Cover Page

Course Title: Machine Learning
Team Members:
- Mohamed Mostafa : 23101594
- Marwan Khaled : 23101599
- Mohamed Adel : 23101899

Generated: 2025-12-19 12:42

Project: Movie Metadata Analytics for Streaming Acquisition Decisions

Section 1: Problem Domain

We model movie performance signals to support streaming acquisition decisions.
Classification target: is_high_rated derived from vote_average.
Regression target: log(1+revenue) for movies with reported revenue.

Section 2: Project Summary

Problem Domain: Streaming platform decision support using real-world movie metadata.
Dataset: Top_10000_Movies.csv (10014 rows, 13 columns).
Dirty data indicators: missing/blank text, zero revenues, skewed distributions, outliers, and semi-structured categorical genres.
ML tasks: (1) Classification of High Rated movies; (2) Regression for log-revenue.
Models: Logistic Regression, Random Forest, SVC, KNN (classification) and Linear/Ridge, Random Forest Regressor, SVR, KNN (regression).
Evaluation: Cross-validation and test metrics; scaling vs normalization comparisons included.

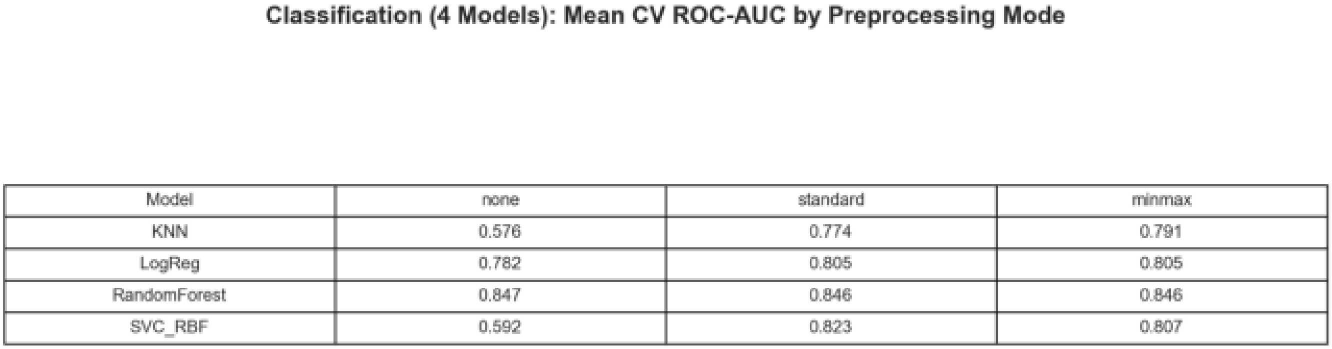
Section 3: Source Code

Source Code: This submission includes Ml_Project.ipynb and a template download script.
Preprocessing uses ColumnTransformer with robust imputation and one-hot encoding.
Models are compared under three preprocessing modes: none, StandardScaler (scaling), and MinMax (normalization).

Section 4: Visualization Snapshots

The following pages include snapshots of model comparisons, EDA, required visualizations, and model evaluation plots.

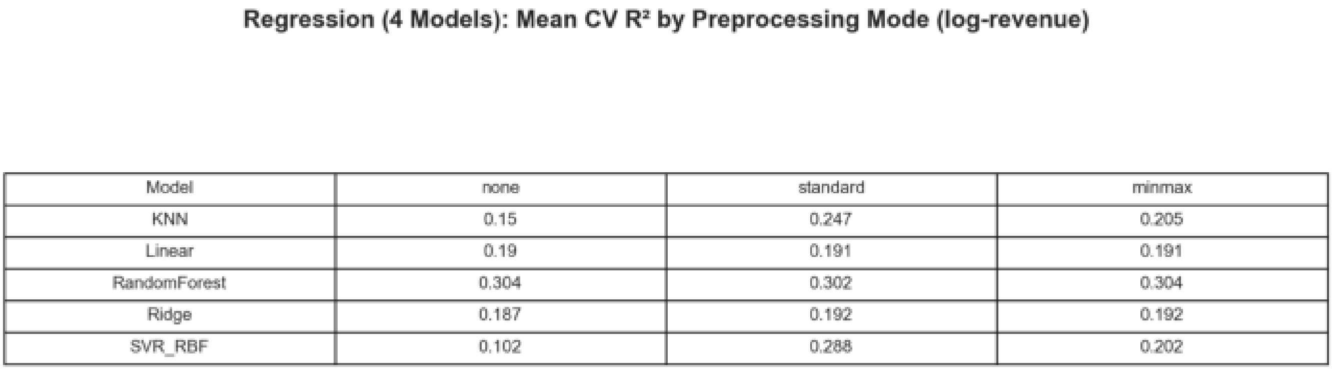
Model Comparison (Classification Table)



Classification (4 Models): Mean CV ROC-AUC by Preprocessing Mode

Model	none	standard	minmax
KNN	0.576	0.774	0.791
LogReg	0.782	0.805	0.805
RandomForest	0.847	0.846	0.846
SVC_RBF	0.592	0.823	0.807

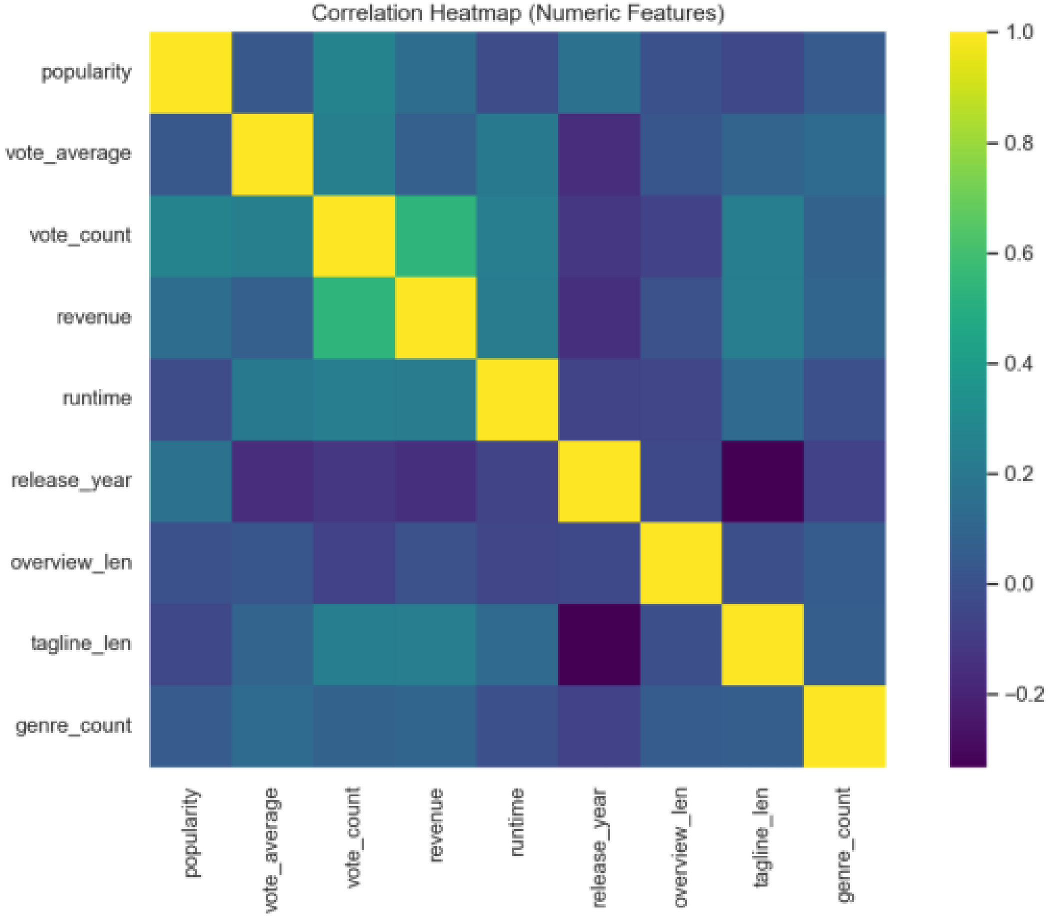
Model Comparison (Regression Table)



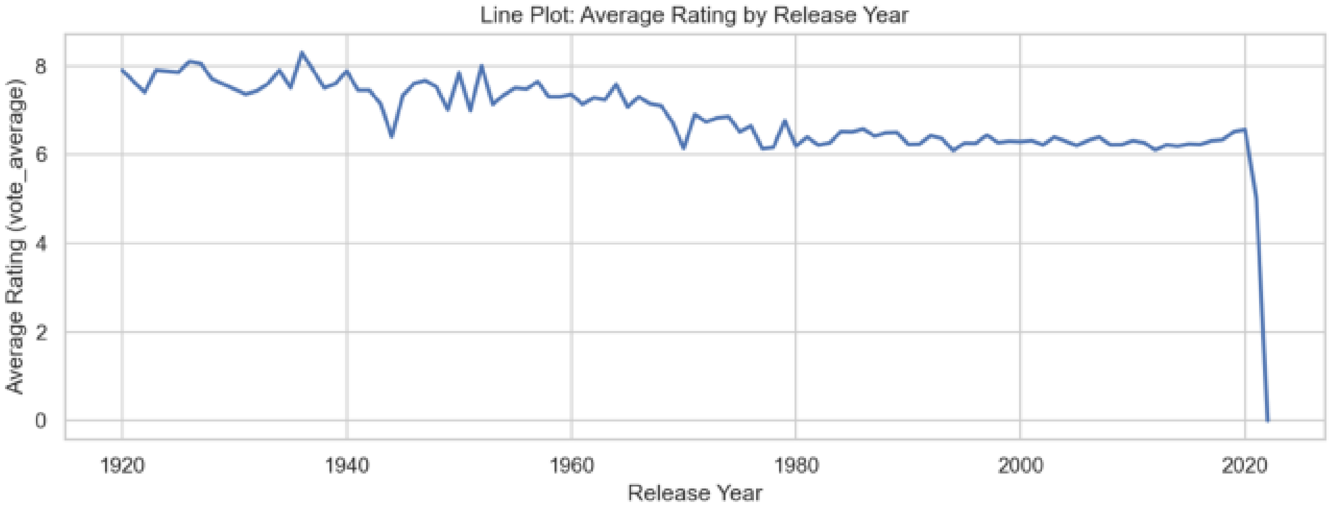
Regression (4 Models): Mean CV R² by Preprocessing Mode (log-revenue)

Model	none	standard	minmax
KNN	0.15	0.247	0.205
Linear	0.19	0.191	0.191
RandomForest	0.304	0.302	0.304
Ridge	0.187	0.192	0.192
SVR_RBF	0.102	0.288	0.202

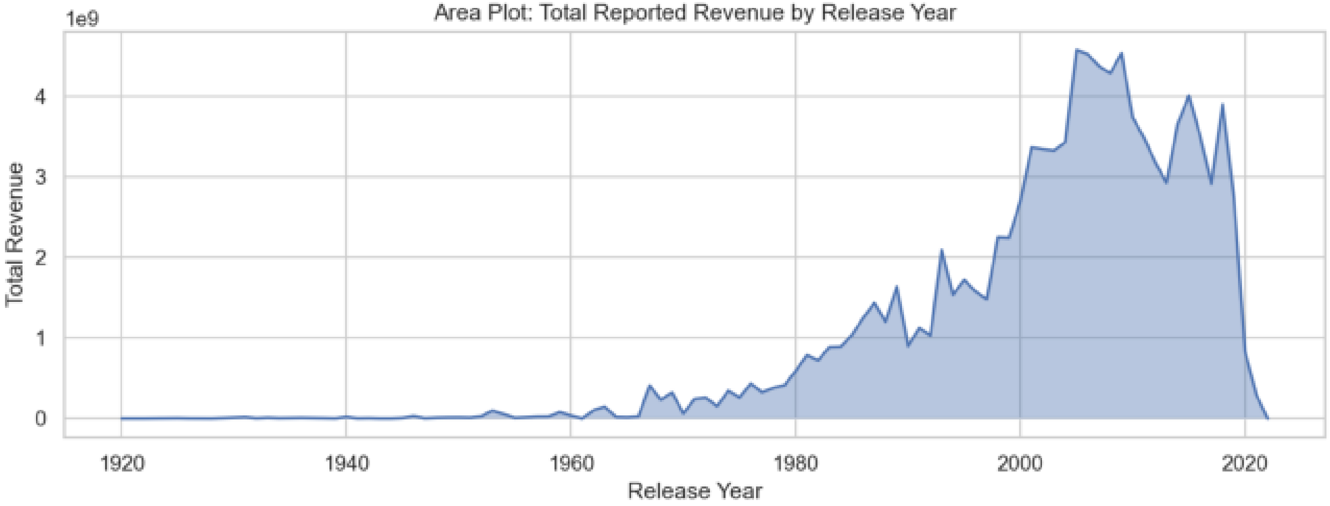
EDA: Correlation Heatmap



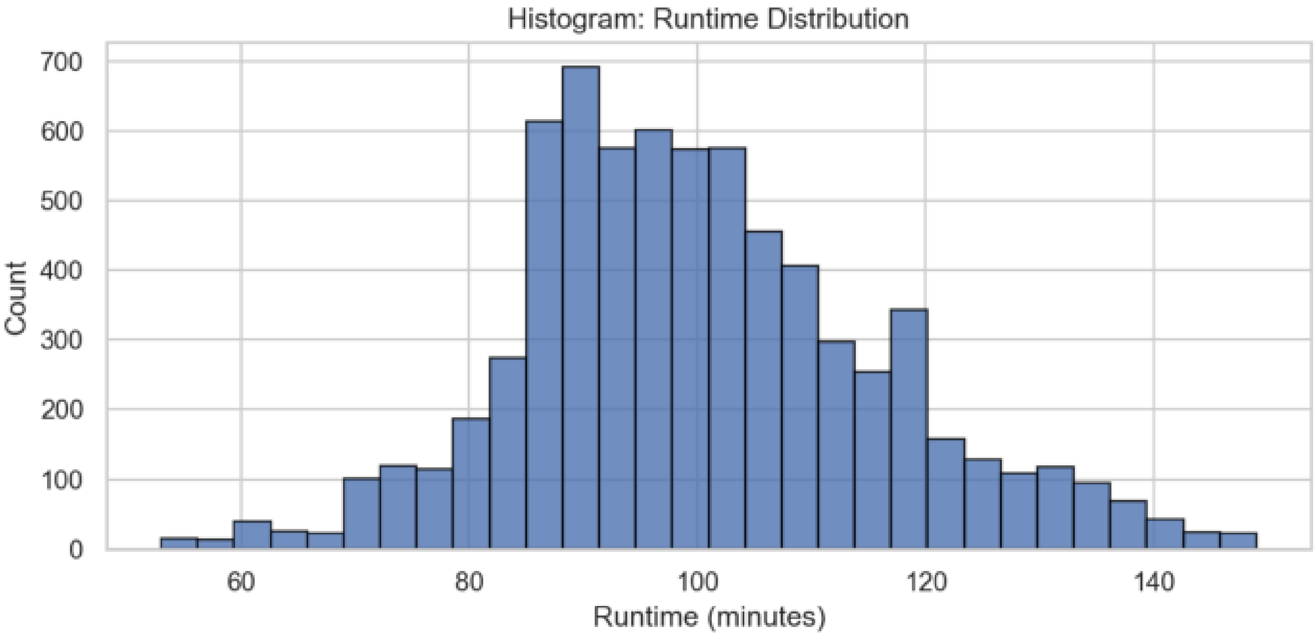
Visualization 1: Line Plot



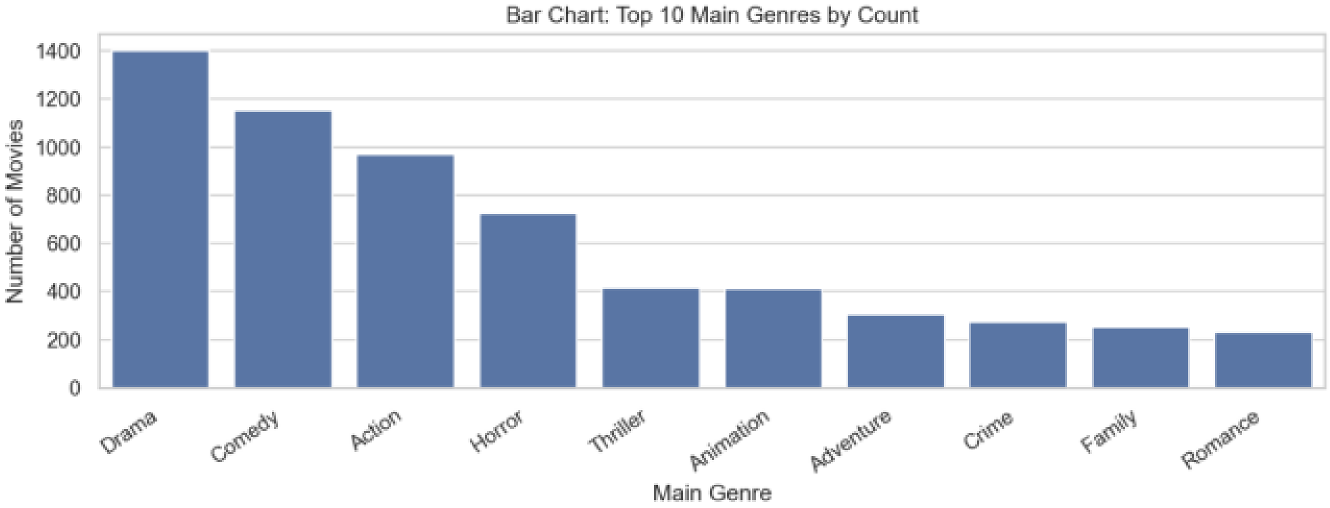
Visualization 2: Area Plot



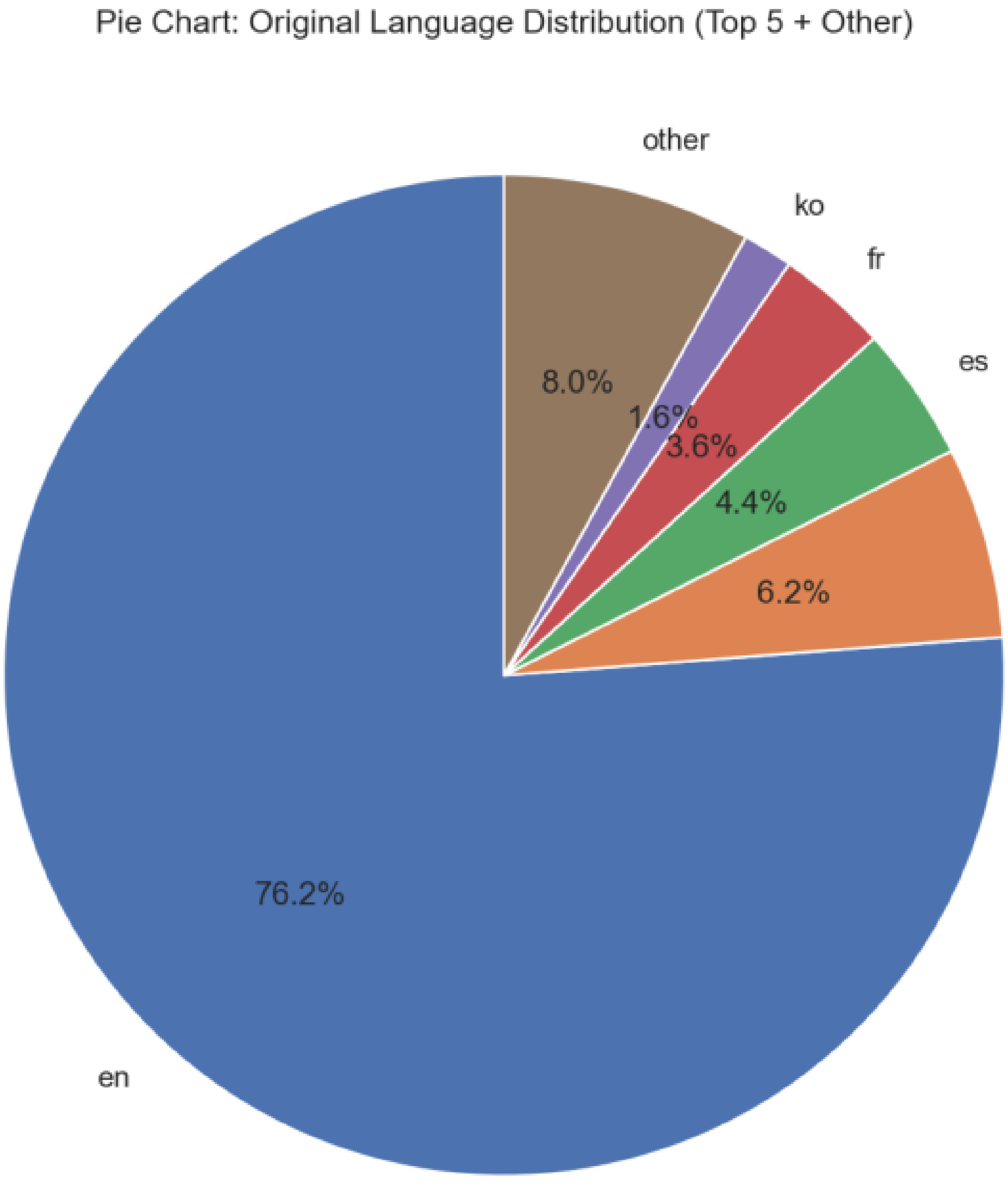
Visualization 3: Histogram



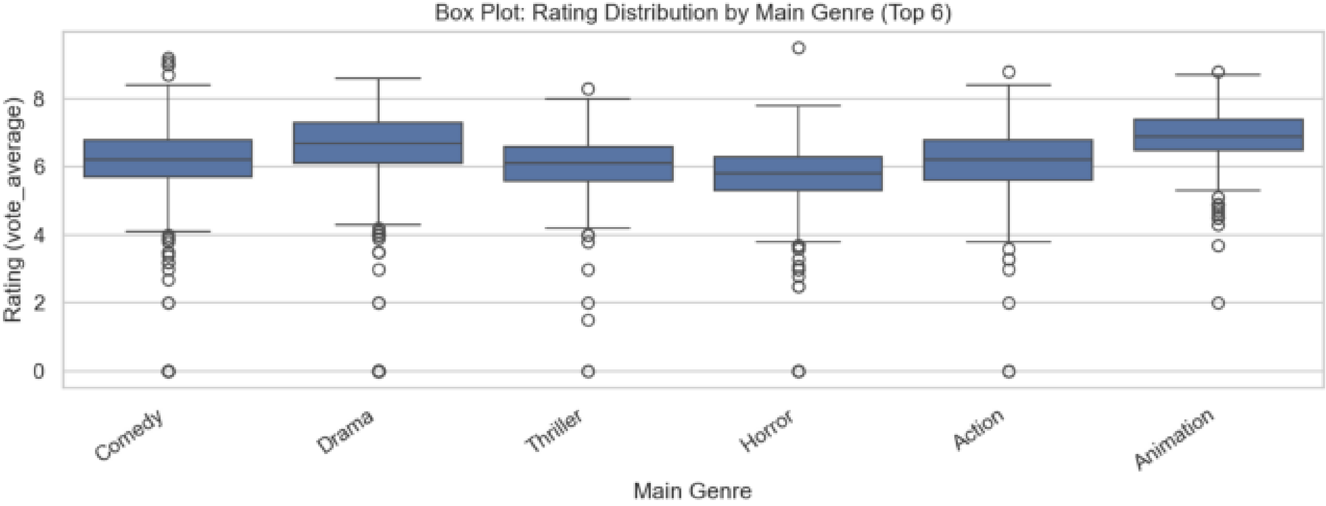
Visualization 4: Bar Chart



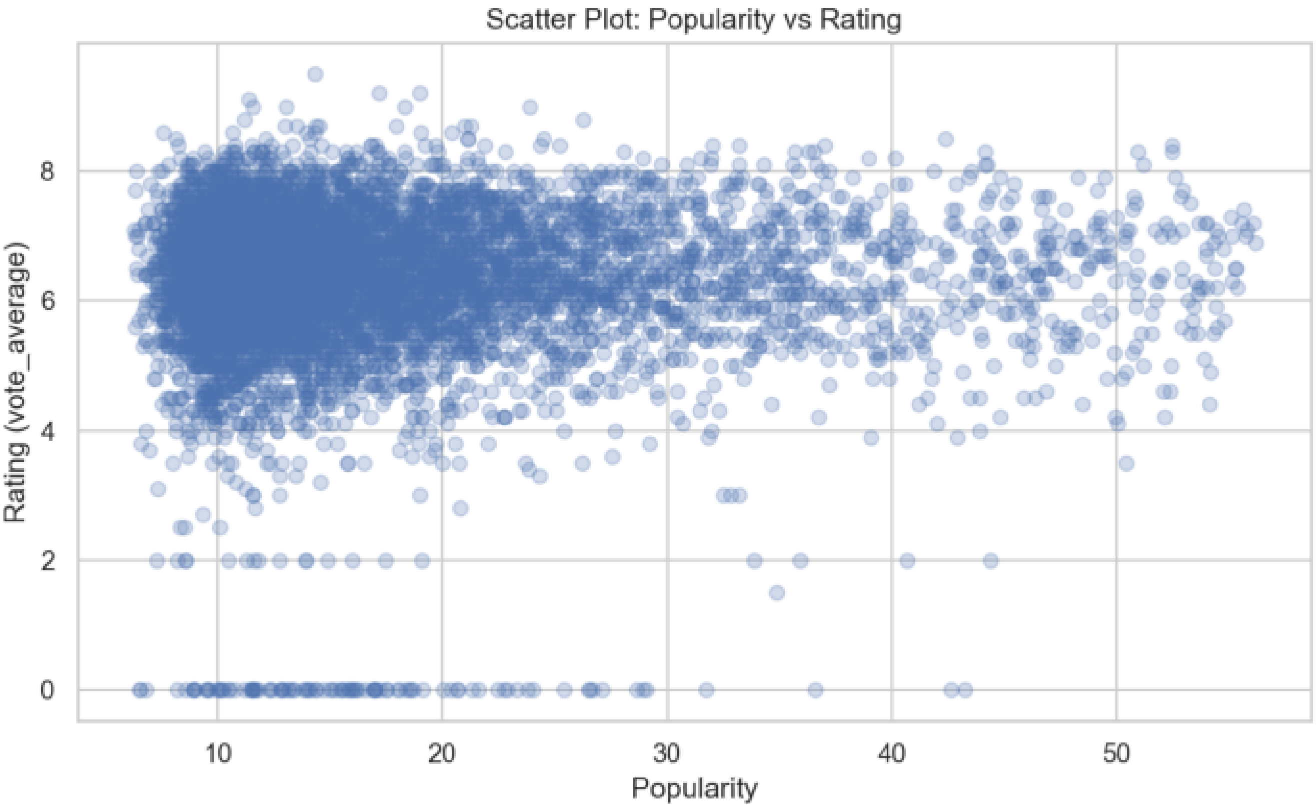
Visualization 5: Pie Chart



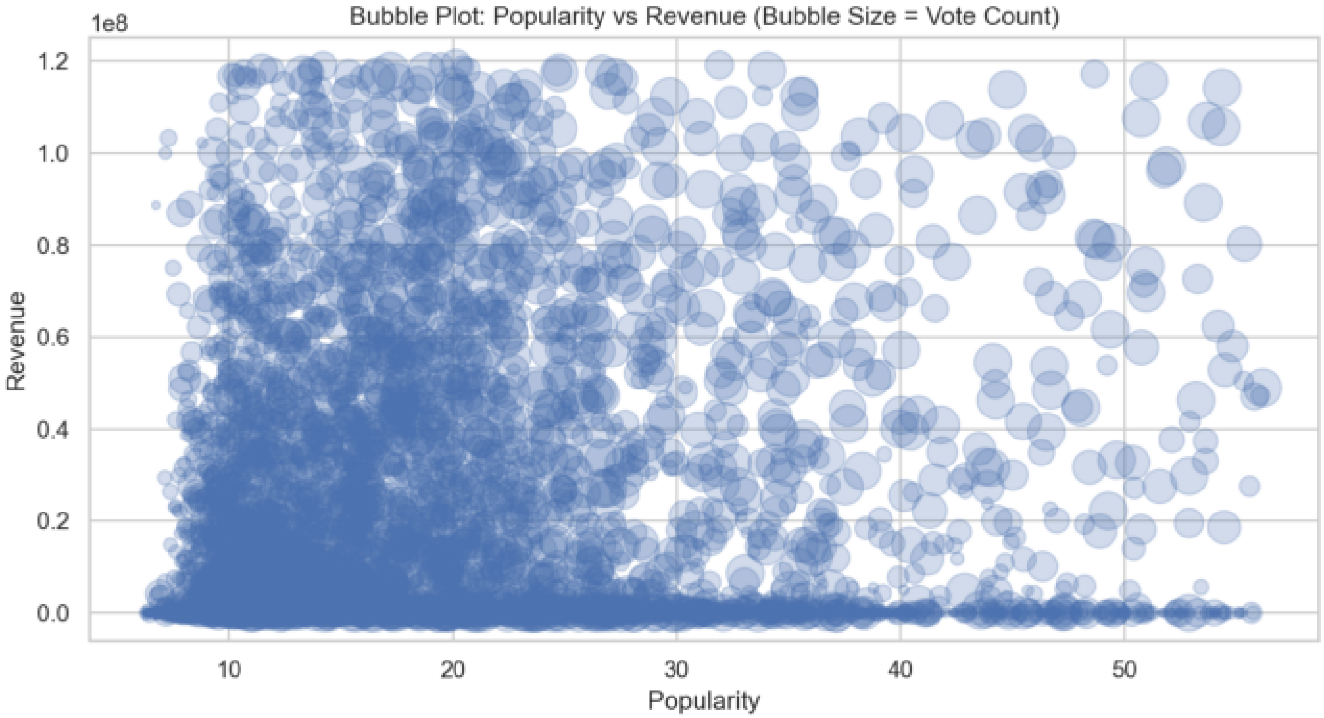
Visualization 6: Box Plot



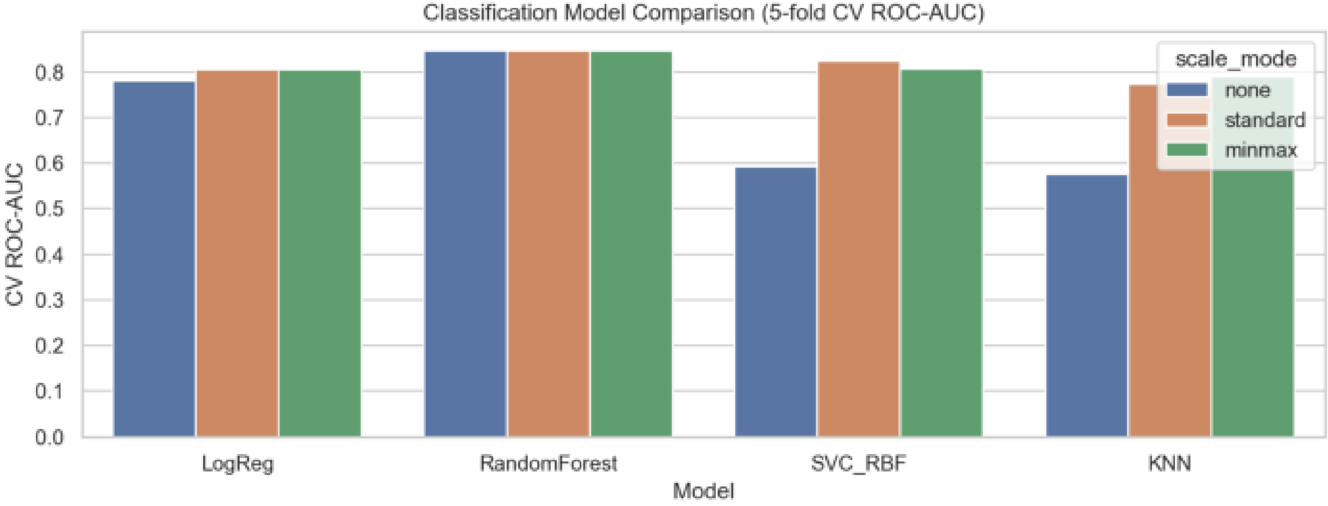
Visualization 7: Scatter Plot



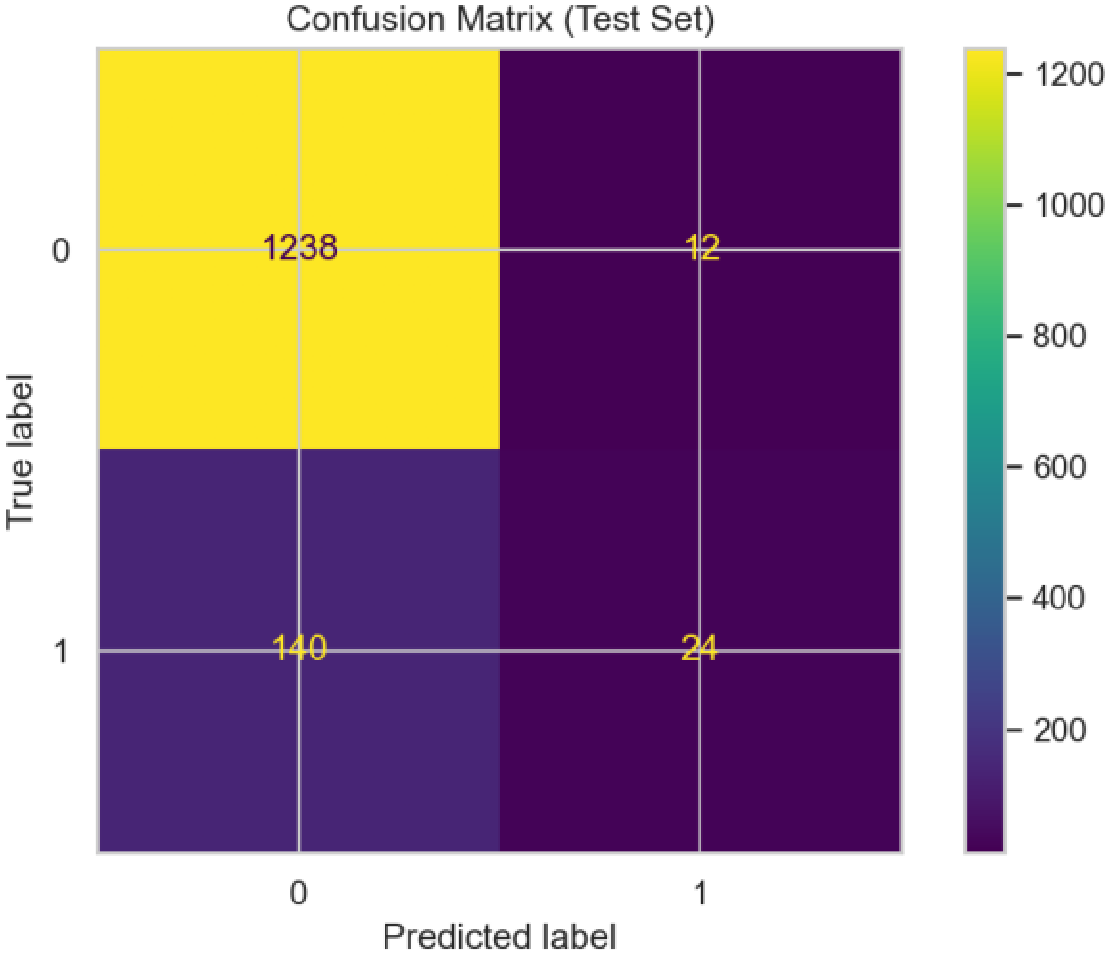
Visualization 8: Bubble Plot



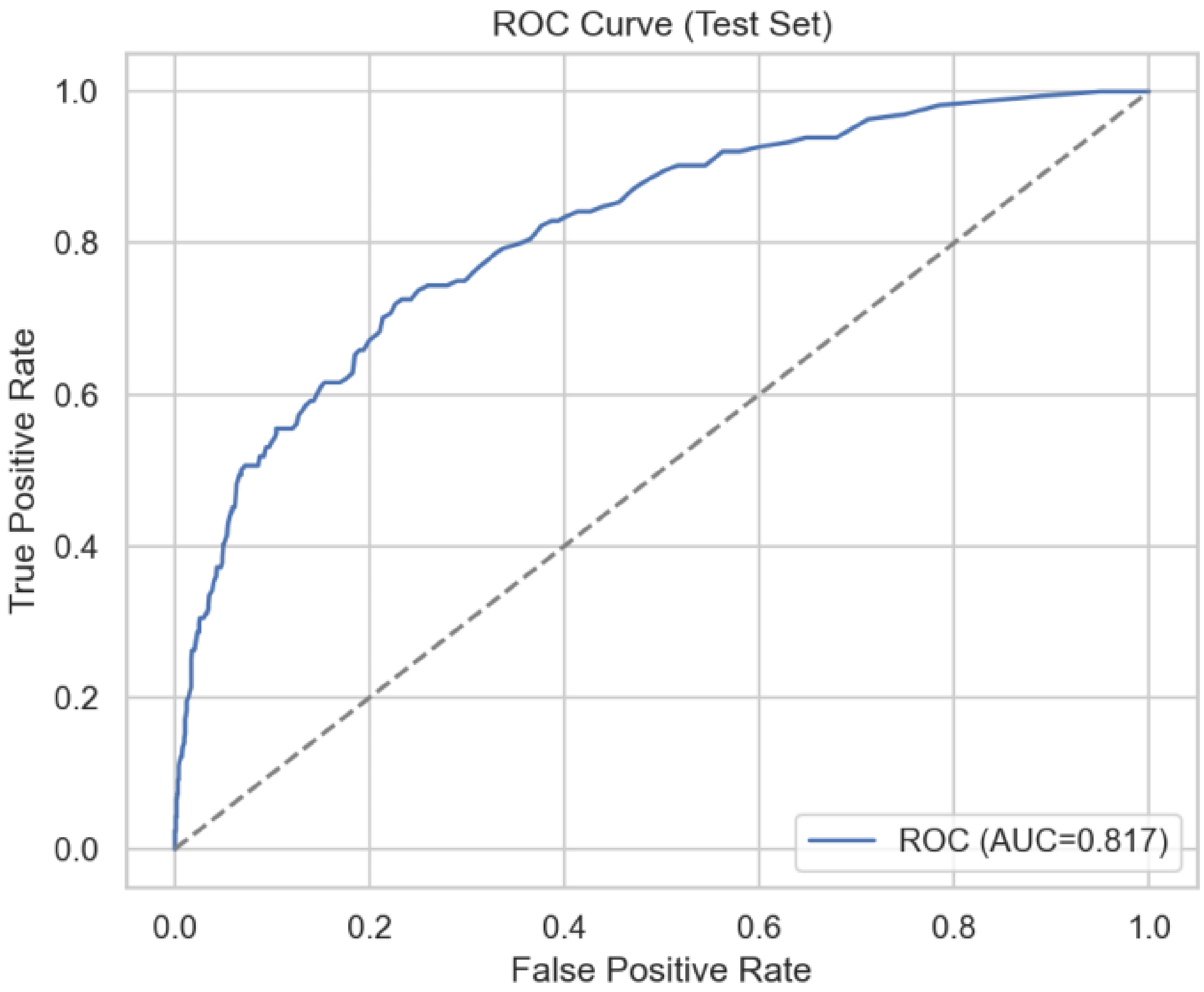
Classification: CV ROC-AUC Comparison



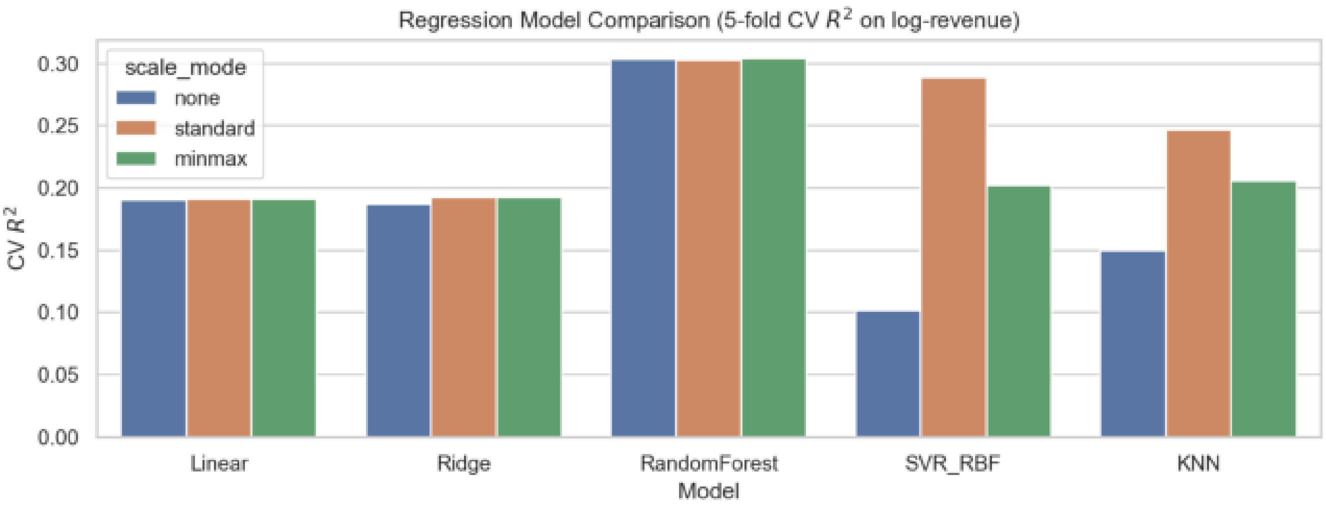
Classification: Confusion Matrix



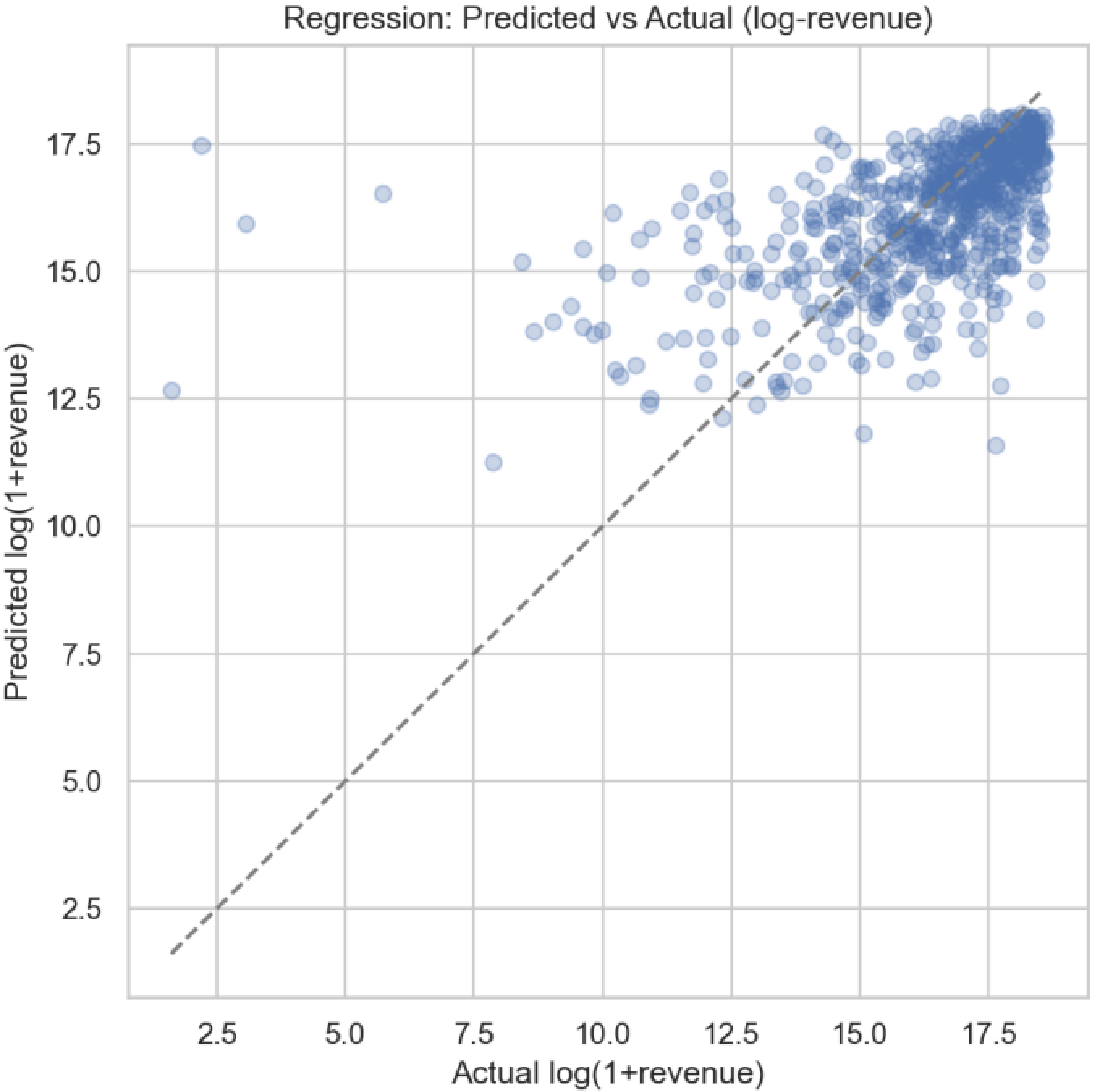
Classification: ROC Curve



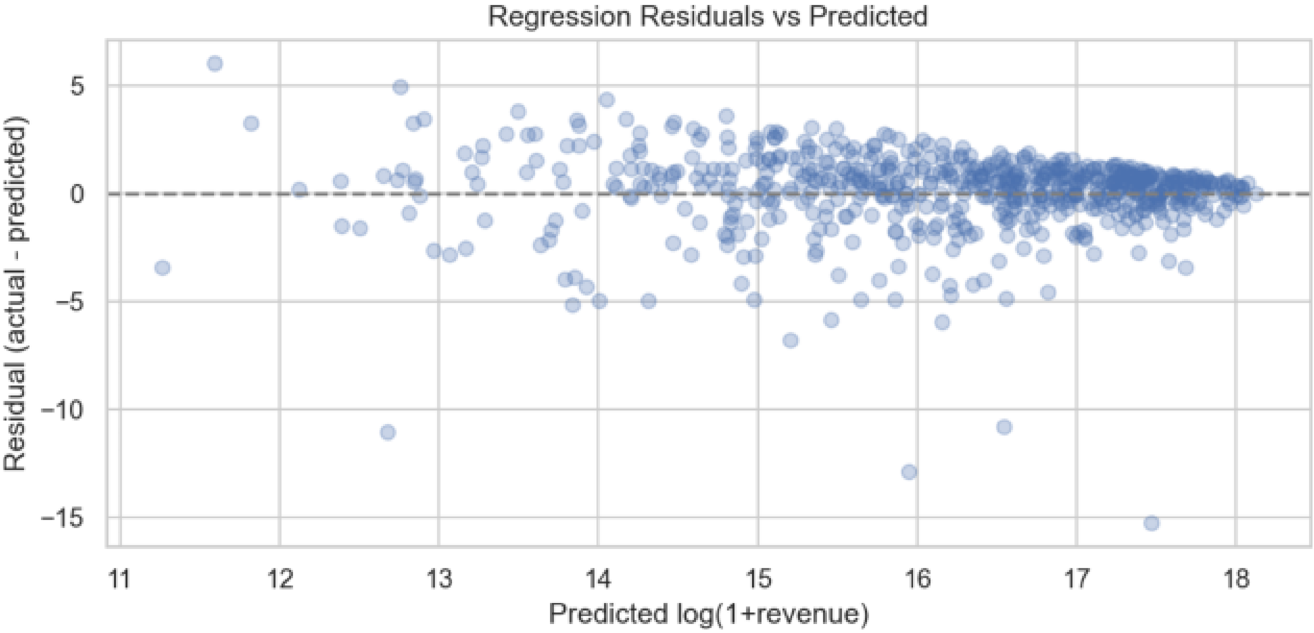
Regression: CV R2 Comparison



Regression: Predicted vs Actual



Regression: Residuals


Regression Residuals vs Predicted
Residual (actual - predicted)
Predicted log(1+revenue)