NYU Stern School of Business

Data Science for Business

TECH-GB 2336 Section 30

Crowdfunding Success Prediction for Startup Incubator

Group 6

Submitted By:

Shashank	Dugad	sd5957@nyu.edu
Anshi	Shah	ans10020@nyu.edu
Run	Zhang	rz2762@nyu.edu
Jessie	Yu	cy2879@stern.nyu.edu

Table of Contents

Executive Summary

Crowdfunding has become a vital funding avenue for startups globally, allowing entrepreneurs to raise capital directly from backers via online platforms. In Turkey, a rapidly growing startup ecosystem, evidenced by a nearly 6-fold increase in investments from 2021 to 2024 ($4.7 billion) and the emergence of seven unicorns, highlights the potential of this mechanism. Turkey's largest startup incubator supports early-stage ventures by refining business plans, connecting them with resources, and facilitating funding opportunities. However, the ecosystem faces a significant challenge: only 23.1% of crowdfunding campaigns succeed, and the national success rate of 28.7% remains below the threshold needed to reliably support startup growth.

This report presents a data-driven project to predict crowdfunding success and identify key factors contributing to the success, leveraging a dataset of 1,628 Turkish campaigns with 38 features. Through EDA, predictive modeling (including XGBoost and Decision Trees), and NLP analysis of campaign descriptions, we aim to equip the largest turkish startup incubator with actionable insights. The work addresses two key questions: Can we predict crowdfunding success rates for startups? And which factors drive success in Turkey's ecosystem? The following sections detail our methodology, findings, and recommendations to enhance funding outcomes.

Business understanding

Context and Motivation

Turkey's startup ecosystem is experiencing rapid growth, supported by government-led initiatives like the Turcorn 100 program and tax-incentivized technology development zones. Despite this, the incubator faces a critical hurdle: 56% of its startups struggle to secure follow-on funding after initial rounds. Traditional VC funding remains highly competitive and selective, leaving a large portion of early-stage ventures underfunded.

Crowdfunding offers a compelling alternative. It provides startups with access to early capital, customer validation, and brand visibility without ceding equity. However, only 28.7% of crowdfunding campaigns in Turkey succeed, mirroring the global success gap. For our team, Turkey's largest incubator, closing this success gap could be a strategic and financial opportunity.

Crowdfunding in Turkey: A Localized Overview

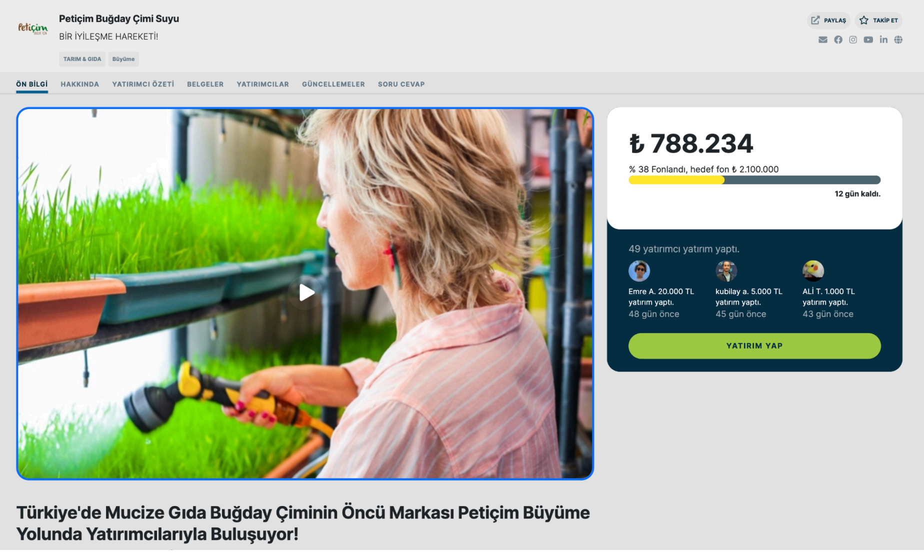
Crowdfunding in Turkey has emerged as a vital alternative financing mechanism, particularly for startups and small to medium-sized enterprises (SMEs) seeking to overcome traditional funding barriers. The process involves entrepreneurs presenting their projects on online platforms to solicit small contributions from a large number of individuals, collectively achieving their funding goals.

The Turkish crowdfunding landscape has been significantly shaped by regulatory developments. In 2017, amendments to the Capital Markets Law No. 6362 established a legal framework for equity-based crowdfunding, further detailed by the Capital Markets Board (CMB) through the Communiqué on Equity-Based Crowdfunding published in October 2019. This regulatory environment has fostered the growth of various crowdfunding platforms across the country.

Notable platforms include Fongogo, Fonbulucu, and Burada with different focuses. These platforms serve as intermediaries, ensuring compliance with regulatory standards and providing the necessary infrastructure for campaign management. The typical crowdfunding process in Turkey encompasses the following stages:

1. Project Ideation and Planning: Entrepreneurs develop their business concepts and prepare detailed plans outlining their objectives and funding requirements.
2. Campaign Submission and Approval: The project undergoes a review process to ensure it meets the necessary criteria and regulatory compliance by the platform.
3. Campaign Launch: Upon approval, the campaign is launched on the platform, making it accessible to potential backers.
4. Promotion and Engagement: Entrepreneurs actively promote their campaigns through various channels, including social media, to attract backers and maintain engagement.
5. Fundraising Period: The campaign remains live for a predetermined duration, during which backers can contribute funds.
6. Campaign Conclusion and Fund Allocation: If the funding goal is met, the collected funds are transferred to the entrepreneur, who will then deliver the rewards.

This structured approach, underpinned by a supportive regulatory framework and a growing number of dedicated platforms, has positioned crowdfunding as a significant enabler of entrepreneurial activity in Turkey.



Peticim Buğday Çimi Suyu

BİR İYİLEŞME HAREKETİ!

₺ 788.234

% 38 Fonlandı, hedef fon ₺ 2.100.000

12 gün kaldı.

49 yatırımcı yatırım yaptı.

Emre A. 20.000 TL yatırım yaptı. 48 gün önce

kubilay a. 5.000 TL yatırım yaptı. 45 gün önce

ALİ T. 1.000 TL yatırım yaptı. 43 gün önce

Türkiye'de Mucize Gıda Buğday Çiminin Öncü Markası Peticim Büyüme Yolunda Yatırımcılarıyla Buluşuyor!

A typical webpage of a crowdfunding campaign in Turkey

Business Problem

Our incubator currently lacks a scalable, data-driven framework to assess and improve the crowdfunding readiness of its startups. With no clear roadmap for identifying high-potential campaigns or optimizing campaign design, promising ideas risk being lost in the noise.

Project Objective

This project primarily seeks to empower the incubator with:

- A predictive model to estimate a startup's probability of crowdfunding success.
- Insights into the most influential factors driving campaign outcomes.
- NLP analysis of campaign descriptions to guide founders on effective storytelling.

Successful campaigns signal market validation, attracting VC investment. By transforming raw data into actionable intelligence, the incubator can better guide founders, allocate internal support, and increase the chances of funding success across its portfolio.

Key Business Stakeholders and Benefits

- Incubator Program Directors: Identify and prioritize support for high-potential campaigns.
- Startup Mentors: Use model insights to coach founders and design crowdfunding campaigns
- Startup Founders: Receive tailored campaign design recommendations and implementation.
- Investors & Government Agencies: Identify campaigns with strong social and economic ROI, and allocate resources for the ecosystems. By boosting funding success, the incubator can drive innovation and job creation, amplifying Turkey's economic impact.

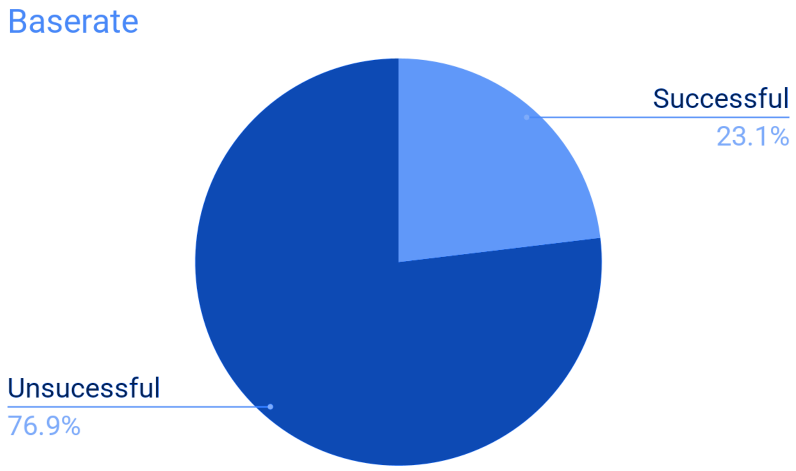
Dataset and Preparation

Dataset Overview

The dataset originates from the UCI Machine Learning Repository, collected in 2022 across 6 Turkish crowdfunding platforms. Authored by Kilinc and Aydin (2023), it contains 1,628 campaigns with 38 features tracking campaign mechanics, creator profiles, and regional dynamics. We've also attached a translated version of it in the appendix.

Among all the campaigns or rows in the dataset, there are 76.9% of them being successful, with 23.1% marked as unsuccessful. There's the problem of imbalanced base rate in the dataset, which is expected as most startups are likely to fail.

Feature Type	Count
Numeric	21
Categorical	17

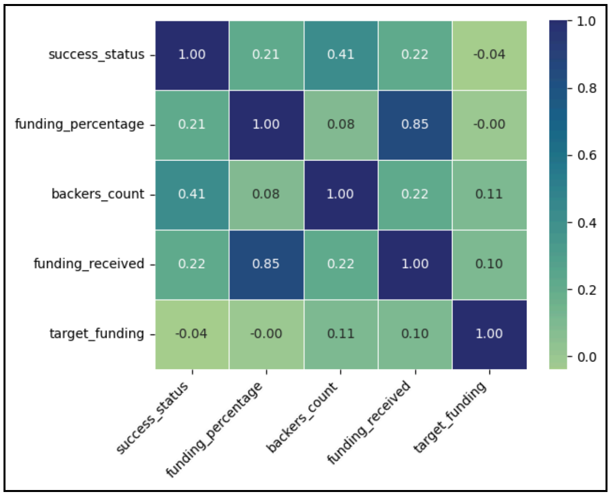


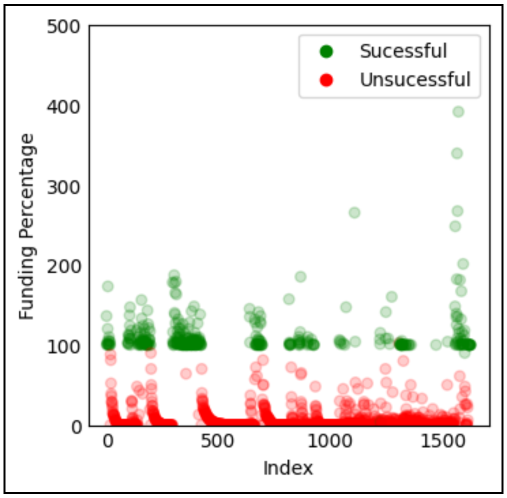
Overall, the features can be roughly classified into three categories. The first is the general information of the startup, including platform name, funding method, region, etc. The second type is the information about the owner, such as the owner's gender, and how many projects are owned by the owner at the same time. The last features' category is marketing, and it has recorded things like whether the startup has a website, and how many social media followers it has.

Because some features are valid only in the Turkish area, this dataset was particularly useful for understanding the startup atmosphere in Turkey around 2022. While it might not be as useful to generalize the findings to other regions and different time horizons, it's still a great source for getting an initial overview of what the most important factors are in a startup's success.

Data Preparation

In the data preparation stage, we first tried to understand the meaning of the features and their necessity in this task. As we plotted the correlations graph below for some features:





It became clear that some features, such as ‘funding_percentage’ and ‘funding_received’, have a strong correlation, and therefore, we may not need all of them in our modeling. Additionally, the ‘funding_percentage’ feature also displayed a high correlation with our target-success_status. As we further plotted the visualization on it, we realized this feature may be the definition of our target ‘success_status’, because any startups above a 100% funding percentage have been marked as successful, otherwise unsuccessful. As a result, we dropped both the features that are highly correlated with other features and the features that are highly correlated with our target variable.

The other part of data preparation was to translate the data. Except for numerical data and specific Turkish names or regions, we used the Google Translate API to translate features like project descriptions into English. It ensures the efficiency of our analysis later on. Also, translating the project description was an essential step for the second part of our modelling on text processing on the dataset, as an endeavor to capture the relationship between different word usage in the project description, and the success status of the project.

With regard to missing data, we've identified that the ‘start_date’ and ‘end_date’ include 611 and 553 missing entries, which we decided to drop for future modelling. The region feature happened to have one single missing value, which we filled with the most frequently appearing value: Marmara. During the data preparation stage, we also turned the categorical features into one-hot representation with drop_first set to be true.

Exploratory Data Analysis

The major Exploratory Data Analysis (EDA) we have done was to find out whether the values of some features have played significant roles in deciding a startup's success through group_by analysis.

The Results were shown in the following table:

Backer	Success %
0	0%
1-10	4.1%
11-50	**60.9%**
51-100	85.0%
101-500	91.7%
501+	-

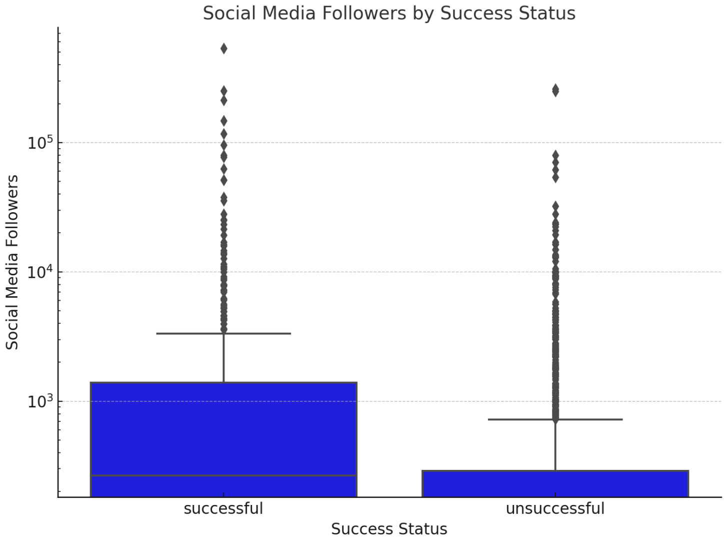
Region	Success %
Akdeniz	**46.2%**
Marmara	**30.6%**
Genel	22.2%
İç Anadolu	20.0%
Unknown	18.2%
Güneydoğu	16.7%

Platform	Success %
Buluşum	**100%**
Ideanest	**100%**
Arıkovanı	66.7%
Fongogo	27.5%
Fonbulucu	11.4%
Crowdfon	4.7%

Category	Success %
Social Responsibility	**100%**
Dance-Performance	50%
Film-Video-Photograph	36%
Health-Beauty	33%
Food and Beverage	33%
Education, Tourism, Publishing	~28%-33%
Technology, Environment, Music	~10-12%
Animal, Fashion, Design	0%

As shown in the above tables, we can already draw some conclusions from the group_by analysis. For instance, when the number of backers has increased to 11-50, the success rate of the startup would jump significantly to 60.9%. And, startups based in some regions and platforms also display privilege compared to other alternatives. More importantly, we also identified that startups working in specific areas are more likely to succeed than others. Those working on social responsibility have displayed a 100% success rate, while technology and fashion design surprisingly underperform. It may have something to do with the denominator as it's more challenging to succeed in competitive areas. We will further verify these conclusions with the machine learning models.

From another analysis, we learn that higher social media followers night have a positive impact on success rate. The boxplot below indicates that greater social media presence might be positively associated with campaign success.



- Successful campaigns have a higher median social media follower count (around 1,200) compared to unsuccessful ones (around 600).
- The distribution for successful campaigns is more right-skewed, with numerous high-follower outliers reaching up to 500k.
- Unsuccessful campaigns also show outliers, but fewer extreme values and a lower overall

Modeling & Evaluation

Strategy

The objective is to predict whether a crowdfunding campaign will be successful or not. The target variable success_status is binary, where:

- 1 indicates a successful campaign
- 0 indicates an unsuccessful campaign

Given the potential for class imbalance (i.e., more unsuccessful projects than successful ones), we employed SMOTE (Synthetic Minority Oversampling Technique) to balance the training dataset.

The target column success_status was label-encoded as binary.

The features were divided into two categories and dealt with accordingly.

- Categorical Features: platform name, crowdfunding type, category, funding method, owner gender, location, region, and campaign media presence indicators.
- Numerical Features: All remaining continuous or discrete numeric c,olumns excluding success_status.

Preprocessing

Numerical features: Imputed with median, then standardized.

Categorical features: Imputed with most frequent and one-hot encoded, dropping the first category to avoid multicollinearity.

A ColumnTransformer was used to encapsulate the preprocessing pipeline. A stratified 80-20 split was used to maintain class balance across sets. To address the imbalance, we applied SMOTE only on the training set.

Models Built

Model 1: Decision Trees

A GridSearchCV with StratifiedKFold (5 splits) optimized for F1-score due to class imbalance. The best parameters found were:

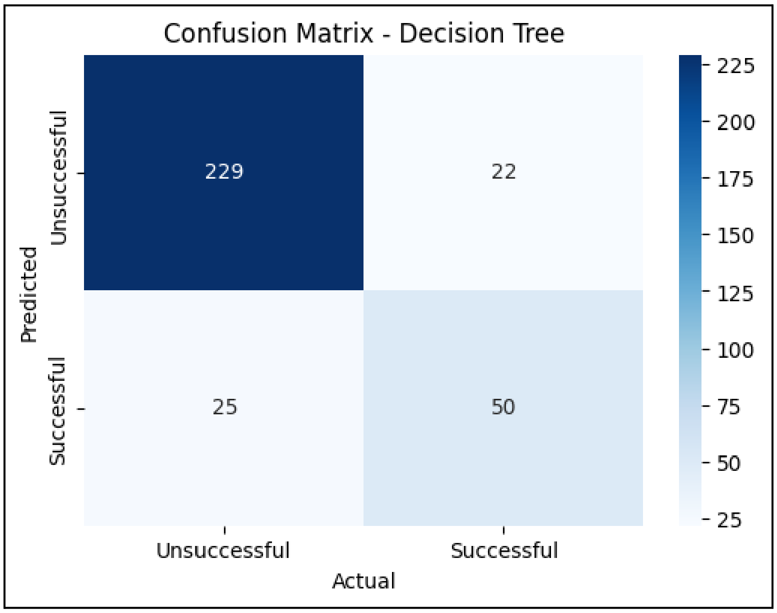
- max_depth: 80
- min_samples_split: 10

Best cross-validated F1 score: 0.8809

Test Set Performance:

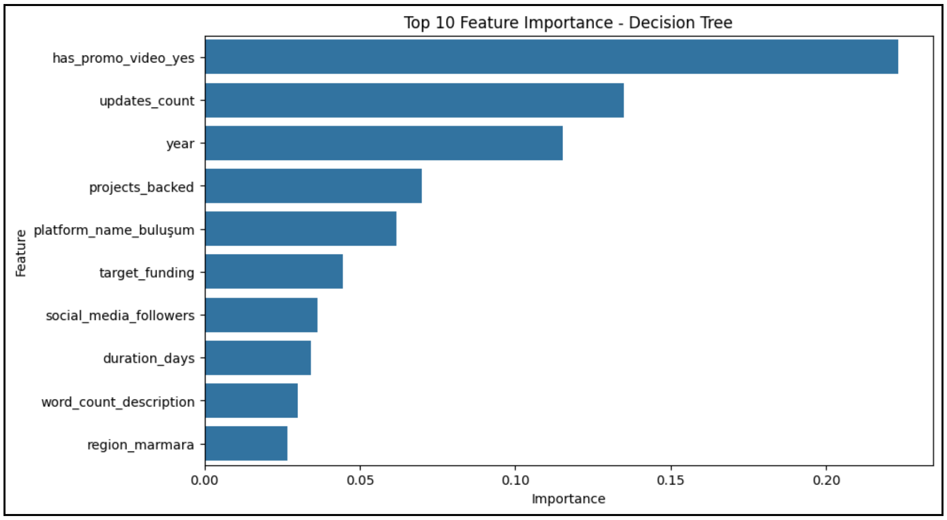
Metric	Score
Accuracy	0.86
Precision (1)	0.69
Recall (1)	0.67
F1-Score (1)	0.68
ROC-AUC	0.8251
PR-AUC	0.5966

Confusion Matrix:



The model shows good performance on majority class (unsuccessful campaigns), but moderately underperforms in detecting successful ones, which is a typical challenge in imbalanced classification.

Feature Importance:



Top 10 features driving decision tree predictions included a mix of numerical metrics and categorical campaign indicators. These features can guide campaigners on key success drivers (e.g., region, platform, video presence).

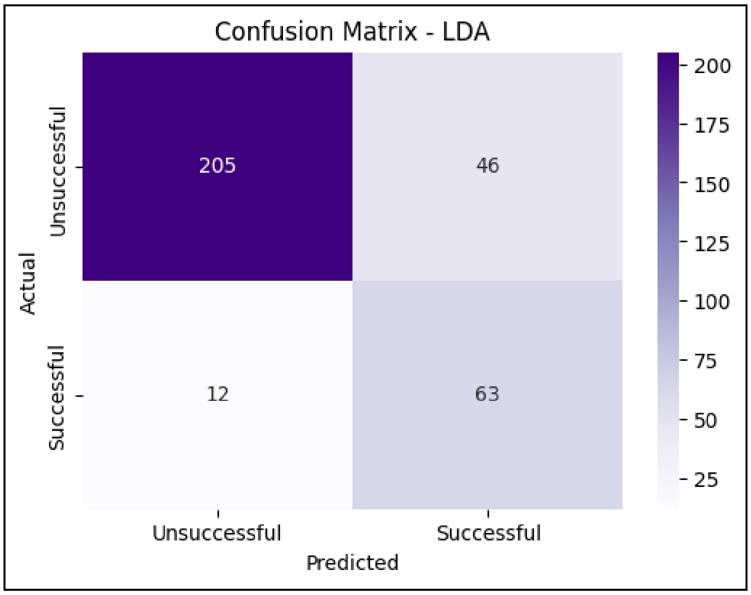
Model 2: Linear Discriminant Analysis

LDA was applied as a linear baseline after converting sparse matrices (from one-hot encoding) to dense format.

Test Set Performance:

Metric	Score
Accuracy	0.84
Precision (1)	0.66
Recall (1)	0.64
F1-Score (1)	0.65
ROC-AUC	0.7994
PR-AUC	0.5482

Confusion Matrix:



LDA provides a simpler, interpretable model but lacks the complexity needed to handle feature interactions or non-linearity.

Model 3: Support Vector Machine

Kernel: Radial Basis Function (rbf)

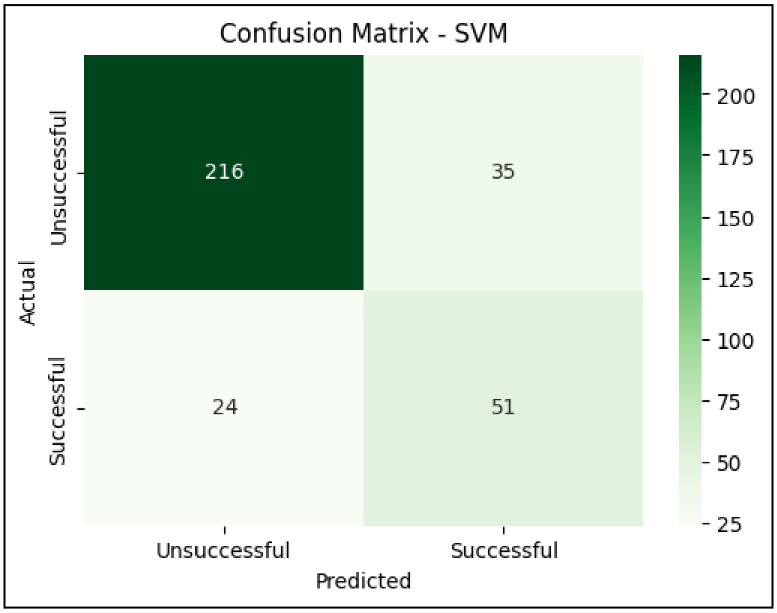
Class Weight: balanced (to address class imbalance)

Probability Estimates: Enabled (probability=True)

Test Set Performance:

Metric	Score
Accuracy	0.84
Precision (1)	0.66
Recall (1)	0.64
F1-Score (1)	0.65
ROC-AUC	0.7994
PR-AUC	0.5482

Confusion Matrix:



The heatmap visually illustrated misclassifications, where some successful campaigns were predicted as unsuccessful. Nevertheless, the performance was solid for a non-tree-based model.

Model 4: XGBoost Classifier

Tree booster with max_depth=5, n_estimators=100, learning_rate=0.1

scale_pos_weight=1 due to prior SMOTE balancing

eval_metric='logloss'

Warnings about deprecated use_label_encoder parameter were safely ignored.

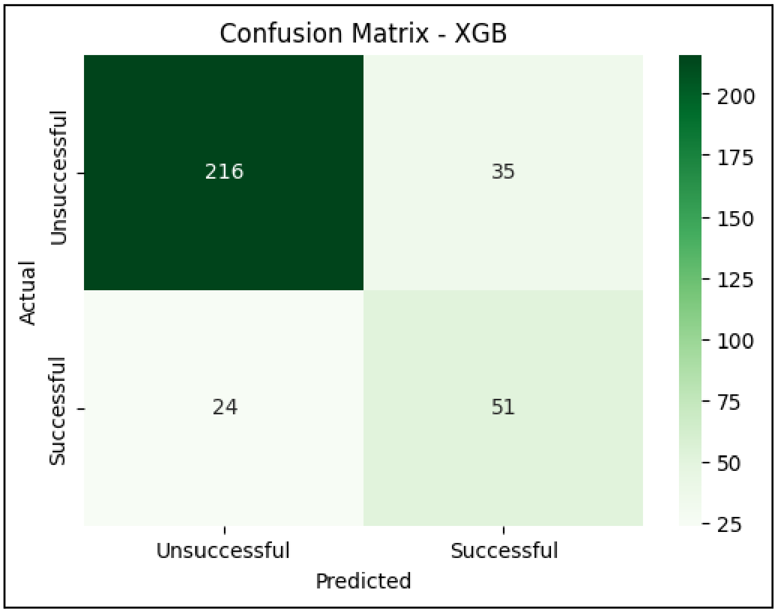
Test Set Performance:

Metric	Score
Accuracy	0.87
Precision (1)	0.74
Recall (1)	0.71
F1-Score (1)	0.72

ROC-AUC	0.9153
PR-AUC	0.8175

XGBoost clearly outperformed all previous models, especially in capturing successful campaigns. It provided high-quality probability estimates, making it suitable for ranking or threshold tuning in deployment scenarios.

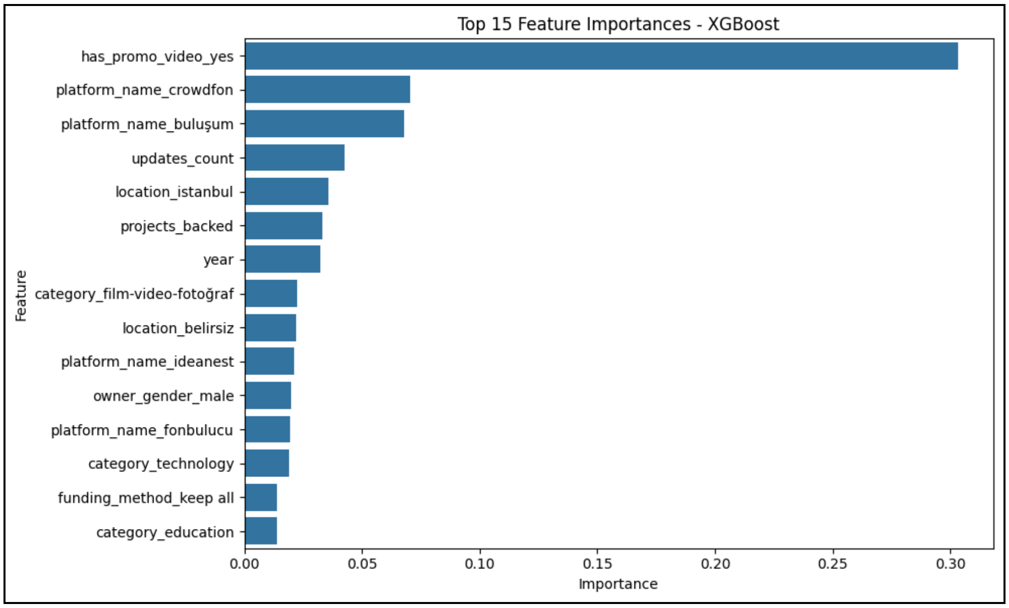
Confusion Matrix:



A strong diagonal in the heatmap confirmed good class separation. Slight misclassification in the minority class (successful) remains but was lower than with SVM or LDA.

Feature Importance:

Top features identified by XGBoost (via gain-based importance) included both engineered categorical variables and original numerical features. This reinforces XGBoost's strength in handling mixed data types and capturing nonlinearities.



Model 5: Logistic Regression

Class weight: balanced

Maximum iterations: 1000

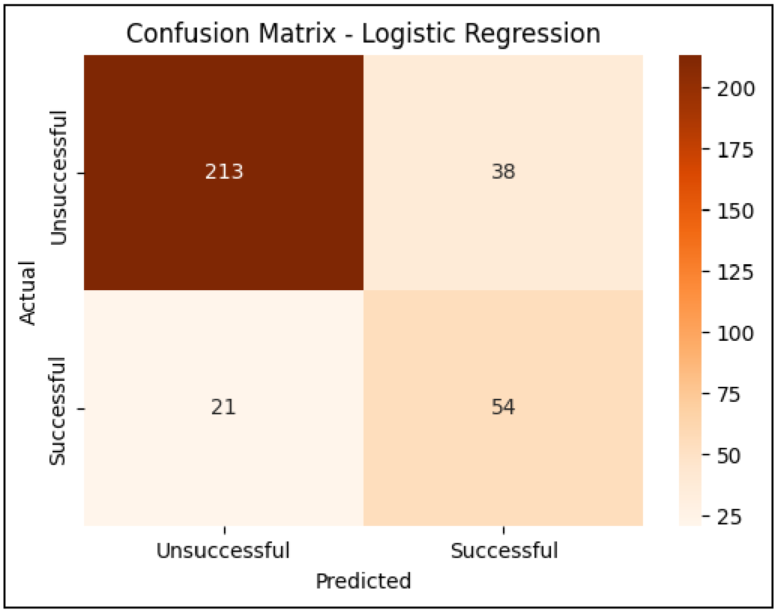
Solver automatically chosen based on data size (default)

Test Set Performance:

Metric	Score
Accuracy	~0.83 (inferred)
Precision (1)	~0.68
Recall (1)	~0.62
F1-Score (1)	~0.65
ROC-AUC	**~0.84**
PR-AUC	**~0.70**

Logistic Regression gave a competitive baseline with relatively good generalization. However, its linear nature made it less expressive compared to XGBoost or SVM for modeling complex feature interactions.

Confusion Matrix:



The heatmap revealed that the model tended to predict the majority class more confidently. Recall for successful campaigns was lower, which is expected for simpler models without ensemble learning.

Comparison and Evaluation

Model	Accuracy	F1-Score (1)	ROC-AUC	PR-AUC
Decision Tree	0.86	0.68	0.8251	0.5966
LDA	0.84	0.65	~0.80	~0.55
SVM	0.85~	0.67~	0.8683~	~0.60
XGBoost	**0.87**	**0.72**	**0.9153**	**0.8175**
Logistic Regression	0.83~	0.65	~0.84	~0.70

The best overall model is XGBoost, based on the highest F1, ROC-AUC, and PR-AUC scores.

Key Takeaways:

- XGBoost offers the best performance and interpretability via feature importances.
- SVM performs competitively but is less interpretable.

- Logistic Regression is fast, easy to implement, and a good linear benchmark.
- Feature importance plots can drive real-world campaign strategies, e.g., choosing the right category or platform.

Text Processing Using NLP

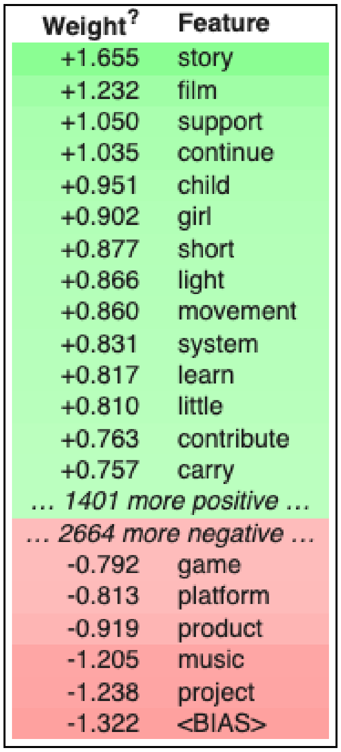
It was explored how natural language used in project descriptions correlates with the success of crowdfunding campaigns. The goal was to extract interpretable linguistic features that provide insight into what kind of messaging resonates with backers.

NLP Pipeline and Vectorization:

To process the campaign descriptions, we employed a standard NLP pipeline:

- Text cleaning: Lowercasing, punctuation removal, and lemmatization.
- Tokenization: Using nltk to split text into individual words.
- Stopword removal: To discard common but uninformative words like “the”, “is”, and “and”.

The processed text was then converted into numerical features using a TF-IDF Vectorizer. TF-IDF (Term Frequency–Inverse Document Frequency) helps identify terms that are not only frequent but also discriminative across documents. It allows the model to focus on meaningful words rather than common ones.\



Weight?	Feature
+1.655	story
+1.232	film
+1.050	support
+1.035	continue
+0.951	child
+0.902	girl
+0.877	short
+0.866	light
+0.860	movement
+0.831	system
+0.817	learn
+0.810	little
+0.763	contribute
+0.757	carry
… 1401 more positive …	
… 2664 more negative …	
-0.792	game
-0.813	platform
-0.919	product
-1.205	music
-1.238	project
-1.322	<BIAS>

Model Interpretation with Linear Classifier:

A linear classifier (such as Logistic Regression) was trained using the TF-IDF features. This model’s feature weights indicate how strongly each word contributes to the classification:

- Positive weights suggest that a word is predictive of a successful campaign.
- Negative weights imply association with unsuccessful campaigns.

Using model interpretation tools such as ELI5, we visualized and ranked the top contributing words.

Key Findings from Text Analysis:

Words with the highest positive weights were – story, support, child, learn, movement, continue, girl, light, and contribute were predominantly associated with emotionally engaging narratives, human-centered goals (e.g., education, community development, storytelling), and social responsibility and continuity. For example, successful projects are often described as a compelling journey –

"Production support campaign for a short film series whose every stage was projected 'online' for the first time in Turkey!"

"Documentary film about the true story of Street Children"

Words like music, product, platform, project, and game were associated with less successful campaigns. Such terms are often generic, linked to commercial or entertainment-focused projects, and less emotionally compelling or mission-driven. These campaigns underperformed in earlier group-by/category-level analysis, reinforcing the pattern.

Interpretation and Strategic Insights:

The analysis shows that language matters: campaigns that tell a story, highlight social impact, or evoke empathy are more likely to succeed. Descriptions filled with commercial or technical jargon fail to engage potential backers. These findings align with behavioral expectations—backers are influenced by narratives that promise emotional returns, not just financial or material ones.

Business Impact and Deployment

Our model enables strategic improvements both within the incubator and across the broader Turkish crowdfunding ecosystem.

Internal Impact: Enhancing Incubator Operations

1. **Startup selection optimized for crowdfunding**
   We can leverage the prediction model to assess incoming ventures for crowdfunding viability. Key factors include regional performance trends, platform success rates, and marketing strategies. This data-driven approach could ensure that resources are allocated to startups with the highest potential for successful crowdfunding campaigns.

2. **Targeted Mentor Coaching**
   Equip mentors with feature-driven insights from our model. Emphasize proven success factors like using promo videos, frequent campaign updates, and early backer engagement to optimize campaign performance.

3. **Personalized founder training**
   Develop playbooks tailored to each founder's strengths and gaps. Use insights from our NLP analysis to improve storytelling and build emotionally resonant campaigns. Provide storytelling templates that highlight keywords associated with success (e.g., "support," "story," "learn").

4. **Crowdfunding readiness scorecard development**
   Create a *standardized assessment tool* to evaluate startups on key success factors, including:
   - Campaign Design Quality: Clarity and appeal of the campaign narrative.
   - Social Media Presence: Follower count and engagement metrics.
   - Multimedia Usage: Inclusion of promotional videos or images.
   - Update Frequency: Regular communication with potential backers.
   - Early Backer Engagement: Initial traction in terms of backer numbers and funding velocity.

   Implementing this scorecard will guide startups in enhancing areas critical to campaign success.

5. **Real-time monitoring dashboards/tools**
   Build campaign dashboards to monitor:
   - Funding Velocity: Rate at which funds are being pledged.
   - Backer Growth: Increase in the number of backers over time.
   - Engagement Rates: Interactions on social media and campaign updates.

   Real-time insights facilitate agile decision-making and strategy adjustments during the campaign lifecycle. Also, this will enable early detection of underperforming campaigns and allows mentors to proactively intervene.

External Impact: Strengthening the Broader Ecosystem

1. **Investor Signaling**
   Campaigns flagged as "high-potential" by the model, especially those with strong narratives and ≥11 backers, can be promoted to VCs and angels as promising follow-on investments.

2. **Public Sector Policy Guidance**
   Regional success disparities (e.g., higher outcomes in Akdeniz and Marmara) can inform government programs aimed at leveling geographic startup opportunity.

3. **Strategic Platform Partnerships**
   Partner with high-performing platforms like Buluşum, Ideanest, and Fongogo to:
   - Expand campaign visibility
   - Offer platform-specific coaching

- Build startup-backers trust via platform endorsement

Future Work: Scaling and Sustaining Impact

1. **Campaign Simulation Dashboard Development**
   Build a tool to forecast funding likelihood under various feature configurations (e.g., with or without promotional videos), assisting startups in optimizing their campaign strategies.
2. **Model Enhancement with Additional Data**
   Incorporate real-time social media analytics and backer demographics into the predictive model to improve accuracy and relevance.
3. **Generalization to Emerging Markets**
   Although there are factors that are specific to Turkey market and the industry distribution might be different, we can also see that early backers traction and marketing effort would significantly impact the success of crowdfunding campaign. There are possibilities to adapt the model and insights to other emerging markets by integrating localized datasets, thereby extending the impact of the research beyond Turkey.

Ethical Considerations

- Bias Monitoring: The model may unintentionally favor certain categories (e.g., social campaigns) or regions. Ongoing fairness checks are required.
- Data Privacy: While most data is public (e.g., campaign descriptions), we must be aware not to expose personally identifiable information in analysis outputs.
- Founder Reputation: Overreliance on a scorecard may discourage startups in lower-ranked categories. Human judgment should complement model predictions.

Appendix

Contribution

Name	Contribution
Shashank	Topic research and finding data sets, developing proposal, applying machine learning algorithms to concoct predictive models, running groupby analysis, drawing insights from models, drafting the modeling and evaluation part for both presentation and report
Anshi	Topic research and finding data sets, developing proposal, defining business problem and context, applying all machine learning algorithms to concoct predictive model, running groupby analysis, drawing insights from models, drafting the modeling and evaluation part for both presentation and report
Run	Topic research and finding data sets, developing proposal, initial exploratory data analysis and updates, uncovering imbalanced data, applying machine learning algorithms, NLP for text analysis, drawing insights from models, drafting the dataset and EDA part for both presentation and report
Jessie	Topic research and finding datasets, developing proposal, dataset translation, initial exploratory data analysis and updates, refining business problems, running decision tree, structuring presentation/report and organizing analysis, drafting business and impact parts for both presentation and report

Links to Data

Original dataset (in Turkish):
https://archive.ics.uci.edu/dataset/1025/turkish+crowdfunding+startups

Translated dataset (Categorical features + Description):
https://drive.google.com/file/d/1fndxw8G0mOFgCWE-AP4Rt8Fmipg1XvYn/view?usp=drive_link

Link to Code

https://colab.research.google.com/drive/1XF1aEycQGw0xYlicbuiH0qmCZVi0F7JI?usp=drive_link