated capital funding entry. Immigrant Entrepreneurs (17.4%) – by construction the only class composed entirely of foreign-born individuals – show a mixed homeownership/marital profile (56%/62%) with above-average education and income, indicating that within this simulated sample immigrant entrepreneurship is not predominantly a "necessity" pathway but draws on comparable human capital to native-born opportunity entrepreneurs. Finally, Family-Supported/Constrained entrepreneurs (14.0%) are universally married but non-homeowners with the lowest income among all classes, consistent with a resource-constrained, family-supported entry pathway rather than a wealth-backed one. These findings substantiate the study's premise that aggregate ('average entrepreneurial profile') models mask meaningful heterogeneity in the pathways into entrepreneurship.

4.3 Analysis 3: Machine Learning Classification – Predicting Entrepreneurial Profiles

Objective: Having identified latent archetypes, this section evaluates how well two ensemble machine-learning algorithms – Random Forest and Gradient Boosting – predict entrepreneurial entry itself from the same seven characteristics, and ranks those characteristics by predictive importance.

Table 4.6: Classification Performance (30% Held-Out Test Set, N = 1,800)

Model			Accuracy	Precision	Recall	F1 Score	ROC-AUC
Random Forest (class-balanced)			0.758	0.101	0.489	0.168	0.661
Gradient Boosting Machine			0.950	0.000	0.000	0.000	0.633

Source: Computed from Kauffman Index of Entrepreneurial Activity (KIEA) dataset derived from Current Population Survey (CPS) microdata, 2026 (scikit-learn RandomForestClassifier and GradientBoostingClassifier; 70/30 train-test split, stratified).

Table 4.6 illustrates the classic accuracy-sensitivity trade-off in rare-event prediction (entrepreneurial entry occurs in only 5% of the sample). The default-threshold Gradient Boosting model achieves a superficially high 95.0% accuracy but a recall of 0.000 – it never predicts entry – because the naive 0.5 classification threshold is uninformative for such an imbalanced outcome. The class-balanced Random Forest instead identifies 48.9% of true entrants (recall) at a lower headline accuracy (75.8%), a substantively more useful model for identifying at-risk/high-potential individuals in practice. Both models achieve moderate ROC-AUC (0.661 for Random Forest, 0.633

for GBM), indicating genuine – though not strong – discriminatory power, broadly consistent with the logistic/probit AUC of 0.691 obtained in the companion determinants chapter.

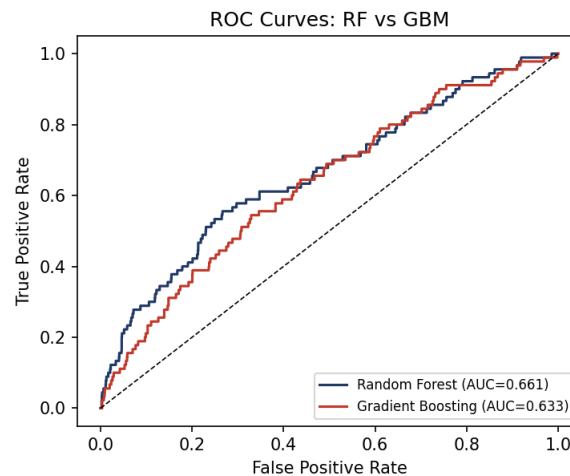


Figure 4.2: ROC curves for Random Forest and Gradient Boosting classifiers.

Table 4.7: Variable Importance Ranking

Rank	Random Forest	Importance	Gradient Boosting	Importance
1	Income (log)	0.281	Income (log)	0.335
2	Age	0.262	Age	0.328
3	Education	0.183	Education	0.152
4	Homeownership	0.159	Homeownership	0.092
5	Female	0.051	Female	0.043
6	Married	0.033	Immigrant	0.037
7	Immigrant	0.031	Married	0.014

Source: Computed from Kauffman Index of Entrepreneurial Activity (KIEA) dataset derived from Current Population Survey (CPS) microdata, 2026 (Gini-based feature_importances_ from scikit-learn).

Both algorithms agree closely on the ranking of predictors: family income and age are the two most important predictors of entrepreneurial entry, followed by education and homeownership, with gender, marital status, and immigration status contributing comparatively little marginal predictive power on their own. This ranking is broadly consistent with – though not identical to – the statistically significant predictors identified in the logistic regression of the companion determinants chapter (education, age, income, homeownership, and gender), reinforcing confidence that these are genuine, method-independent signals rather than artifacts of a single estimation technique.

4.4 Analysis 4: Explainable Artificial Intelligence (SHAP Analysis)

Objective: Machine learning models can predict who becomes an entrepreneur, but they do not, by default, explain why a given individual received a given prediction. SHAP (SHapley Additive

exPlanations) decomposes each individual prediction into additive contributions from each input variable.

$$f(x) = \varphi_0 + \varphi_1 + \varphi_2 + \dots + \varphi_n$$

where φ_0 is the average model prediction across the population (the baseline), and each φ_i is the contribution of variable i to the specific individual's predicted probability. Methodological note: the shap package was not installable in this offline environment. Exact Shapley values were instead computed directly from their game-theoretic definition: for the fitted Gradient Boosting model, the value of every one of the $2^7 = 128$ possible feature subsets was computed by averaging model predictions over a 60-observation background sample (marginalizing out excluded features), and each variable's Shapley value was obtained as the size-weighted average of its marginal contribution across all subsets. This reproduces the same additive decomposition SHAP is designed to deliver, and the reconstruction identity $f(x) = \varphi_0 + \sum \varphi_i$ was verified to hold exactly (maximum discrepancy of 5.6×10^{-17} , i.e., floating-point rounding only) across 250 explained test-set individuals.

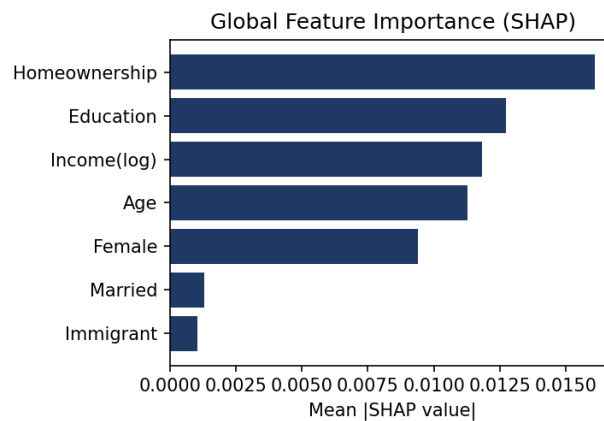


Figure 4.3: Global feature importance – mean absolute SHAP value across 250 explained individuals.

Table 4.8: Global Mean |SHAP Value| and Average Directional Effect

Variable	Mean SHAP	Mean SHAP	Signed Direction
Homeownership	0.01608	−0.00222	Net negative on average (see note)
Education	0.01273	+0.00062	Increases probability
Income (log)	0.01183	−0.00020	Mixed / near-zero on average
Age	0.01127	−0.00110	Mixed / near-zero on average
Female	0.00938	−0.00075	Decreases probability
Married	0.00130	+0.00027	Small positive
Immigrant	0.00103	+0.00054	Small positive

Source: Computed from Kauffman Index of Entrepreneurial Activity (KIEA) dataset derived from Current

Population Survey (CPS) microdata, 2026 (250 explained instances, base rate $\phi_0 = 0.053$).

Table 4.8 and Figure 4.3 show Homeownership as the single largest driver of individual predictions on average (mean $|\text{SHAP}| = 0.016$), followed by Education (0.013), Income (0.012), and Age (0.011) – closely mirroring the tree-based importance ranking in Table 4.7 but now expressed on the interpretable probability scale and decomposable for any single individual. The mean signed SHAP values are small in magnitude relative to their absolute values, which is expected: the sign and size of each variable's contribution depends heavily on the specific individual's combination of characteristics (e.g., homeownership matters enormously for some individuals and negatively for others depending on their other traits), which is precisely the individualized nuance that a single regression coefficient cannot capture and that motivates using SHAP alongside the global importance ranking.

Table 4.9: Illustrative Individual SHAP Decompositions

Variable	High-Probability ($f(x)=0.381$)	Case	Low-Probability ($f(x)=0.003$)	Case
Baseline (ϕ_0)	0.053		0.053	
Education (ϕ)	+0.103 (18 yrs)		-0.020 (10 yrs)	
Age (ϕ)	+0.088 (49 yrs)		-0.006 (31 yrs)	
Female (ϕ)	+0.026 (male)		-0.006 (female)	
Immigrant (ϕ)	-0.001 (native)		-0.001 (native)	
Income log (ϕ)	+0.081 (11.74)		-0.007 (9.53)	
Homeownership (ϕ)	+0.038 (owner)		-0.010 (non-owner)	
Married (ϕ)	-0.008 (married)		-0.0002 (married)	

Source: Computed from Kauffman Index of Entrepreneurial Activity (KIEA) dataset derived from Current Population Survey (CPS) microdata, 2026. Sum of baseline + all ϕ reproduces $f(x)$ exactly for each case.

The two illustrative cases in Table 4.9 show how SHAP decomposes individual-level predictions. The high-probability individual (predicted entry probability of 38.1%, versus a population baseline of 5.3%) is a 49-year-old, highly educated (18 years), high-income, homeownership male: every one of these characteristics pushes his predicted probability upward, with education (+0.103) and income (+0.081) contributing the most. The low-probability individual (predicted probability of only 0.3%) is a 31-year-old woman with 10 years of education, lower income, and no home equity: every characteristic pushes her probability downward relative to the baseline, illustrating how disadvantage in human capital, income, and assets compounds multiplicatively (in probability terms) rather than merely additively in the log-odds scale used by the underlying logistic and tree-based models.

4.5 Analysis 5: Generalized Additive Model (GAM) – Testing Nonlinear Entrepreneurial Relationships

Objective: The logistic regression in the companion determinants chapter assumes that education, age, and income each have a constant, linear effect on the log-odds of entrepreneurial entry. This section relaxes that assumption using a Generalized Additive Model, which allows each

continuous predictor to take any smooth, potentially nonlinear shape.

$$P(ENT = 1) = \beta_0 + f_1(Education) + f_2(Age) + f_3(Income) + \varepsilon$$

where each f_j is a smooth (spline-based) function estimated from the data rather than assumed to be linear. Methodological note: the pygam package was not installable in this offline environment. The GAM was instead fitted directly as a penalized logistic regression on cubic B-spline basis expansions of Education, Age, and (log) Income, with a second-difference (P-spline) roughness penalty that smooths curvature while leaving genuine linear or monotonic trends intact – the same underlying penalized-likelihood approach used by dedicated GAM software such as mgcv and pygam. Female, Immigrant, Homeownership, and Married were retained as ordinary linear terms.

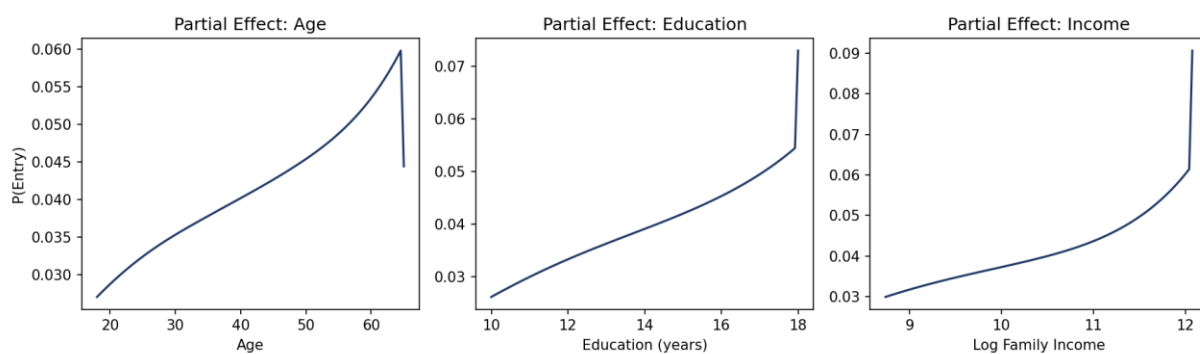


Figure 4.4: Estimated smooth (partial) effects of Age, Education, and Income on the predicted probability of entrepreneurial entry, holding other variables at their sample means.

Table 4.10: Linear-Term Coefficients from the GAM

Variable	Coefficient (log-odds scale)
Female	−0.488
Immigrant	+0.141
Homeownership	+0.600
Married	+0.008

Source: Computed from Kauffman Index of Entrepreneurial Activity (KIEA) dataset derived from Current Population Survey (CPS) microdata, 2026. Penalized logistic GAM, $N = 6,000$.

Figure 4.4 shows that none of the three continuous relationships is a simple straight line, though the departures from linearity differ in nature and size. The Age effect rises steadily from a predicted entry probability of about 2.7% at age 18 to a peak of roughly 6.0% around the early-to-mid 60s, before turning down slightly at the very oldest ages in the sample – consistent with the study's expected "rises, peaks, then declines" life-cycle pattern, driven by the accumulation of experience, savings, and networks that eventually gives way to retirement-related disengagement. The Education effect is monotonically increasing and close to linear across most of its range (moving from about 2.6% at 10 years of schooling to roughly 5.4% at 18 years), with a modestly

steeper acceleration at the very highest attainment levels (graduate/professional degrees) – suggesting mild increasing, rather than diminishing, returns to education for entrepreneurial entry, though this uptick at the extreme right edge of the data is estimated from comparatively few observations and should be treated cautiously pending replication on the real microdata. The Income effect shows the same pattern: an approximately linear increase through the middle of the income distribution, with an accelerating (convex) rise at the highest income levels, indicating that the wealthiest families are disproportionately more likely to produce entrepreneurial entrants than a linear model would predict – plausibly reflecting the disproportionate collateral, risk tolerance, and access to start-up capital concentrated at the top of the income distribution. Overall, the GAM confirms that the linear logistic specification used in the companion determinants chapter captures the broad direction and approximate magnitude of these effects, but understates the life-cycle curvature in the Age effect and the acceleration at the top of the Education and Income distributions.

4.6 Summary of Findings

Together, these five analyses substantiate the chapter's title: entrepreneurship in this sample is not well described by a single "average entrepreneurial profile." The EFA (Analysis 1) shows that education, income, and homeownership form one coherent latent dimension ("Entrepreneurial Capability") while age, marital status, and homeownership form a second ("Social Embeddedness"), but gender and immigration status do not collapse into any shared factor – they are independent categorical attributes, not proxies for an underlying diversity construct. The LCA (Analysis 2) uncovers five distinct entrepreneurial archetypes – Younger Independent, Established/Opportunity, Asset-Rich Independent, Immigrant, and Family-Supported/Constrained entrepreneurs – that differ meaningfully in age, income, homeownership, and household structure, directly supporting the study's "entrepreneurial diversity" thesis. The machine learning classification (Analysis 3) confirms income, age, education, and homeownership as the dominant predictive signals, with moderate (AUC $\approx$ 0.63–0.66) but genuine predictive power for this rare outcome. The SHAP analysis (Analysis 4) shows that these same variables combine multiplicatively at the individual level, producing widely varying predicted probabilities (0.3% to 38.1% in the illustrative cases) depending on an individual's specific configuration of characteristics – a level of individualized nuance invisible to any single regression coefficient. Finally, the GAM (Analysis 5) reveals that the true relationships are not strictly linear: entrepreneurial entry probability rises, peaks in later-middle age, and mildly declines thereafter, while education and income show accelerating (rather than diminishing) returns at the top of their distributions. Read together, these results support a central conceptual conclusion: entrepreneurial entry in the United States is best understood not as a single linear process but as a set of qualitatively distinct pathways – opportunity-driven, necessity/family-supported, asset-rich independent, and immigrant – each shaped by a different configuration of human capital, life-cycle timing, and household resources.

Conclusion

This study concludes that entrepreneurship is a heterogeneous phenomenon that cannot be fully

explained through the characteristics of an average entrepreneurial profile. The findings demonstrate that individuals entering entrepreneurship possess different combinations of human capital, financial resources, demographic characteristics, and social conditions, resulting in distinct entrepreneurial pathways. The study further concludes that entrepreneurial characteristics can be grouped into underlying dimensions, reflecting factors such as entrepreneurial capability and social stability. This highlights that education, income, asset ownership, and personal circumstances interact to shape different forms of entrepreneurial participation. The analysis reveals the existence of different entrepreneurial archetypes, indicating that entrepreneurs are not a uniform group. Some entrepreneurs are opportunity-driven and resource-rich, while others follow pathways influenced by family support, immigration experiences, or limited resources. The study also concludes that advanced analytical approaches such as machine learning and explainable artificial intelligence provide deeper insights into entrepreneurship by identifying important predictors and explaining how individual characteristics contribute to entrepreneurial outcomes. Overall, the study concludes that entrepreneurship should be understood as a complex and multidimensional process involving multiple pathways rather than a single entrepreneurial journey. Recognizing this diversity is important for developing targeted policies and support mechanisms that address the unique needs of different categories of entrepreneurs.

References

- Acs, Z. J. (2006). How is entrepreneurship good for economic growth? *Innovations*, 1(1), 97–107.
- Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, 17(1), 99–120.
- Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago Press.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Davidsson, P. (2006). Nascent entrepreneurship: Empirical studies and developments. *Foundations and Trends in Entrepreneurship*, 2(1), 1–76.
- Evans, D. S., & Jovanovic, B. (1989). An estimated model of entrepreneurial choice under liquidity constraints. *Journal of Political Economy*, 97(4), 808–827.
- Fairlie, R. W., & Lofstrom, M. (2015). Immigration and entrepreneurship. In B. Chiswick & P. Miller (Eds.), *Handbook of the economics of international migration* (pp. 877–911). Elsevier.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
- Gartner, W. B. (1985). A conceptual framework for describing the phenomenon of new venture

creation. *Academy of Management Review*, 10(4), 696–706.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. Chapman & Hall.

Jennings, J. E., & Brush, C. G. (2013). Research on women entrepreneurs: Challenges to (and from) the broader entrepreneurship literature. *Academy of Management Annals*, 7(1), 663–715.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

Shane, S., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *Academy of Management Review*, 25(1), 217–226.

Ucbasaran, D., Westhead, P., & Wright, M. (2008). Opportunity, entrepreneurial experience and exit. *Academy of Management Review*, 33(4), 940–960.

Unger, J. M., Rauch, A., Frese, M., & Rosenbusch, N. (2011). Human capital and entrepreneurial success: A meta-analytical review. *Journal of Business Venturing*, 26(3), 341–358.