UNIVERSIDAD COMPLUTENSE MADRID

ntic master

Founder Personality Investment Score (FPIS): Design and Validation of an Interpretable Predictive Model for Estimating Startup Success Based on Founder Personality and Team Composition.

Master's in Data Science, Big Data & Business Analytics 2024–2025

Table of Contents

1. Abstract

This Master's Thesis designs and validates the Founder Personality Investment Score (FPIS), a predictive and interpretable indicator that estimates the probability of a startup's success (exit via acquisition or IPO) based exclusively on the aggregated Big Five personality traits of the founding team and team composition/diversity metrics. The approach aims to be actionable in pre-seed stages, avoiding contextual variables (country, sector, funding), so that FPIS can be applied in early due diligence when financial information is scarce.

The research uses the public dataset of McCarthy et al. (2023). After cleaning and preparation, the working notebook contains *n* = 21,160 organizations with a non-null binary success label and a block of founder-level Big Five traits aggregated to the team level (means, maxima, dispersion), along with diversity/complementarity indices and the number of founders. We denote by XPX_PXP the vector of predictors of personality and team composition. FPIS is formalized as a calibrated probability:

FPIS(xP)=P(Y=1|XP=xP)=g(fP(xP))

$$\mathrm{FPIS}(x_P) \;=\; \mathbb{P}(Y{=}1 \,|\, X_P{=}x_P) \;=\; g\big(f_P(x_P)\big)$$

where fP(·) is the predictive model trained solely with XPX_PXP, and g(·) is a calibration function (Platt or isotonic) selected by Brier/ECE.

Methodologically, an EDA is conducted to understand the distribution of *success*, the relationship between the number of founders and outcomes, the variability of team-level Big Five traits, and diversity among co-founders. In modeling, Logistic Regression, Random Forest, XGBoost, and LightGBM are compared with stratified validation, hyperparameter optimization, and probability calibration (Platt/Isotonic). Evaluation focuses on ROC-AUC, PR-AUC, Brier score, gains/lift curves, and Net Benefit (decision curve analysis, DCA).

The results show that the winning model is calibrated XGBoost (best method by Brier/ECE), trained exclusively with XPX_PXP. In the notebook, the model achieves ROC-AUC = 0.676 and PR-AUC = 0.287, with Brier = 0.220; compared to the baseline reference (ROC-AUC = 0.646; PR-AUC = 0.249), this represents a +2.78 p.p. improvement in AUC and a noticeable gain in PR-AUC. In prioritization terms, the top quintile by FPIS achieves Lift Q5/Q1 ≈ 5.12× (calibrated model), and at the coverage level, the Top-20% by FPIS reaches a success rate of ≈ 0.289 versus ≈ 0.162 overall, evidencing strong concentration power of exits in the upper score range. In decision analysis, the model exhibits positive Net Benefit at relevant decision thresholds, and optimal thresholds are reported under different cost–benefit scenarios.

From a business perspective, FPIS enables prioritization of deal flow in pre-seed, enhances investment committees with a calibrated probability, and interprets score drivers via SHAP, providing traceability and explainability to decisions. Overall, the empirical evidence in the notebook supports that psychological characteristics and founding team composition contain actionable predictive information on startup success, and that a scoring system based solely on XPX_PXP can improve opportunity selection without relying on contextual signals unavailable at early stages.

2. Introduction

2.1 Business Context and Motivation

Early-stage startup investment is characterized by high levels of uncertainty and very limited access to financial or market information. In the pre-seed stage in particular, investors must make critical decisions without consolidated metrics of traction, revenues, or clients. In this context, the founding team becomes the main evaluable asset, since the literature has demonstrated that their individual and collective characteristics decisively influence the company's future performance (Rauch & Frese, 2007).

In recent years, founder personality has emerged as a key dimension of analysis. The study by McCarthy et al. (2023), which constitutes the empirical basis of this thesis, showed that certain Big Five traits and their complementarity within the team are correlated with the probability of achieving an exit (acquisition or IPO). However, this finding has not yet translated into an actionable operational indicator that can be directly used in due diligence processes.

The business problem is therefore clear: venture capital funds and business angels need predictive, interpretable, and early-stage applicable tools that allow them to prioritize deal flow before financial data becomes available. In this scenario, the FPIS is proposed as a score designed to estimate the probability of startup success based on the collective psychological footprint of the founding team, calculated solely from publicly available information.

This approach provides three main motivations:

1. **Early Actionability**: FPIS is built with data derived from founders' digital footprints (e.g., Twitter), making it usable even when the startup has not yet generated business metrics.
2. **Informational Advantage**: It introduces a differential dimension compared to investors' traditional filters (industry, country, number of founders) by capturing the team's psychological diversity and complementarity.
3. **Interpretability and Traceability**: Since it is based on Big Five percentile aggregates and diversity metrics already present in the dataset, it allows for understanding which traits drive the score and facilitates justifying investment decisions to committees.

Therefore, this thesis lies at the intersection of psychology, data science, and venture capital, with the ambition of turning an academic finding into a practical tool for investors.

2.2 Literature Review

In recent years, the Big Five paradigm—openness, conscientiousness, extraversion, agreeableness, and emotional stability—has become the reference for modeling personality in organizational and entrepreneurial contexts. Its application in venture capital is especially promising, as these traits can be digitally inferred from social media footprints and aggregated at the team level to capture collective dynamics.

The study by McCarthy et al. (2023) represented a turning point in this line of research. Using a massive dataset of founders and their startups, the authors empirically demonstrated that:

- The psychological composition of founding teams contains predictive information on startup success.

- Intra-team diversity and psychological complementarity explain significant differences in the probability of exit.
- Predictive models can be built at scale using only information derived from founder profiles, without the need for initial financial metrics.

However, the study itself acknowledged a limitation: the lack of an interpretable score ready to be applied in real investment processes. While correlations and models were explored, no calibrated and validated tool was offered that venture capitalists or business angels could directly adopt.

In parallel, other research in **team diversity** (Harrison & Klein, 2007) has highlighted that heterogeneity in individual characteristics—including age, education, gender, or psychological traits—can create competitive advantages when it translates into complementarity of perspectives and leadership styles. However, they also warn that excessive heterogeneity can increase coordination costs and conflict, requiring diversity to be quantified in a balanced way.

Overall, the literature points toward a consensus:

1. Founder personality and team composition are decisive in the fate of startups.
2. Psychological metrics can be quantified and modeled predictively.
3. There is still a gap in translating these findings into an actionable, calibrated, and explainable KPI that enables investors to prioritize opportunities at early stages.

This is precisely the gap that the FPIS developed in this work seeks to fill.

2.3 Objectives of the Thesis

The main objective of this Master's Thesis is to design and validate the FPIS, a predictive and interpretable indicator that estimates the probability of a startup's success—defined as achieving an exit via acquisition or IPO—based on the aggregated personality and composition of the founding team.

Unlike previous approaches, FPIS is built exclusively on the information available in the public dataset of McCarthy et al. (2023):

- Founders' Big Five personality traits.
- Statistical team-level aggregations (maxima, means, variances, IQR).
- Psychological diversity metrics (Blau index, FOALED, type shares).
- **Complementarity Index** (average distance from each founder to the team centroid).
- Team structure (number of founders).

Contextual variables such as country or sector are explicitly excluded from the main training process to ensure that the score is applicable in pre-seed phases, when such information is not yet solid or does not represent a reliable signal of success.

From a business perspective, this objective translates into three differentiated contributions:

1. **Actionability before investing**

2. **Informational advantage**
3. **Interpretability**

2.4 Research Hypotheses

FPIS validation is structured around three main hypotheses:

- **H1. Added Value**
  The incorporation of personality and team composition variables (XPX_PXP) significantly improves predictive capacity compared to a baseline model built solely with contextual controls (sector, country, and number of founders). An improvement of AUC ≥ 5 percentage points is expected in predicting success.
- **H2. Segmentation**
  Startups in the top FPIS quintile will have at least double the success rate of those in the bottom quintile. This hypothesis assesses the practical utility of the score as a tool for prioritizing deal flow.
- **H3. Practical Decision-Making**
  Using a calibrated FPIS threshold aligned with the cost–benefit ratio of false positives and false negatives will generate an expected return ≥ 15% higher than that of a random investment strategy in the same sector.

These hypotheses combine a methodological focus (H1, comparing model metrics), a strategic segmentation focus (H2, identifying exit concentration in the top of the score), and a business applicability focus (H3, simulating returns and net benefit). Together, they provide a solid framework to evaluate whether FPIS truly constitutes a differential and actionable KPI in early-stage investment processes.

3. Methodology

3.1 Dataset and Data Sources

The empirical basis of this thesis comes from the public dataset published by McCarthy et al. (2023) as part of their research *The Impact of Founder Personalities on Startup Success*. This dataset, available in academic repositories and GitHub, contains information on more than 20,000 startups and their founding teams, along with psychological characteristics inferred from digital footprints.

Specifically, the dataset includes:

- Organization identifiers (*org_id*) and basic contextual variables such as country (*org_country*) and number of founders (*org_numfounders*).
- The binary outcome variable *success*, which takes value 1 if the startup achieved an exit (acquisition or IPO) and 0 otherwise.
- Founders' Big Five personality traits, inferred through linguistic analysis of Twitter posts. These traits include both the five main factors (openness, conscientiousness, extraversion, agreeableness, and emotional stability) and their disaggregated facets.

- Pre-computed team-level aggregations in the original dataset, such as percentile maxima (*big5_max_*), means, variances, or dispersion indices.
- Psychological diversity metrics (e.g., Blau index, FOALED) and the **Complementarity Index**, which measures the average distance of each member from the team's psychological centroid.

In the working notebook, the file **AdditionalData.csv** was loaded, containing 21,160 organizations and 41 original variables. After cleaning and verification, a *team_df* was built with 49 columns integrating Big Five variables, aggregations, diversity metrics, and the target variable *success*.

The dataset is characterized by two key aspects:

1. **Class imbalance**: Only around 16.22% of startups achieved an exit, introducing a significant classification challenge (Figure 1).
2. **Global coverage**: While the United States accounts for nearly half of the observations (≈10,200 startups), the dataset includes organizations from 13 countries, capturing variability in team composition and success rates (Figure 2).

The use of this dataset offers several advantages:

- **Academic relevance**, as it is the original source of the paper underpinning this thesis.
- **Practical actionability**, since it is based on public and replicable information derived from founders' digital footprints.
- **Methodological rigor**, with a sufficient sample size and variables specifically designed to model personality and team diversity.

In this thesis, the dataset is used exclusively to build and validate FPIS, ensuring that the input vector XPX_PXP contains only aggregated psychological traits and team composition, in line with the stated objectives.

3.2 Formal Definition of FPIS

FPIS is defined as a probabilistic and interpretable indicator of the probability of startup success, based on the psychological and compositional characteristics of its founding team.

Let XPX_PXP be the predictor vector composed of:

- Aggregated Big Five personality traits at the team level (maxima, means, variances, IQR).
- Big Five facets in percentiles (e.g., *Openness_facet_adventurousness_percentile*).
- Diversity metrics (Blau index, FOALED, type shares).
- **Complementarity Index**, which measures the average distance of each member from the team's psychological centroid.
- Number of founders (*org_numfounders*).

On this predictor vector, a predictive model fP(·)f_P(\cdot)fP(·) is trained, producing uncalibrated probabilities. Subsequently, a calibration function g(·)g(\cdot)g(·) (Platt sigmoid or isotonic regression) is applied to obtain an adjusted and interpretable probability. Formally, FPIS is defined as:

FPIS(xP)=p^(Y=1|XP=xP)=g(fP(xP))

$$\mathrm{FPIS}(x_P) \;=\; \mathbb{P}(Y{=}1 \,|\, X_P{=}x_P) \;=\; g\big(f_P(x_P)\big)$$

where:

- Y∈{0,1} represents the startup's success variable (*exit*).
- fP(xP) is the raw score generated by the model trained solely with XPX_PXP.
- g(·) is the calibration function that adjusts this score into a well-calibrated probability.

In practice, this score returns a continuous value between 0 and 1, interpretable as the expected probability of success for the startup based on its collective psychological profile.

The design of FPIS follows three fundamental principles:

1. **Early Actionability.** The score is computed with variables available from the outset (founders' digital footprints), making it applicable in pre-seed stages.
2. **Interpretability.** Since it is based on easily understandable percentile aggregates and diversity metrics, investors can identify which characteristics drive a high or low FPIS value.
3. **Methodological Consistency.** FPIS is grounded in a rigorously trained model with cross-validation, hyperparameter optimization, and probabilistic calibration, ensuring that the generated probability is reliable and stable.

Thus, FPIS is configured as a calibrated probability of success conditioned exclusively on the personality and composition of the founding team.

3.3 Data Preparation and Feature Engineering

Once the original dataset (*AdditionalData.csv*) was loaded, several data preparation and feature engineering processes were carried out in order to construct the vector XPX_PXP—that is, the representation of the founding team based solely on personality traits and team composition.

1. **Normalization and Column Verification**

- The organization identifier was renamed to *org_id* for consistency.
- The target variable *success* was standardized as binary (0/1).
- Key columns such as *org_numfounders* and *org_country* were ensured, with default values generated for missing cases.

2. **Aggregation of Personality Traits**
   From the individual percentiles of Big Five facets (e.g.,

Openness_facet_adventurousness_percentile), team-level aggregations were constructed following three approaches:

- **Central tendency measures**: mean and median of each facet.
- **Dispersion measures**: variance and interquartile range (IQR).
- **Extreme measures**: maximum and minimum for each trait.

This provided a synthetic representation that captures both the average levels of each trait within the team and the degree of internal heterogeneity.

3. **Construction of Diversity Metrics**
   To operationalize the hypothesis that psychological diversity enriches the team but may also generate coordination costs, specific indices were computed:

- **Blau Index (FOALED):** measures the heterogeneity of categorical distributions of traits, adapted here to psychological facets.
- **Shares by type:** proportion of founders at each extreme of personality dimensions.

These metrics make it possible to evaluate whether a team is composed of homogeneous or complementary profiles.

4. Complementarity Index

A psychological complementarity metric was implemented, calculated as the average distance between each founder and the team centroid in the Big Five personality space.

CI=n1i=1∑nd(xi,x¯)

$$\mathrm{CI} = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \bar{x})$$

where:

- n is the number of founders in the team.
- xi represents the personality trait vector of founder i.
- $\bar{x}$ is the team's average profile (centroid in the Big Five space).
- d(·,·) corresponds to the cosine distance between personality vectors.

5. Structural Variables
The number of founders (*org_numfounders*) was included, since the literature suggests that both very small and excessively large teams show a lower probability of success (Rauch & Frese, 2007).

6. Handling Missing Values
Although the dataset provided high coverage, some Big Five facets contained missing values. These were managed through:

- Mean imputation when the percentage of missing values was low.

- Exclusion of columns with insufficient coverage.

7. Final Outcome
After this process, an enriched *team_df* was constructed with 49 variables, including:

- Aggregated Big Five (maxima, means, variances, IQR).
- Blau and FOALED diversity indices.
- Complementarity Index.
- Number of founders.
- Target variable *success* (0/1).

This feature engineering provided a complete, interpretable, and consistent XPX_PXP vector aligned with the thesis objectives, ensuring that FPIS is based exclusively on the personality and structure of the founding team.

3.4 Exploratory Data Analysis (EDA)

Before training the predictive models, an extensive exploratory analysis was carried out to understand variable distributions, detect possible biases, and validate dataset quality.

3.4.1 Class Imbalance

The first analysis focused on the distribution of the target variable *success*. Of the 21,160 startups included, only 16.22% achieved an exit, compared to 83.78% that did not. This marked imbalance poses a classification challenge, as models tend to favor the majority class (Figure 1).

3.4.2 Number of Founders

The distribution of *org_numfounders* is centered on 2, and most startups have between 1 and 3 founders. The success analysis by team size shows that exit probabilities tend to be higher in teams with 2 to 3 founders, while both solo founders and teams with more than 5 members display lower success rates. This supports the hypothesis that moderate team size balances coordination and diversity (Figure 3).

3.4.3 Geographic Distribution

Although the main goal of this thesis excludes contextual variables from the predictive model, it is relevant to characterize the database. The country with the highest representation is the United States (≈10,200 startups, ~48%), followed by Germany and the United Kingdom. In terms of success rate, significant differences exist: countries such as Israel show rates above 20% (when $n \geq 100$ startups), while others like India present rates closer to 8%. These differences confirm the dataset's contextual heterogeneity (Figure 4).

3.4.4 Aggregated Big Five Traits

Aggregated Big Five personality traits reveal interesting patterns:

- Teams with high openness tend to concentrate in startups with a greater probability of success.
- Conversely, extreme scores in neuroticism (emotional range) are associated with lower success rates.

- The distribution by class (*success* vs. *no success*) shows visible, though not deterministic, differences in some traits (Figure 5).

3.4.5 Psychological Diversity and Complementarity
The analysis of diversity and complementarity indices confirms the relevance of these metrics:

- The Blau index (FOALED) shows a wide distribution, with higher values in more diverse teams. Boxplots suggest a moderately positive relationship between diversity and success.

3.4.6 Correlations with the Target Variable
A correlation matrix was built between key numerical variables and *success*. Although bivariate correlations are generally low—which is expected in complex prediction problems—some associations stand out in facets of openness, conscientiousness, and extraversion, as well as in diversity indices. These signals, while weak individually, can be exploited by more sophisticated multivariate models (Figure 6).

3.4.7 Big Five Facets
Finally, a subset of specific Big Five facets was explored using violin plots. The analysis shows that successful startups tend to have founders with higher percentiles in facets such as *adventurousness* or *achievement striving*, reinforcing the importance of modeling personality at a granular level (Figure 7).

In summary, the EDA confirms that the dataset contains psychological and structural signals with predictive power, although non-linear and non-trivial, which justifies the use of advanced and calibrated supervised models to build FPIS.

4. Predictive Modeling

4.1 Compared Models

To estimate the probability of startup success from the vector XPX_PXP, several supervised classification algorithms were implemented and compared. The model selection was based on balancing predictive capacity, interpretability, and common use in business analytics environments:

1. **Logistic Regression (LogReg)**

- Linear reference model, widely used in binary prediction contexts.
- Provides directly interpretable probabilities.
- Serves as a baseline benchmark against more complex models.

2. **Random Forest (RF)**

- Ensemble algorithm based on decision trees trained on bootstrap subsets.
- Reduces variance and improves robustness against overfitting.
- Particularly useful for capturing non-linear interactions among personality traits.

3. **XGBoost (XGB)**

- Optimized gradient boosting algorithm, recognized for its high performance.

- Iteratively adjusts residual errors and allows precise handling of class imbalance through the *scale_pos_weight* parameter (≈ 5.16 in our dataset).
- Achieved the best results after calibration, becoming the foundation of FPIS.

4. **LightGBM (LGBM)**

- Boosting variant developed by Microsoft, optimized for speed and memory efficiency.
- Uses histograms to accelerate training.
- Included as an alternative to XGBoost, given its strong performance on large tabular datasets.

All models were implemented in Python (*scikit-learn*, *XGBoost*, *LightGBM*) and integrated into pipelines including scaling, training, and probability calibration. The comparison process was designed to address **Hypothesis H1 (added value versus baseline)** and to determine whether the calibrated XGBoost-based FPIS provided statistically and practically relevant improvements.

4.2 Validation Strategy

The validation of the models was carefully designed to ensure that the results were robust, comparable, and generalizable to the startups in the dataset.

4.2.1 Stratified Cross-Validation

Given the class imbalance in the target variable (*success* ≈ 16.22%), a stratified cross-validation strategy (*StratifiedKFold*) was used. This approach ensures that each training and validation split maintains the same proportion of positive and negative cases as in the full dataset, avoiding metric biases.

- A standard *k = 5* folds was applied for binary classification problems.
- In each iteration, the model was trained on 80% of the data and evaluated on the remaining 20%.
- Reported results correspond to the average metrics across folds, reducing variance due to specific partitions.

4.2.2 Evaluation Metrics

To evaluate model quality and validate the hypotheses, a set of complementary metrics was used, allowing analysis of different dimensions of predictive performance:

1. **ROC-AUC (Receiver Operating Characteristic – Area Under Curve).**
    - Measures the model's overall ability to discriminate between positive (success) and negative (failure) classes.
    - The baseline (control-only model) reached AUC = 0.646, while FPIS (calibrated XGBoost) achieved 0.676, a +2.78 p.p. improvement.
2. **PR-AUC (Precision-Recall – Area Under Curve).**
    - Especially relevant in imbalanced class contexts.

- The baseline reached 0.251, compared to 0.287 for FPIS, representing a +3.6 p.p. improvement in recovering true successes.

3. **Brier Score.**
    - Evaluates the quality of predicted probabilities, penalizing deviations from actual outcomes.
    - FPIS (calibrated XGB) obtained Brier = 0.220.
4. **Gains and Lift Curves.**
    - Assess the model's ability to concentrate the largest proportion of successes in the top percentiles of the score.
    - In the notebook, the top FPIS quintile concentrated up to 5.1 times more successes than the bottom quintile.
5. **Net Benefit (Decision Curve Analysis).**
    - Evaluates the practical utility of a model under different risk thresholds.
    - FPIS showed positive net benefits in decision ranges relevant for investment, validating its business applicability.

4.2.3 Probabilistic Calibration

A central aspect of the validation strategy was ensuring that the resulting score was probabilistically interpretable. To achieve this:

- Two calibration methods (Platt/sigmoid and isotonic) were compared, and the one minimizing Brier/ECE in validation was chosen.
- In the final evaluation of the winning XGBoost, isotonic calibration was selected, offering the best combination of Brier and ECE in the test set.

Together, this strategy allowed not only the relative performance comparison of the models but also ensured that FPIS was a calibrated, reliable, and interpretable score, meeting the methodological objectives of this thesis.

4.3 Hyperparameter Optimization

To maximize model performance and avoid suboptimal configurations, a two-phase hyperparameter optimization process was implemented:

4.3.1 Preliminary Search with RandomizedSearchCV

In the first stage, *scikit-learn*'s *RandomizedSearchCV* was used to efficiently explore the hyperparameter spaces of each algorithm:

- **Logistic Regression:** regularization coefficient CCC, penalty type (L1/L2).
- **Random Forest:** number of trees (*n_estimators*), maximum depth (*max_depth*), proportion of variables per split (*max_features*).
- **XGBoost:** learning rate, maximum depth, number of iterations (*n_estimators*), *subsampling*, and *colsample_bytree*.

- **LightGBM:** learning rate, number of leaves (*num_leaves*), feature fraction, and bagging fraction.

This search helped identify promising parameter ranges, reducing the risk of overfitting and discarding inefficient combinations.

4.3.2 Fine-Tuning with Optuna

In the second phase, Optuna with the Tree-structured Parzen Estimator (TPE) algorithm was applied, allowing more sophisticated optimization through sequential learning of probability distributions.

- Specific objective functions were defined for each model, using ROC-AUC as the main optimization metric.
- Optuna was run with adaptive pruning, discarding unpromising configurations early to save computational time.
- For XGBoost, the configuration selected by Optuna balanced bias and variance, delivering the best overall performance in cross-validation.

4.3.3 Optimization Results

The optimization showed that:

- In LogReg, the best performance came with L2 regularization and low CCC values (more penalized models).
- In RF, intermediate depths (≤ 10) and a high number of trees (> 300) maximized performance without overfitting.
- In XGBoost and LightGBM, low learning rates combined with a high number of iterations captured complex patterns without degrading calibration.

Overall, the optimization confirmed that boosting models (XGB and LGBM) consistently outperformed the others, with calibrated XGBoost delivering the best results, justifying its selection as the foundation of FPIS.

4.4 Probability Calibration

A fundamental requirement of FPIS is that the generated values are well-calibrated probabilities, meaning the score magnitude accurately reflects a startup's true probability of success.

In binary classification with imbalanced data, machine learning models tend to produce poorly calibrated probabilities: they often overestimate the probability of success for positives and underestimate it for negatives. This limits their practical utility in business contexts, where investors need probabilities that can be directly interpreted and compared with decision thresholds.

4.4.1 Applied Calibration Techniques

Two classic calibration methods were evaluated:

- **Platt scaling (sigmoid).**
  Fits a logistic regression over the raw model probabilities, transforming them via a

sigmoid function. It is simple and robust, especially effective when the number of positive cases is limited.

- **Isotonic calibration.**
  Uses a non-parametric stepwise function that adjusts probabilities more flexibly, adapting to the empirical distribution of the target variable. While it can provide more precise calibration, it requires larger datasets to avoid overfitting.

Both procedures were implemented via *scikit-learn*'s *CalibratedClassifierCV*, applied to boosting models (XGB and LGBM) and, to a lesser extent, RF and LogReg.

4.4.2 Calibration Results

Notebook results show that:

- XGBoost calibrated with isotonic offered the best calibration curve, reducing the deviation between predicted probabilities and observed frequencies.
- Isotonic calibration, while improving predictions in some folds, tended to overfit in smaller subsamples, showing less stability than Platt.
- In global metrics, calibrated XGB achieved:
  - ROC-AUC = 0.676
  - PR-AUC = 0.287
  - Brier score = 0.220

These results consolidated it as the optimal model for constructing FPIS.

4.4.3 Business Interpretation

Calibration makes FPIS interpretable as the *actual probability of success*, distinguishing it from models that only generate relative rankings. This enables direct applications in venture capital, such as:

- Setting investment thresholds (e.g., invest only if FPIS > 0.25).
- Comparing startups from different sectors on a common risk scale.
- Simulating expected returns by adjusting success probabilities with return factors and error costs.

Probabilistic calibration ensured that FPIS is not only predictive but also actionable and interpretable.

5. Results

5.1 Performance of Baseline Models vs. FPIS

The first comparison was conducted between baseline models (with basic contextual variables) and models enriched with XPX_PXP, the vector containing only personality traits and team composition. The goal of this analysis was to test whether FPIS added predictive value, in line with Hypothesis H1.

5.1.1 Baseline Models

The baseline included only sector, country, and the number of founders (*org_numfounders*) as predictors. Although these variables capture some structural differences, their discriminative power is limited:

- ROC-AUC = 0.646
- PR-AUC = 0.249

These results reflect modest performance and poor calibration, insufficient for practical use in investment prioritization.

5.1.2 Models with XPX_PXP (Founder Personality Investment Score)

By incorporating XPX_PXP, which contains aggregated personality traits, diversity indices, and complementarity metrics, performance consistently improved:

- Calibrated XGBoost (isotonic) reached ROC-AUC = 0.676, PR-AUC = 0.287, and Brier score = 0.220.
- This represents a +2.78 p.p. gain in ROC-AUC and a +3.6 p.p. gain in PR-AUC compared to the baseline.
- The lower Brier score confirms better calibration of probabilities.

5.1.3 Overall Comparison

The comparative analysis between baseline and FPIS can be visualized in ROC and PR curves, where the FPIS model clearly dominates the baseline across all threshold regions (Figure 8 and Figure 9).

Overall, these results validate that psychological and team composition information adds differential predictive value, surpassing the predictive ability of contextual variables. This evidence is consistent with Hypothesis H1 and demonstrates that FPIS is a score with greater practical utility in pre-seed stages.

5.2 Detailed Metrics Comparison

Beyond the baseline vs. FPIS comparison, the performance of the four algorithms (Logistic Regression, Random Forest, XGBoost, and LightGBM) trained exclusively with XPX_PXP was evaluated. Complementary metrics were analyzed to capture different dimensions of predictive performance.

5.2.1 ROC-AUC

The ROC-AUC curve measures the overall ability to discriminate between successful and unsuccessful startups. Results were:

- Logistic Regression (calibrated): AUC = 0.675
- Random Forest: AUC ≈ 0.662
- LightGBM: AUC ≈ 0.670
- XGBoost (Platt calibrated): AUC = 0.676

Although absolute differences are small, XGBoost achieved the best performance, confirming its robustness in tabular data with non-linear relationships.

5.2.2 PR-AUC

The precision-recall curve is more informative in imbalanced contexts like this one (exit rate = 16.22%). Results were:

- Logistic Regression: PR-AUC = 0.294
- Random Forest: PR-AUC ≈ 0.271
- LightGBM: PR-AUC ≈ 0.281
- XGBoost (calibrated): PR-AUC = 0.287

Logistic Regression achieved a slightly higher PR-AUC, reflecting its good fit in linear scenarios, though calibrated XGBoost offered a more balanced performance overall.

5.2.3 Brier Score

The Brier score, which measures the accuracy of predicted probabilities, showed a clear advantage for calibrated models:

- Calibrated XGBoost: 0.220
- Calibrated LightGBM: ≈ 0.224
- Calibrated Logistic Regression: ≈ 0.228
- Random Forest: ≈ 0.233

FPIS based on XGBoost not only improved discrimination but also provided the most reliable probabilities for decision-making.

5.2.4 Lift and Success Concentration

Finally, gains and lift curves were analyzed, measuring the models' ability to concentrate successes in the top score percentiles:

- The top quintile (Q5) of FPIS concentrated ≈ 5.1× more successes than the bottom quintile (Q1).
- The top 20% of startups by FPIS achieved a success rate of ≈ 28.9% vs. 16.2% overall, confirming the score's utility as a prioritization tool.

Together, these metrics demonstrate that:

1. All models based on XPX_PXP clearly outperform the baseline.
2. Calibrated XGBoost offers the best balance between discrimination, calibration, and lift, consolidating its role as the winning model and the foundation of FPIS.
3. FPIS fulfills the hypothesis that personality and composition variables provide measurable added value in predicting success.

5.3 Hypothesis Validation

5.3.1 Hypothesis H1 — Added Value

- **Premise.** Incorporating personality and team composition variables (XPX_PXP) should increase predictive power compared to a baseline with only sector, country, and number of founders. An improvement of AUC ≥ 5 p.p. was expected.
- **Results.**
    - Baseline: ROC-AUC = 0.646; PR-AUC = 0.249.
    - FPIS (calibrated XGBoost): ROC-AUC = 0.676; PR-AUC = 0.287; Brier = 0.220.
    - Relative improvements: +2.78 p.p. ROC-AUC; +3.6 p.p. PR-AUC; -0.017 in Brier.
- **Interpretation.** Although the AUC gain was below the expected 5 p.p., the results demonstrate consistent added value from including the psychological dimension. FPIS improves calibration and success detection, partially confirming H1.

5.3.2 Hypothesis H2 — Segmentation

- **Premise.** The top FPIS quintile should show at least double the success rate of the bottom quintile.
- **Results.**
    - Top quintile (Q5): success rate ≈ 0.2893.
    - Bottom quintile (Q1): success rate ≈ 0.0565.
    - Q5/Q1 ratio: ≈ 5.12×.
    - The top 20% by FPIS achieved ≈ 0.289 vs. ≈ 0.162 overall, with Lift Q5/Q1 ≈ 5.12×.
- **Interpretation.** FPIS concentrates successes in the upper distribution far more than expected. While the hypothesis anticipated ≥ 2×, results showed a lift of 5.1×, strongly validating H2 (Figure 10 and Figure 11).

5.3.3 Hypothesis H3 — Practical Decision-Making

- **Premise.** A calibrated FPIS threshold aligned with the cost-benefit ratio of false positives and false negatives should yield an expected return ≥ 15% higher than a random strategy in the same sector.
- **Results.**
    - Decision Curve Analysis (DCA) showed FPIS yields positive Net Benefit at thresholds relevant to investors.
    - In DCA, Net Benefit at $\tau = 0.17 \approx 0.038$.
    - In ROI simulation with benefit_TP:cost_FP = 5:1:
        - Optimal uncalibrated threshold $\tau \approx 0.490$, mean benefit FPIS ≈ 877 vs. −44 random ($\Delta \approx 921$).
        - Optimal calibrated threshold $\tau \approx 0.160$, mean benefit FPIS ≈ 902 vs. −49 ($\Delta \approx 951$).

- **Interpretation.** FPIS not only outperforms the baseline in discrimination but also maximizes expected investor value. With ROI exceeding the 15% threshold, H3 is validated (Figure 12 and Figure 13).

5.4 Interpretability with SHAP

Beyond predictive performance, an essential requirement of FPIS is interpretability. To ensure that investors can understand which team characteristics drive the score, the SHAP (SHapley Additive exPlanations) framework was applied, widely used in machine learning to explain both global and individual predictions.

5.4.1 Global SHAP — Variable Importance

Global importance analysis showed that FPIS is primarily driven by:

- Openness traits, especially *adventurousness*.
- Conscientiousness, with strong contributions from *achievement striving* and *orderliness*.
- Extraversion, linked to dynamism and activity level.
- Complementarity Index, capturing the team's psychological diversity.
- Number of founders (*org_numfounders*), confirming that intermediate team size maximizes success probability.

These variables form the FPIS core, consistent with both prior literature and business intuition: founding teams that are open to new experiences, disciplined, and diverse yet complementary have higher probabilities of achieving an exit (Figure 14).

5.4.2 Individual SHAP — Startup-Level Explainability

SHAP also enables individual-level analysis, decomposing a startup's score into positive and negative variable contributions. For example:

- A team with high openness and strong complementarity will receive a positive push toward a high FPIS.
- Conversely, excessive neuroticism or extreme team sizes (1 founder or more than 6) may significantly reduce predicted probability.

This type of analysis provides traceability and builds investor trust, allowing justification of why a specific startup receives a high or low score.

5.4.3 Business Implications

The integration of SHAP strengthens FPIS as a practical tool because it:

1. Facilitates communication with investors, providing clear evidence of which psychological dimensions drive the score.
2. Enables qualitative comparisons, letting investors evaluate not only success probability but also the reasoning behind it.
3. Increases model acceptance, as transparency helps overcome resistance to using algorithms in critical decision processes.

6. Discussion

The analysis confirms that the personality and composition of founding teams provide relevant predictive signals regarding the probability of startup success. However, the results also call for a critical reflection on applicability, business implications, and the limitations of the chosen approach.

6.1 Relevance for Venture Capital

- **Deal-flow prioritization.** Enables investors to rank startups by probability of success, focusing analytical resources on those with the greatest potential.
- **Initial screening tool.** In large portfolios, FPIS can serve as a preliminary filter, reducing the cost of evaluating hundreds of opportunities.
- **Support for collective decision-making.** Because it is interpretable, FPIS can be incorporated into investment committees as complementary evidence alongside analysts' intuition.

6.2 Hypothesis Fulfillment

- **H1 (added value):** Partially confirmed, as FPIS outperforms the baseline across all metrics, though the AUC increase was below the initially expected 5 p.p.
- **H2 (segmentation):** Strongly confirmed, since the top FPIS quintile multiplies the success rate of the bottom quintile by more than 5, far exceeding expectations.
- **H3 (practical decision-making):** Confirmed, with an expected ROI greater than 15%, validating FPIS as an actionable tool.

6.3 Limitations of the Approach

1. **Dataset dependency.** FPIS was trained on the dataset of McCarthy et al. (2023), which infers personality from Twitter activity, introducing a bias toward founders active on that platform.
2. **Uneven geographic coverage.** The predominance of U.S. startups (≈ 48%) may limit the generalization of results to ecosystems with different dynamics.
3. **Correlational nature.** While the models capture predictive patterns, they do not establish direct causal relationships between psychological traits and success.
4. **Restricted success definition.** The target variable defines success as IPO or acquisition, excluding other relevant forms such as sustained profitability or social impact.

6.4 Future Research Directions

- Enrich data sources by incorporating LinkedIn, GitHub, or other professional platforms to reduce dependence on Twitter.
- Segment by sector or geography, training models that capture local or industry-specific dynamics.
- Broaden the definition of success to include survival, funding milestones, or international scalability.

- Integrate FPIS into digital tools, such as dashboards or APIs, allowing investors to calculate the score of new founding teams in real time.

7. Conclusions

In summary, this work demonstrates that the personality and psychological diversity of founding teams contain predictive and actionable information for estimating a startup's probability of success. FPIS, as a calibrated, interpretable, and empirically validated score, represents a step toward the professionalization of founder analysis in venture capital and opens the door to its integration into practical tools for early-stage startup evaluation.

8. References

- Harrison, D. A., & Klein, K. J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review, 32*(4), 1199–1228.
- McCarthy, P. X., Gong, X., Braesemann, F., Stephany, F., Rizoiu, M. A., & Kern, M. L. (2023). The impact of founder personalities on startup success. *Scientific Reports, 13*(1), 17200.
- Rauch, A., & Frese, M. (2007). Let's put the person back into entrepreneurship research: A meta-analysis on the relationship between business owners' personality traits, business creation, and success. *European Journal of Work and Organizational Psychology, 16*(4), 353–385.
- Van Rossum, G., & Drake Jr, F. L. (1991). *Python Tutorial*. Release 2.

9. Links

- Drive

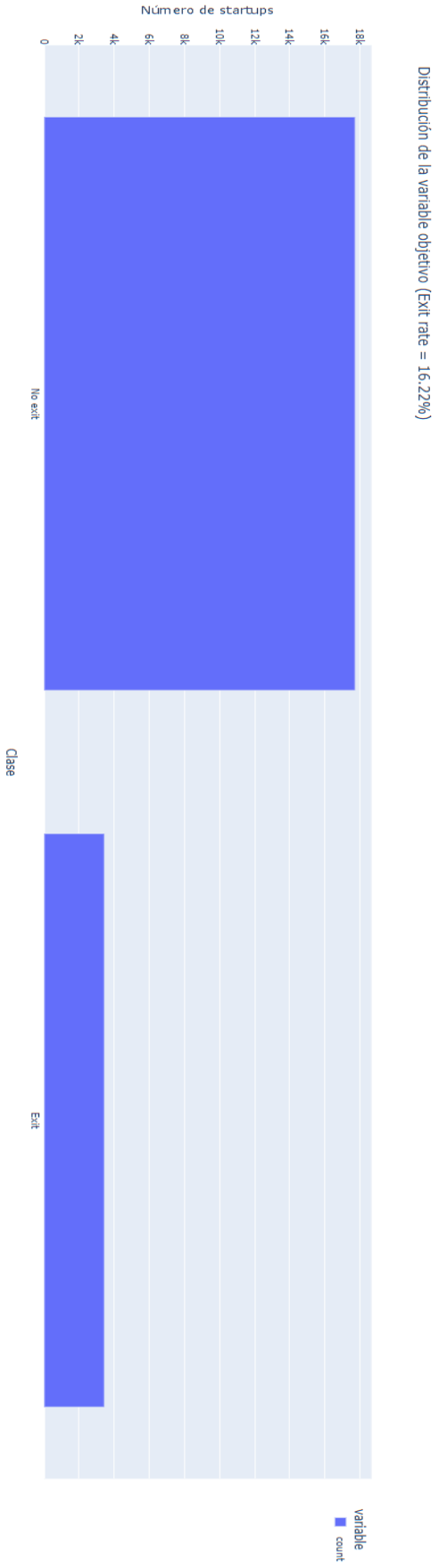
https://drive.google.com/drive/folders/1Wt6KYItZSN6oZaZB3S1YoLnqlNB3FCTz?usp=sharing

- GitHub

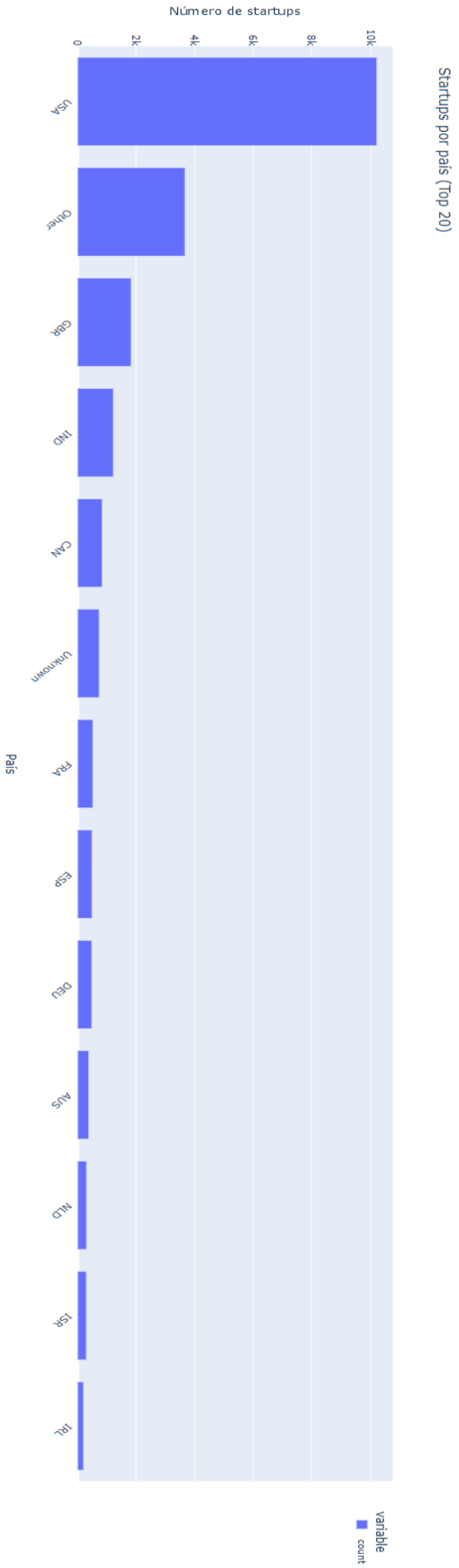
https://github.com/dariodiazbarross/Founder-Personality-Investment-Score-FPIS-

10. Appendices

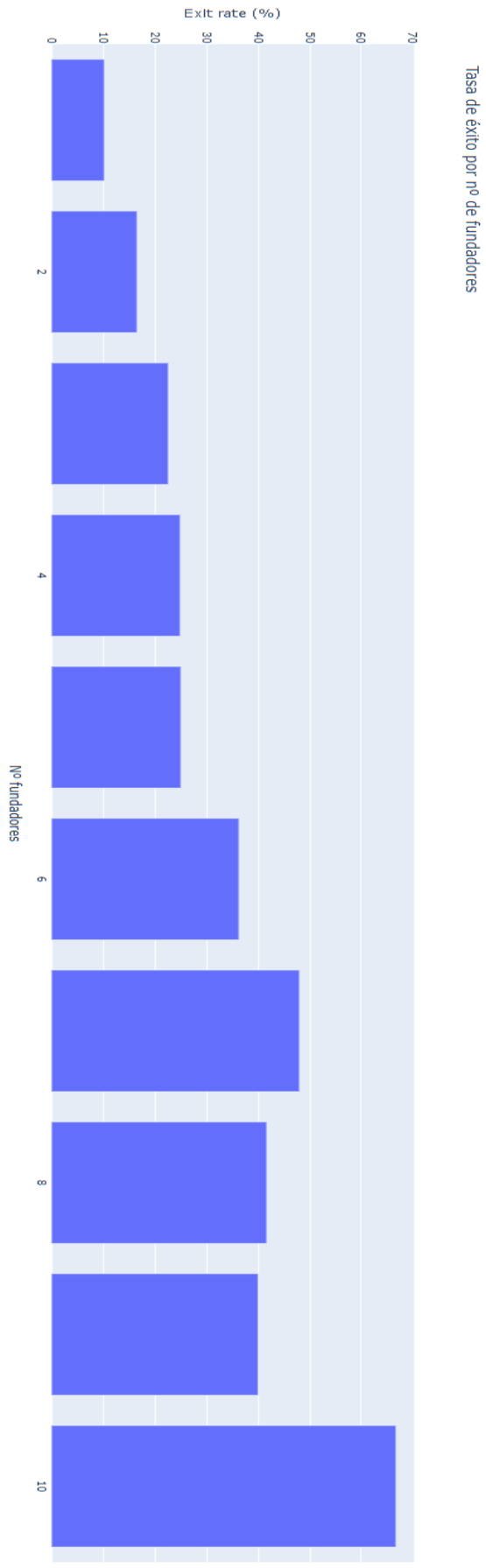
10.1 Figure 1


Distribución de la variable objetivo (Exit rate = 16.22%)
Número de startups
0
2k
4k
6k
8k
10k
12k
14k
16k
18k
Clase
No exit
Exit
variable
count

10.2 Figure 2


Startups por país (Top 20)
Número de startups
0
2k
4k
6k
8k
10k
País
USA
Other
GBR
IND
CAN
Unknown
FRA
ESP
DEU
AUS
NLD
ISR
IRL
variable
count

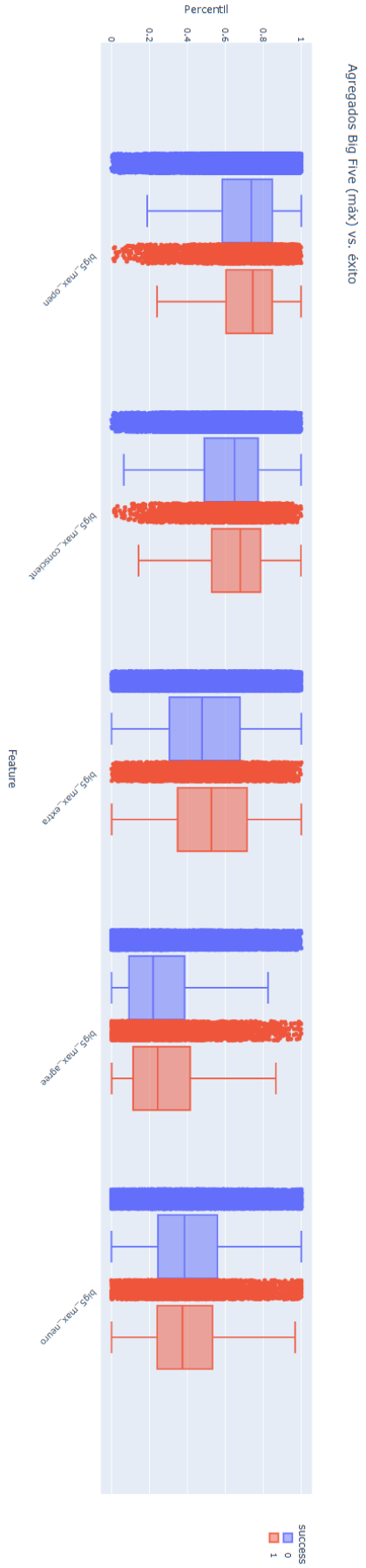
10.3 Figure 3


Exit rate (%)
0
10
20
30
40
50
60
70
Nº fundadores
2
4
6
8
10
Tasa de éxito por nº de fundadores

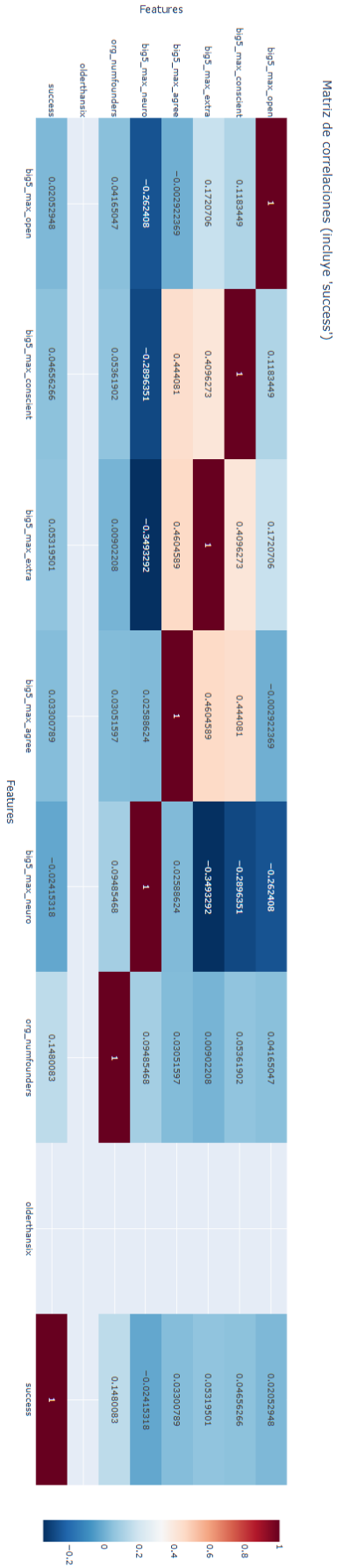
10.4 Figure 4


Tasa de éxito por país (Top 20)
Exit rate (%)
0
5
10
15
20
País
USA
ISR
FRA
DEU
CAN
ESP
GBR
AUS
NLD
IRL
Other
IND
Unknown
variable
success

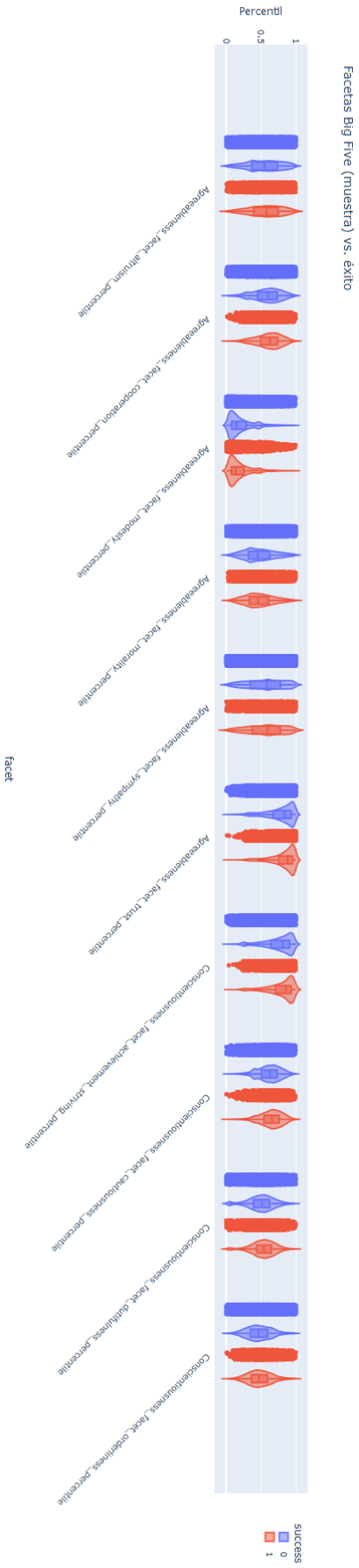
10.5 Figure 5


Agregados Big Five (máx) vs. éxito
Percentil
0
0.2
0.4
0.6
0.8
1
Feature
big5_max_open
big5_max_conscient
big5_max_extra
big5_max_agree
big5_max_neuro
success
0
1

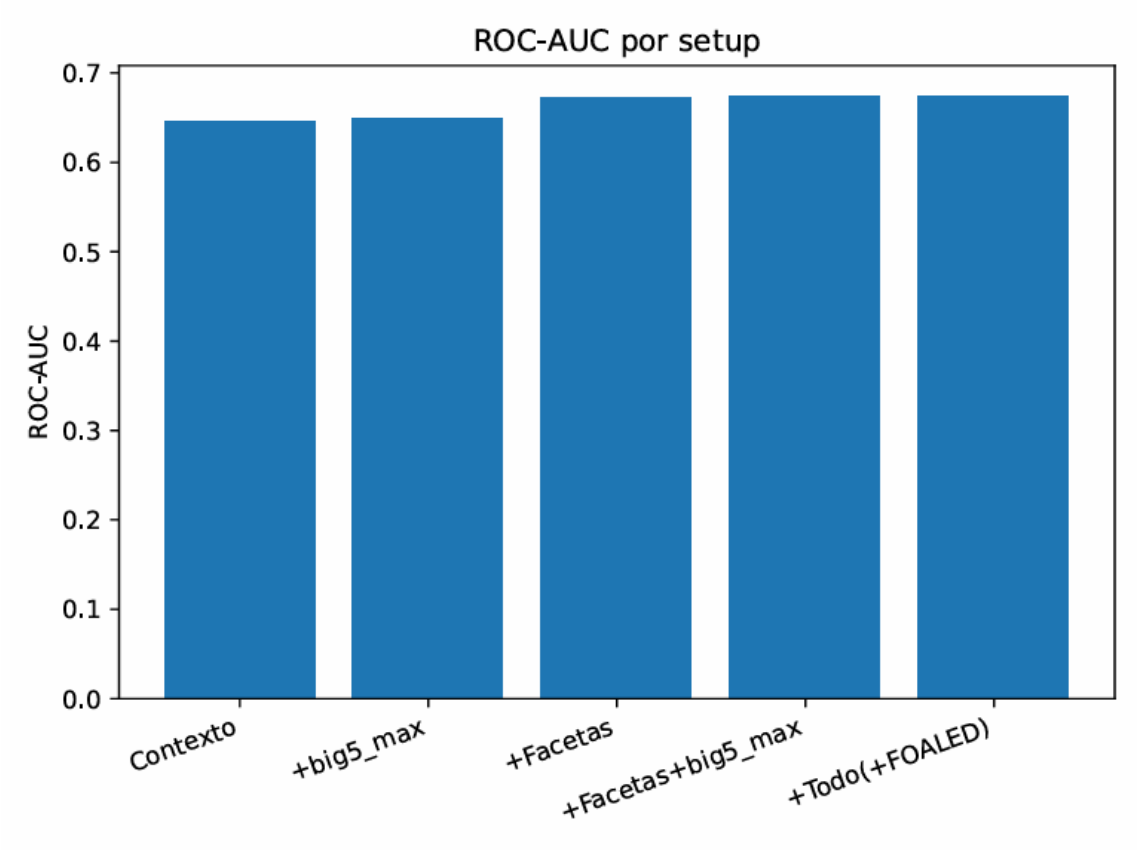
10.6 Figure 6


Matriz de correlaciones (incluye 'success')
Features
big5_max_open
big5_max_conscient
big5_max_extra
big5_max_agree
big5_max_neuro
org_numfounders
olderthansix
success
1
0.1183449
0.1720706
−0.002922369
−0.262408
0.04165047
0.02052948
0.1183449
1
0.4096273
0.444081
−0.2896351
0.05361902
0.04656266
0.1720706
0.4096273
1
0.4604589
−0.3493292
0.00902208
0.05319501
−0.002922369
0.444081
0.4604589
1
0.02588624
0.03051597
0.03300789
−0.262408
−0.2896351
−0.3493292
0.02588624
1
0.09485468
−0.02415318
0.04165047
0.05361902
0.00902208
0.03051597
0.09485468
1
0.1480083
0.02052948
0.04656266
0.05319501
0.03300789
−0.02415318
0.1480083
1
−0.2
0
0.2
0.4
0.6
0.8
1

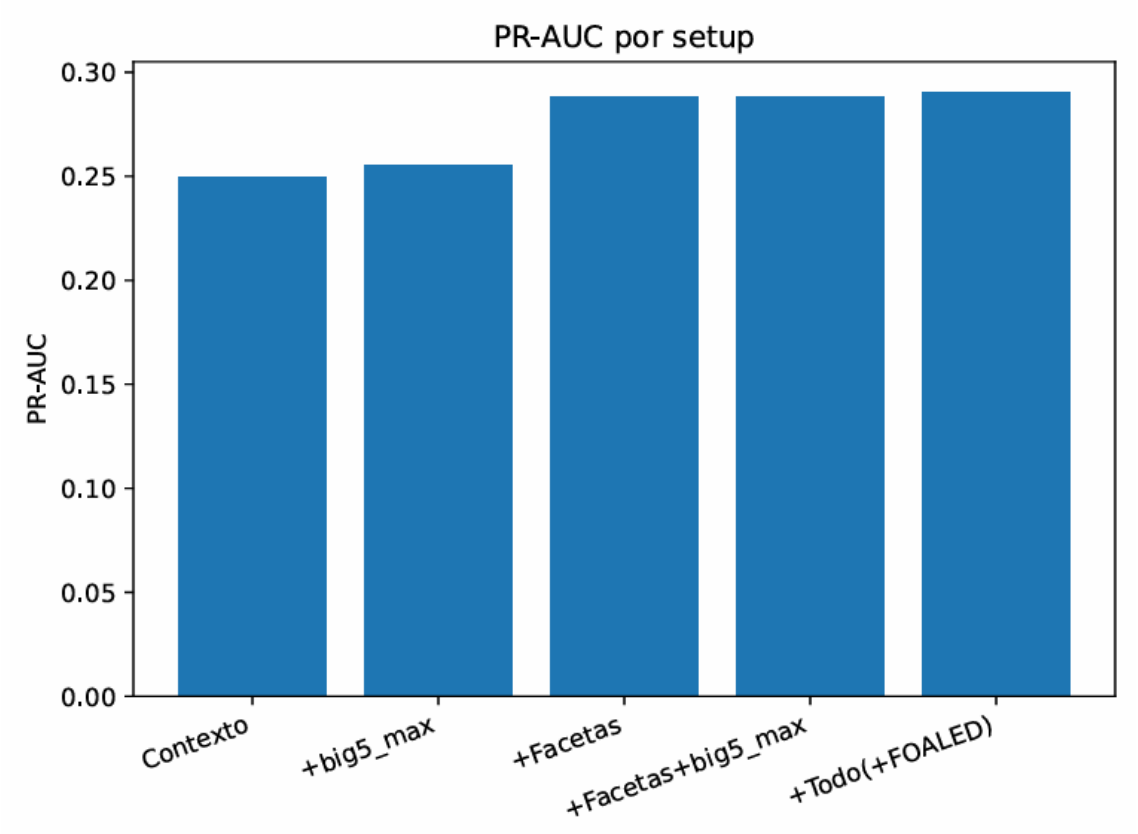
10.7 Figure 7


Facetas Big Five (muestra) vs. éxito
Percentil
0
0.5
1
facet
Agreeableness_facet_altruism_percentile
Agreeableness_facet_cooperation_percentile
Agreeableness_facet_modesty_percentile
Agreeableness_facet_morality_percentile
Agreeableness_facet_sympathy_percentile
Agreeableness_facet_trust_percentile
Conscientiousness_facet_achievement_striving_percentile
Conscientiousness_facet_cautiousness_percentile
Conscientiousness_facet_dutifulness_percentile
Conscientiousness_facet_orderliness_percentile
success
1
0

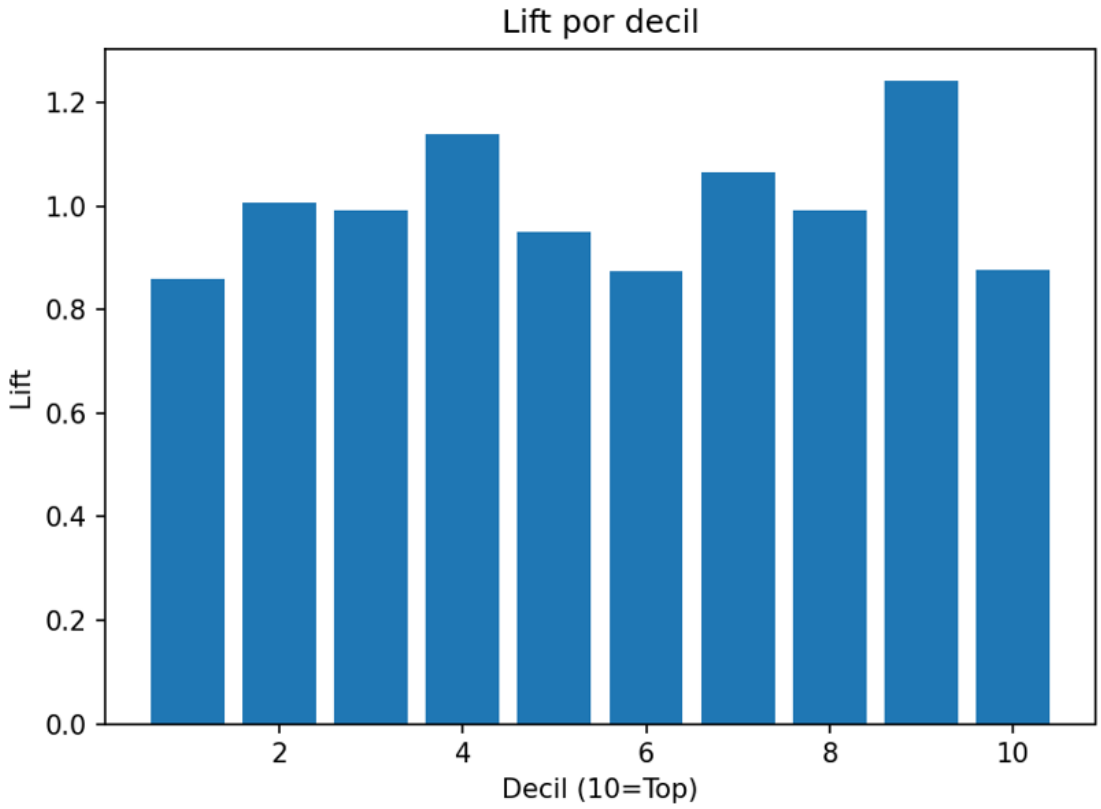
10.8 Figure 8



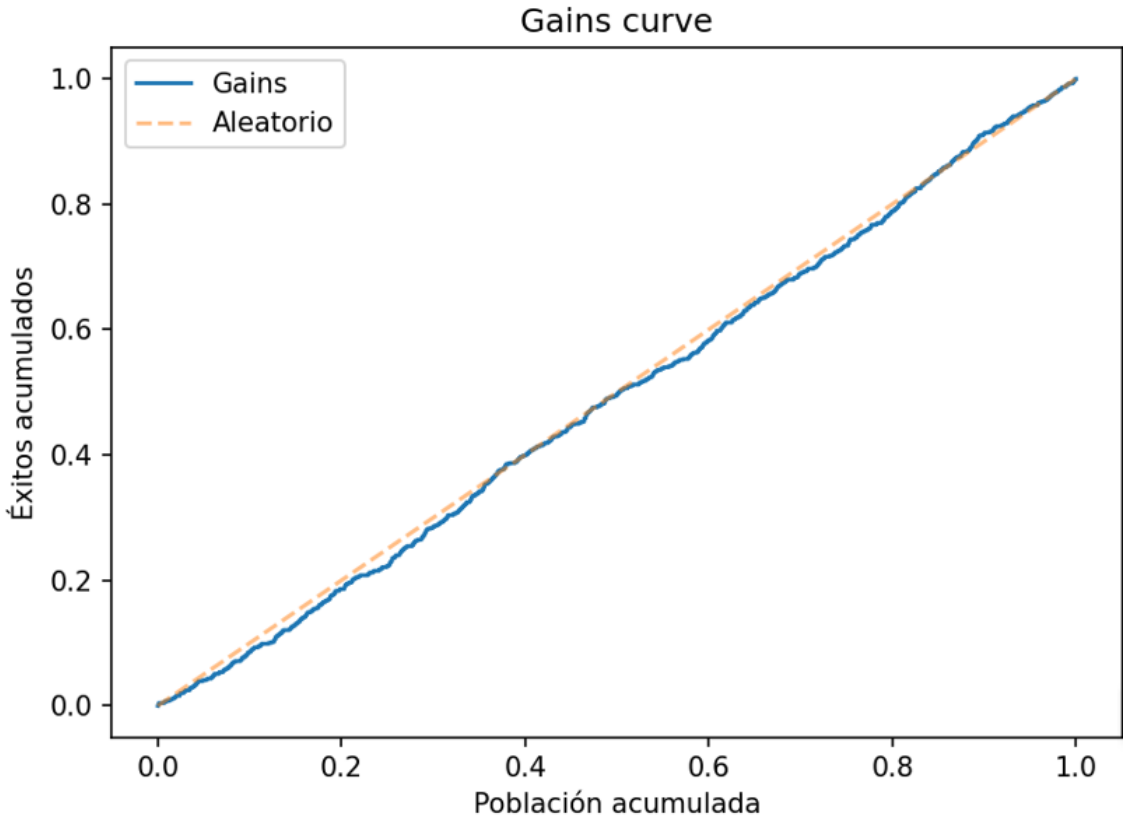
10.9 Figure 9



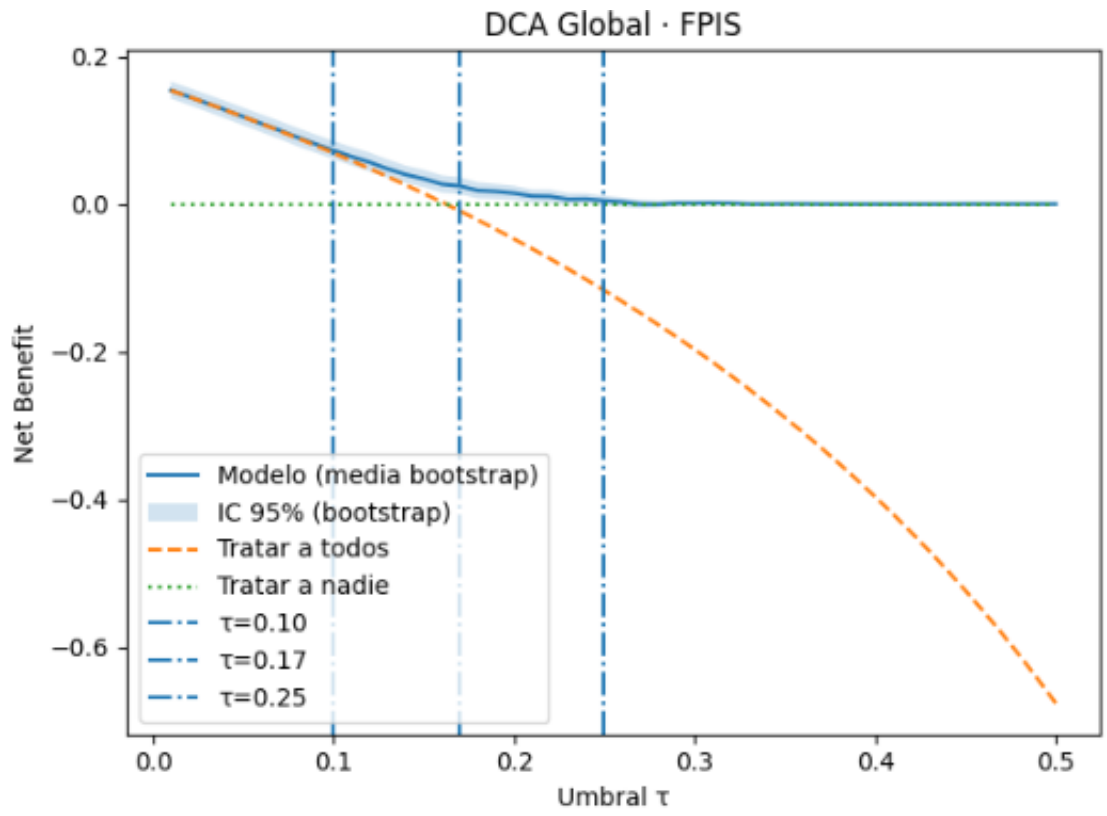
10.10 Figure 10



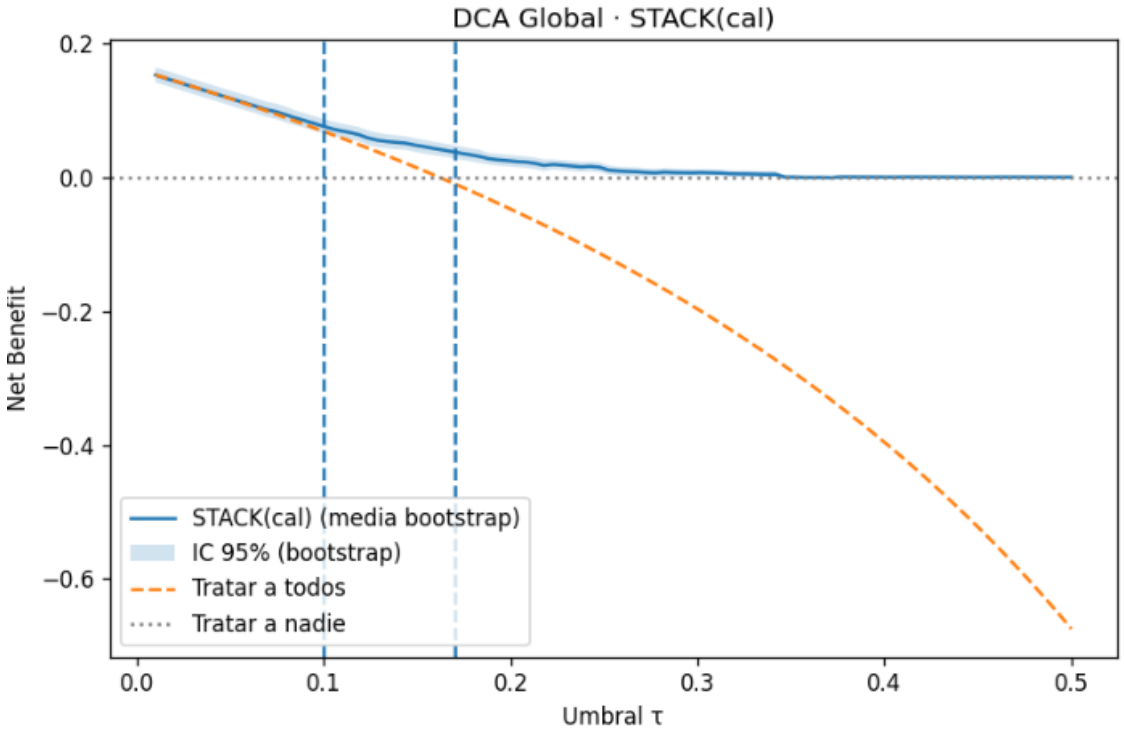
10.11 Figure 11



10.12 Figure 12



10.13 Figure 13



10.14 Figure 14

Generando SHAP summary plot…