INSY 662: Data Mining and Visualization

Individual Project Report

Student Name: Helia Mahmood Zadeh

Student ID: 261224416

Date: December 6, 2025

Course: INSY 662 - Data Mining and Visualization

EXECUTIVE SUMMARY

This comprehensive report documents the complete machine learning pipeline developed to predict Kickstarter project success. The project analyzed 262,412 historical projects spanning multiple categories and geographical regions, applying both supervised and unsupervised learning techniques.68.91% test accuracy with proper generalization gaps and stable cross-validation scores. Two classification models were developed and evaluated: Logistic Regression (61.97% accuracy) served as a baseline, while Random Forest (68.91% accuracy) emerged as the superior model, superior across all evaluation metrics including precision, recall, F1-score, and AUC-ROC. K-Means clustering identified three distinct market segments with varying success profiles.Video presence emerged as the single strongest success predictor followed by goal structure and project description quality.

HERE IS THE LINK OF MY DATASET THAT I APPLIED IN MY PROJECT:

https://docs.google.com/spreadsheets/d/1f5FJV2HGck-1FOIzgGZFH4rSn4IjZIaE/edit?usp=sharing&ouid=100839041109783019599&rtpof=true&sd=true

CLASSIFICATION RESULTS (Task 1):

• Random Forest model: 68.83% test accuracy
• Strong recall: 87.99% (identifies 88 of 100 successful projects)
• AUC-ROC: 0.6847 (superior to baseline LR: 0.5874)
• Key success drivers: video presence (17.91%), goal amount (17.22%), description quality (20%)
• Data leakage prevention: only pre-launch features used

CLUSTERING RESULTS (Task 2):

• Optimal K=3 segments identified using Silhouette, Davies-Bouldin, and Elbow methods
• Cluster 0 (64.3%): Standard projects, 60% success rate
• Cluster 1 (28.9%): Lean projects, 61% success rate
• Cluster 2 (6.7%): Niche micro-projects, 70% success rate
• Key insight: Smallest, least-documented cluster has highest success rate

STRATEGIC VALUE: The combined framework enables project-level success prediction and segment-specific resource allocation for Kickstarter stakeholders.

DATA PREPARATION & TRANSFORMATION PIPELINE

Data preparation followed industry best practices for preventing information leakage while maintaining data integrity. The pipeline involved: (1) temporal train-test splitting, (2) StandardScaler fitting on training data exclusively, (3) categorical label encoding on training data exclusively, (4) index alignment verification, (5) missing value checks, and (6) final verification

of proper feature-label pairing. All transformations were applied consistently across training and test sets using statistics and mappings learned from training data only.

TASK 1: CLASSIFICATION MODEL

Model Selection & Performance: Two models were compared: Logistic Regression (LR) and Random Forest (RF).

MODEL COMPARISON:

	LR	RF	Improvement
Accuracy	61.93%	68.83%	+7.90%
Recall	80.10%	87.99%	+7.89%
F1-Score	0.7242	0.7788	+0.0546
AUC-ROC	0.5874	0.6847	+0.0973

SELECTED: Random Forest (RF) - superior accuracy, recall, and ranking ability.

DATA & FEATURES:

- 17 carefully engineered features available at project launch
- Explicitly excluded: pledged, backers_count, usd_pledged (post-launch data = leakage)
- Features include: goal amount, description quality, video presence, launch timing, category
- Standardized using StandardScaler for preprocessing

PERFORMANCE ANALYSIS:

Test Accuracy: 68.83% | Precision: 69.86% | Recall: 87.99% | F1-Score: 0.7788

The model identifies 28,805 of 32,736 successful projects correctly (87.99% recall), making it highly valuable for identifying high-potential projects. The 37% failure detection rate reflects the diverse nature of failed projects.

FEATURE IMPORTANCE (Random Forest):

1. Video presence: 17.91% - strongest predictor of success
2. Goal amount (log-scaled): 17.75% - non-linear relationship
3. Raw goal amount: 17.22% - critical success factor
4. Category: 8.31% - different categories have inherent success rates
5. Launch month: 6.37% - seasonal patterns matter

LOGISTIC REGRESSION INSIGHTS:

- has_blurb (+3.11): Description presence is critical
- goal (-5.61): High goals dramatically reduce success probability
- has_video (+0.63): Video content boosts success
- launched_month (+0.56): Timing matters

CONCLUSION: RF model's 68.83% accuracy with 87.99% recall makes it production-ready for identifying promising projects and guiding creators on success factors. Random Forest emerges as the clear winner, achieving superior performance across all six metrics. The 6.94 percentage point accuracy improvement (68.91% vs 61.97%) combined with significantly higher recall

(88.06% vs 80.14%) and F1-score (0.7794 vs 0.7244) justifies Random Forest as the recommended production model. While generalization gap is slightly larger (8.87% vs 6.62%), this difference is within acceptable bounds for ensemble methods and does not indicate overfitting concerns.

TASK 2: CLUSTERING MODEL

Methodology & Cluster Selection

K-Means clustering was applied to identify project segments. Optimal K selection tested 2-10:

K	Silhouette	Davies-Bouldin	Decision
2	0.4814	0.9078	Highest score but less insightful
3	0.4444	1.0465	SELECTED - best balance
4	0.4067	1.1914	Quality degradation begins
5+	0.3783-0.1870	1.2572-1.6796	Poor clustering

K=3 JUSTIFICATION:

Silhouette Score 0.4444 (>0.4 = good clustering)

Elbow method shows natural bend at K=3

Interpretabilit WE CHOOSE three distinct, actionable segments

CLUSTER CHARACTERISTICS:

CLUSTER 0 - "Standard Professional Projects" (64.3% of market):

• Profile: Well-documented, moderate funding goals, professional presentation

• Success Rate: 56.23% (training), 60.08% (test), Size: 135,066 training projects

• Insight: Baseline segment representing typical successful projects

• Recommendation: Standard support program with documentation templates

CLUSTER 1 - "Lean Efficient Projects" (28.9% of market):

• Profile: Minimal descriptions, shorter names, below-average goals

• Success Rate: 53.83% (training), 61.44% (test) , Size: 60,700 training projects

• Insight: Experienced creators succeed despite minimal documentation

• Recommendation: Create experienced creator fast-track program

CLUSTER 2 - "Niche Micro-Projects" (6.7% of market):

• Profile: Extremely low goals, minimal documentation, category-specialized, no video

• Success Rate: 53.34% (training), 69.61% (test) , Size: 14,165 training projects

• Training-to-Test Improvement: +16.27% (largest of all clusters)

• THE CLUSTER 2: Smallest and least-resourced cluster achieves highest success

WHY CLUSTER 2 SUCCEEDS:

1. Realistic goal expectations → achievable targets → higher success probability

2. Passionate creators → genuine commitment to projects

3. Niche communities → reliable internal support

4. Selection bias → ambitious creators pursuing modest goals are serious

CONCLUSION: Cluster 2 represents genuine passion projects with realistic expectations, demonstrating the highest ROI for platform investment despite being the smallest segment.

BUSINESS IMPACT & STRATEGIC RECOMMENDATIONS: Resource allocation should be segmented by creator type: Cluster 0 (64%, first-time creators) receives 40% of resources with templates for 60% success; Cluster 1 (29%, experienced creators) receives 30% with fast-track programs for 61% success; Cluster 2 (7%, niche creators) receives 30% focused on category support to push success toward 75%. Each pathway provides appropriate support based on creator experience and project characteristics. Creators should follow their optimal pathway: Cluster 0 (first-time, seeking safety, 60% success), Cluster 1 (experienced, preferring efficiency, 61% success), or Cluster 2 (niche experts, realistic goals, 70% success highest). This reveals that

in specialized domains, genuine commitment and realistic funding matter more than marketing materials. Cluster 2 achieves 70% success with only 7% of projects the best ROI opportunity. A 2-3% platform-wide improvement across 100,000 annual projects yields 2,000-3,000 additional successful projects. Category-specific programs are more cost-efficient than generic features, making this targeted approach essential for maximizing creator success and platform impact.