INSY662 Individual Project Report

## Introduction

This analysis employs classification and clustering techniques to predict campaign success and understand project characteristics that drive positive outcomes. The insights derived aim to help both platform managers and project creators optimize campaign performance through data-driven decision making.

## Classification Task

The primary goal of the classification task is to predict whether a Kickstarter project will be "successful" or "failed" based on data available at the time of project launch. The dataset has been filtered to include only projects with "successful" or "failed" states and eliminated features unavailable at launch time, such as pledged amounts and backer counts.

Feature engineering focused on transforming available data into meaningful predictors. Financial metrics underwent USD conversion using static rates, followed by log transformation to address significant skewness in distributions. Temporal features are developed by calculating launch duration and creation-to-launch time periods. Categorical variables including project categories and countries were one-hot encoded to maintain interpretability while maximizing predictive power.

The feature selection process was guided by comprehensive statistical analysis of available pre-launch data. Chi-square tests revealed that category (Cramer's V = 0.594) has the strongest association with success outcomes. The goal amount emerged as the most important predictor, showing distinct distribution patterns between successful and failed projects through ANOVA testing ($p = 0.0007$). Media-related features, including video presence (Cramer's V = 0.136) and feature images (Cramer's V = 0.158), demonstrated significant relationships with

project outcomes. Project description characteristics, measured through name and blurb length, showed modest but statistically significant effects (eta-squared = 0.009 and 0.004 respectively). Additionally, temporal features are also being added for their good performance in terms of feature importance, including launch duration and creation-to-launch period. All selected features exhibited low multicollinearity, ensuring robust model performance.

After evaluating multiple algorithms including Random Forest, Logistic Regression, and Gradient Boosting through cross-validation, the Gradient Boosting classifier emerged as optimal. Grid search optimization identified ideal parameters (learning rate: 0.1, max_depth: 5, n_estimators: 200), striking an effective balance between model complexity and generalization ability. The final model achieved 80% accuracy on the test set, with particularly strong performance in identifying successful campaigns (85% recall, 81% precision).

**Clustering Task**

The clustering task aimed to group Kickstarter projects into clusters to uncover shared characteristics and derive actionable insights. This analysis provides a deeper understanding of project attributes that influence success.

The feature selection process focused on attributes that meaningfully differentiate projects: converted financial metrics, media presence indicators, and promotional features. Additional features including funding ratios and log-transformed financial variables are included to capture project characteristics more effectively. The clustering process utilized the K-Means algorithm, with k=5 clusters providing optimal separation based on silhouette score analysis. The resulting segments reveal distinct project archetypes with varying success patterns:

The Standard Success Track (32.30% of projects) demonstrates that well-prepared campaigns with moderate goals ($7,500 median) and strong video presence can succeed without

special platform features. The Premium Track (13.37%) achieves a 91.21% success rate through comprehensive platform utilization and staff recognition. The Risk Zone (25.05%) identifies critical failure patterns, particularly overambitious goals ($10,000 median) without sufficient platform support. The Entry Level segment (25.94%) provides a proven pathway for new creators with modest goals ($652 median), while the Media Excellence segment (3.33%) demonstrates the power of comprehensive media strategy with a perfect success rate.

Based on the insights generated from the cluster model, recommendations are provided for both the user and Kickstarter company. Project creators should start with modest goals under $1,000 to build track records and gradually progress to higher targets. Comprehensive media presence becomes crucial for projects over $5,000, and creators should seek staff recognition for higher-goal projects. The analysis demonstrates that effective platform feature utilization significantly impacts success rates. For platform management, implementing an automated risk assessment system based on the classification model's predictions is recommended. Projects showing risk factors should receive enhanced support and guidance. The platform should develop a graduated approach to project goals and feature requirements, guiding creators through proven pathways from entry-level to premium projects.

**Conclusion**

This analysis provides a robust framework for understanding and improving Kickstarter campaign outcomes. The classification model enables proactive risk assessment, while the clustering analysis reveals strategic pathways to success. Together, these insights offer practical tools for enhancing platform effectiveness and optimizing creator outcomes. The identified patterns and success factors form a comprehensive framework for strategic decision-making in the Kickstarter ecosystem.