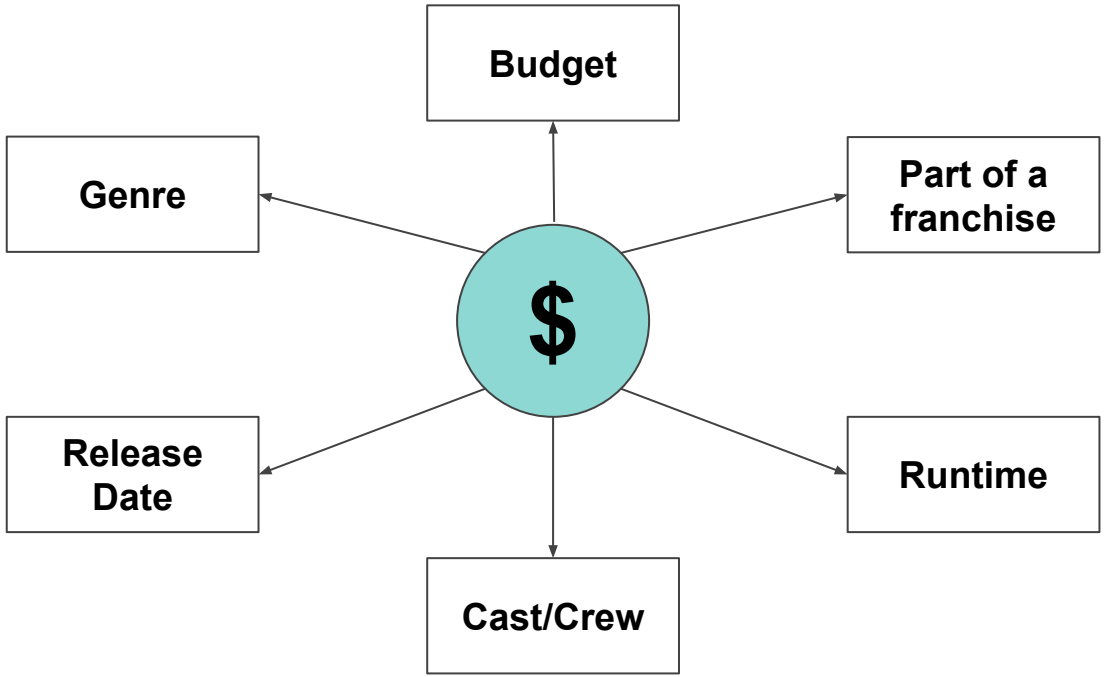
Using Machine Learning to Predict Movie Success

Dr. Deepti Srivastava Tilly

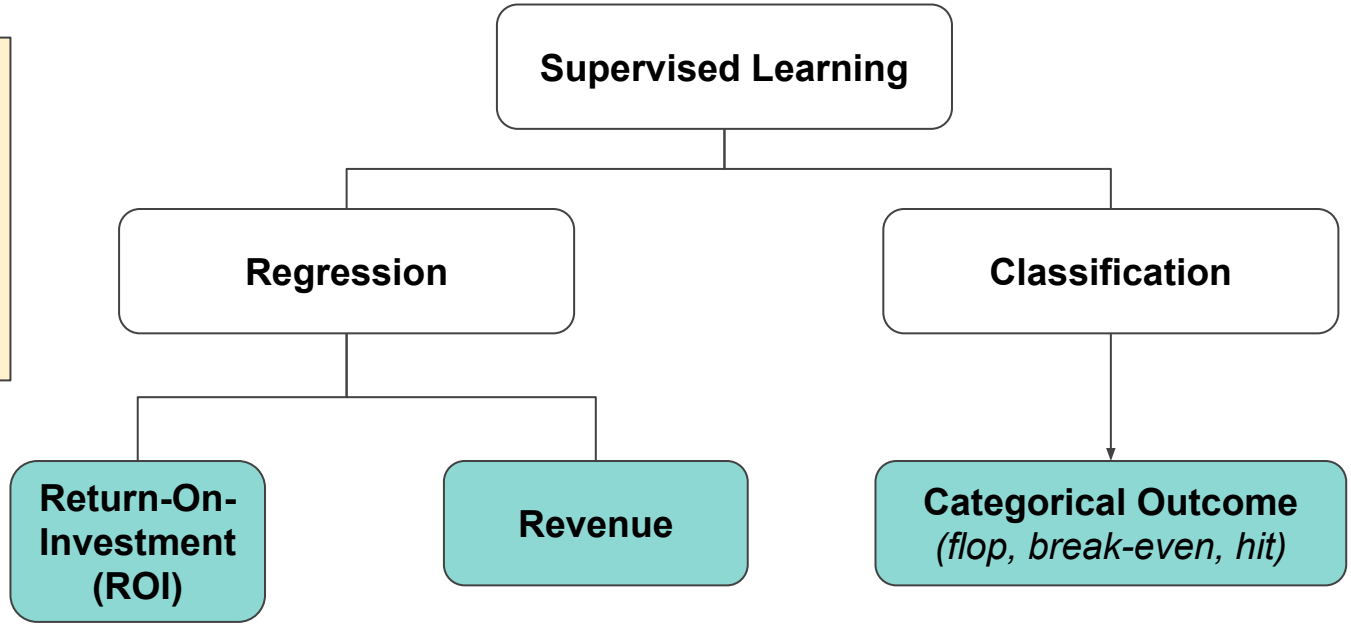
Understanding the key drivers of movie success can be useful in making key business decisions.



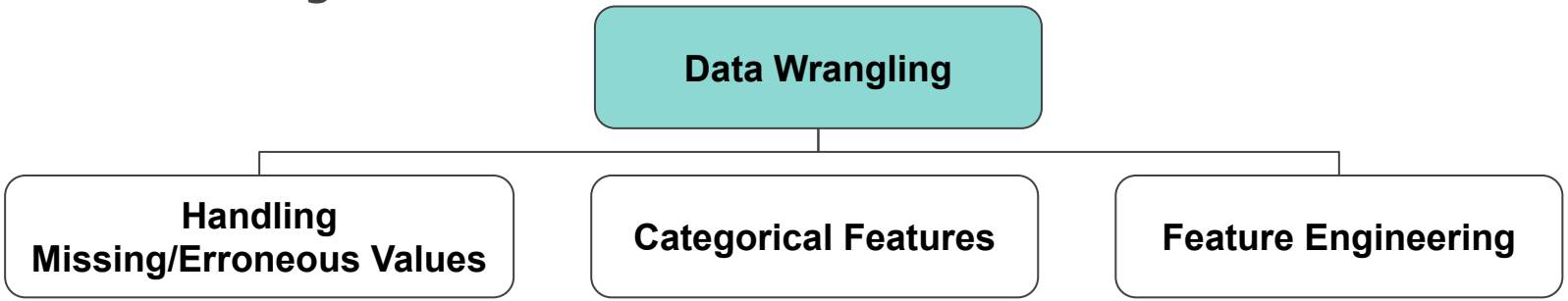
Using Machine Learning, studios can help mitigate potential losses and investment risk by better analyzing and predicting movie success.

Using data from The Movie Database, supervised learning can be used to predict movie success in several different ways.

Prediction of ROI (calculated as revenue/budget) can provide more actionable analysis for producers and investors



Data wrangling involved handling missing values/erroneous data and feature engineering to prepare the data for model building.

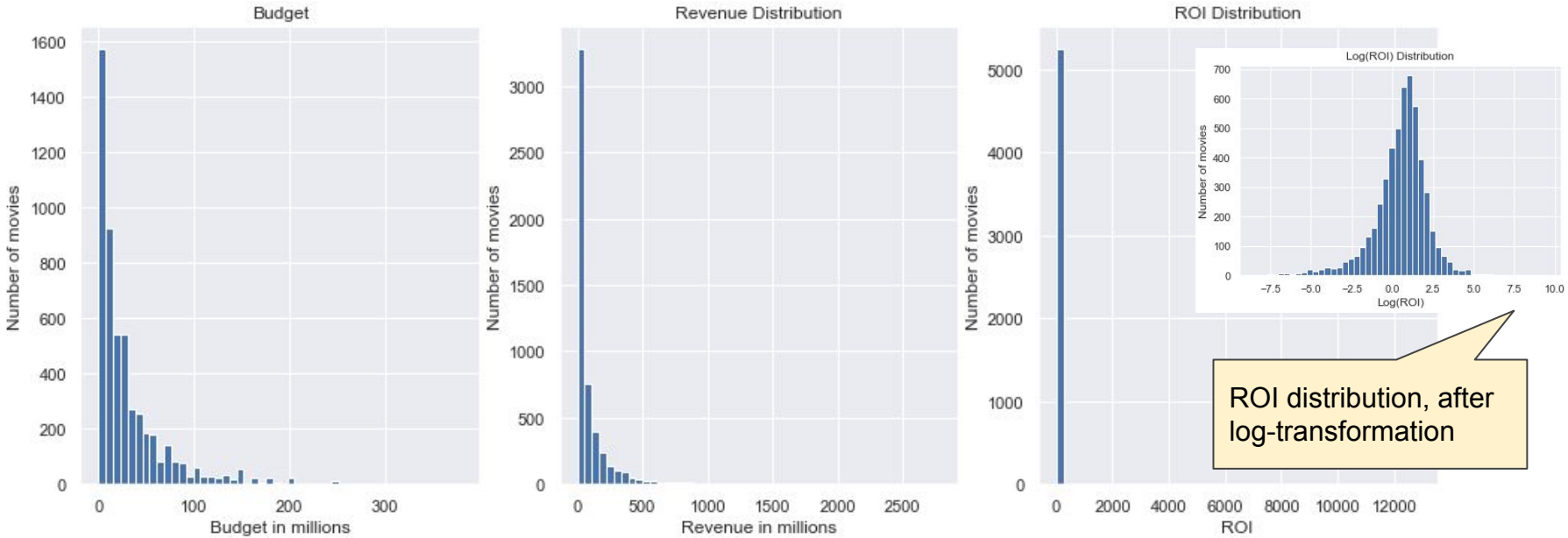


- ~75% of the movies showed budget and revenue values of 0 and single digits, indicative of missing values and erroneous data capture.
- Movies with a budget below $10,000 and a revenue below $1500 were dropped.
- Missing values from other columns were dropped.

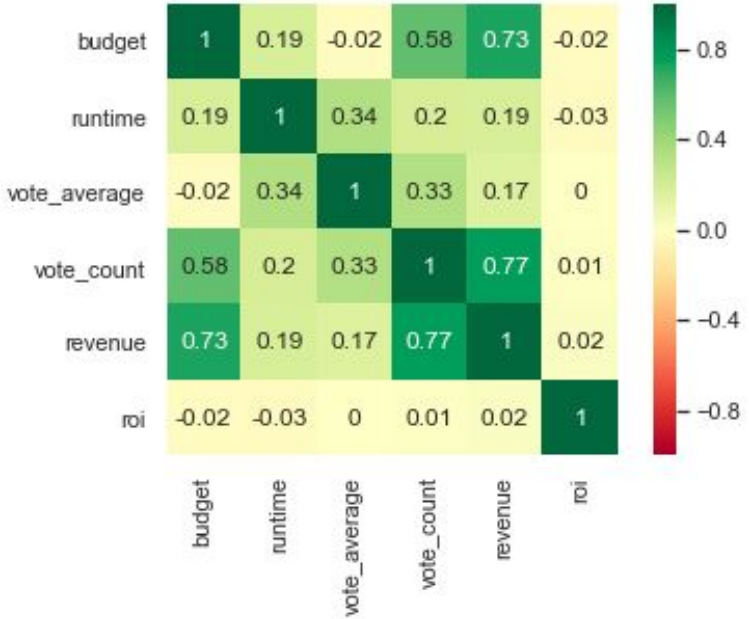
- Categorical features such as genre and release season were converted into dummy variables using one hot encoding.
- Belongs_to_collection was converted into a binary variable.

- Using top grossing actor and director lists from IMDB, 2 new binary features were created to identify if a movie had a top actor or top director.
- Release dates were converted to release seasons.
- Categorical Outcome was calculated based on the ratio of revenue to budget.

EDA showed the distribution for ROI to be extremely right-skewed, indicating that a log transformation might be needed for ML models.



The correlation matrix showed budget and vote_count to have the highest linear correlation with revenue, while VIF analysis showed signs of multicollinearity within the data.



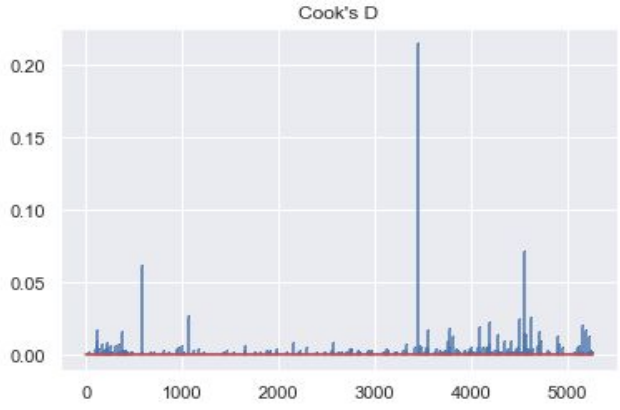
	VIF Factor	features
2	38.221020	vote_average
1	36.378651	runtime
0	3.405193	budget
10	3.131515	drama
3	2.413300	vote_count

- Vote_average and runtime to have high inflation factors, indicating that they are highly correlated with each other.
- In order to avoid problems of multicollinearity, vote_average was removed as a feature.

All numerical features (budget, runtime, vote_average, and vote_count) were scaled to have a minimum value of 0 and maximum value of 1 prior to correlation analysis.

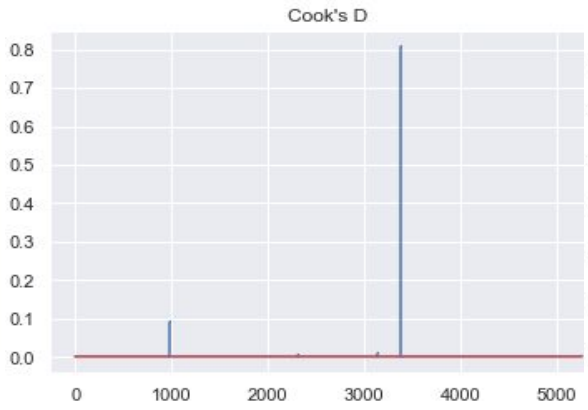
Outliers and high-influence points were identified using Cook's Distance and removed from the dataset to improve model performance.

Revenue



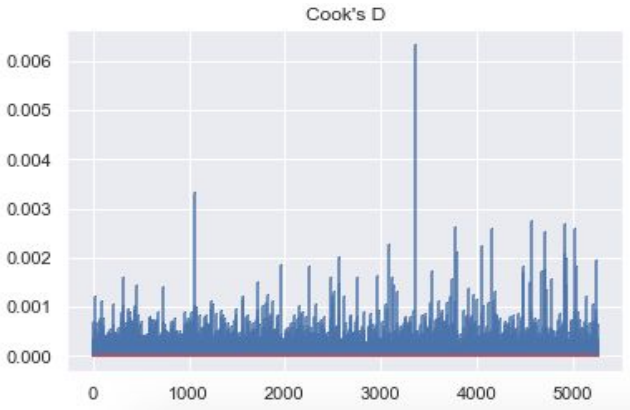
Avatar was identified as a high influence point for revenue and removed from the regression model data.

ROI



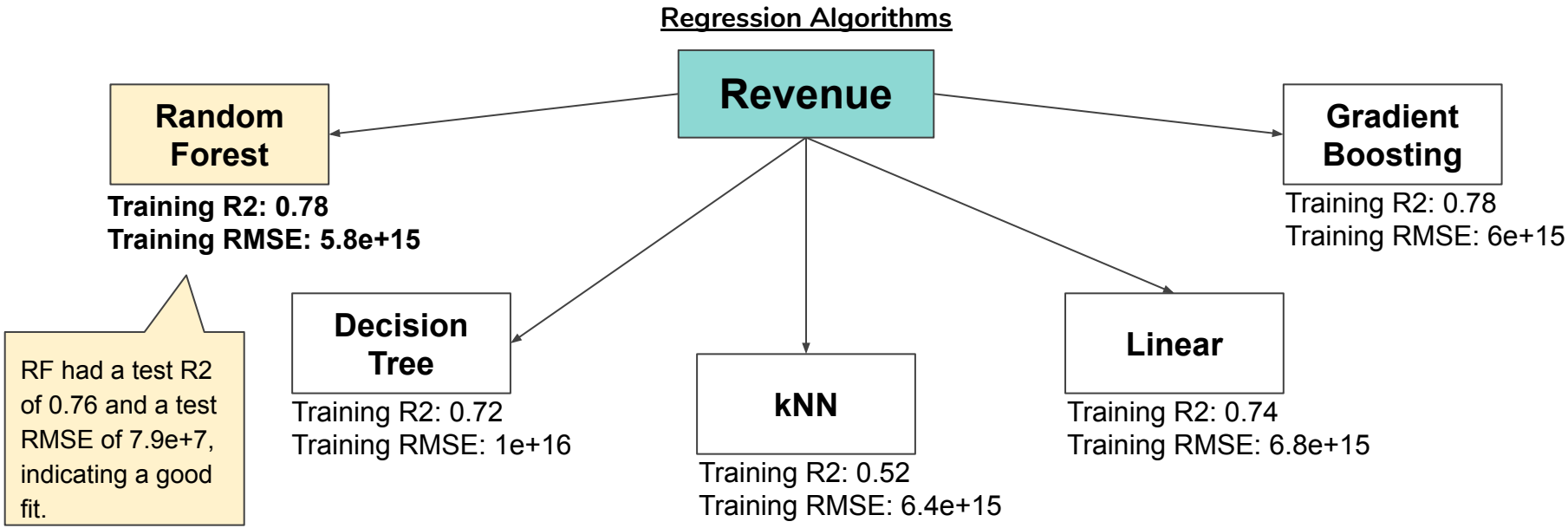
Paranormal Activity and Blair Witch's Project were identified as high influence points for ROI

Categorical Outcome



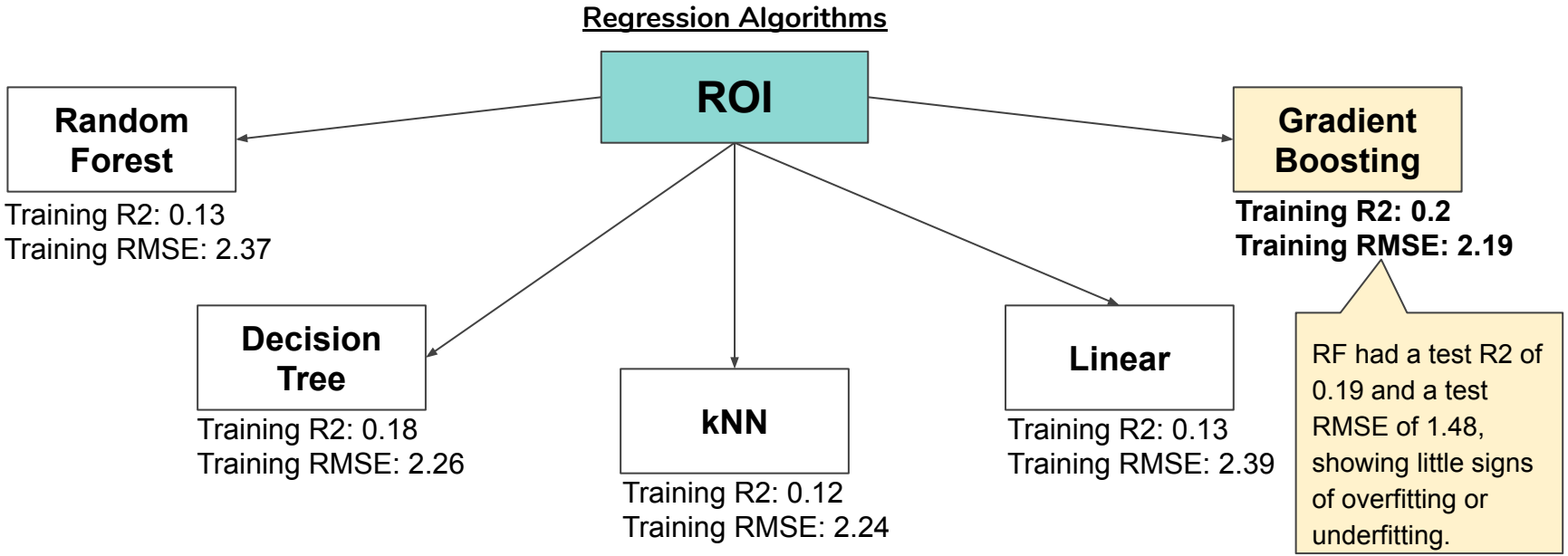
No outliers were detected for the prediction of categorical outcome.

For revenue prediction, Random Forest was found to have the best model performance based on the R^2 and RMSE values on the training set.



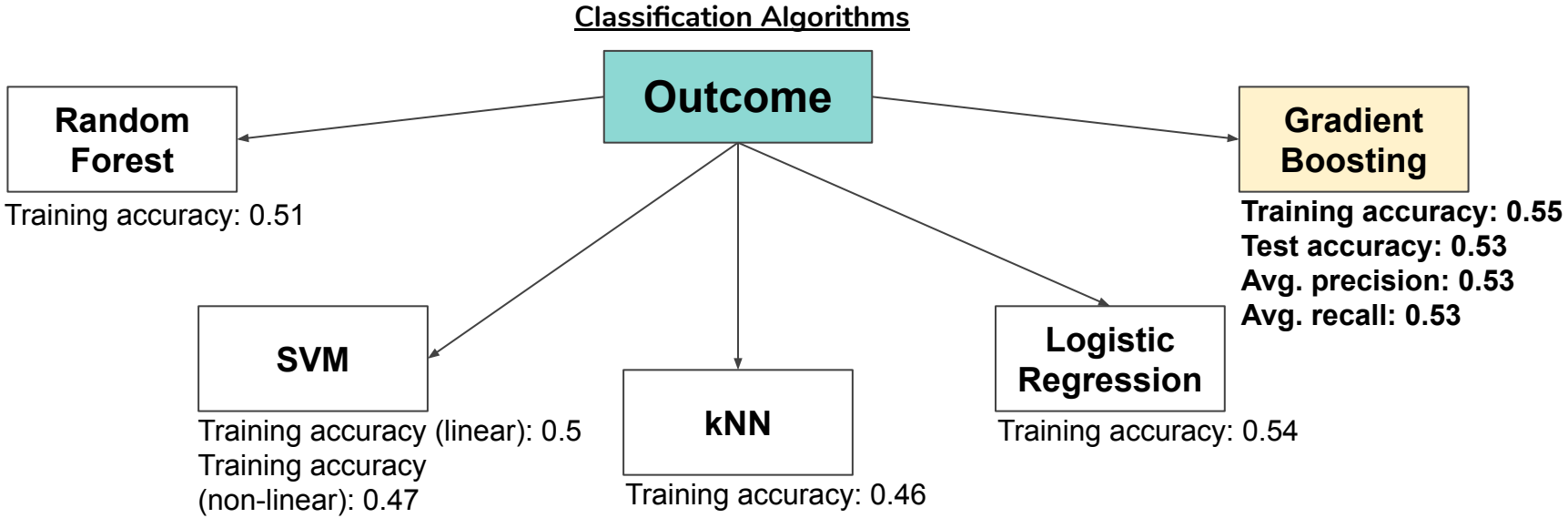
Feature importance was assessed based on Gini Importance, with all features ultimately being deemed important and retained in the final model for hyperparameter tuning.

While log-transformation significantly improved model performance for ROI, even the best-performing model (Gradient Boosting) yielded low R2 values of ~0.2.



Feature importance was assessed based on Gini Importance, with all features ultimately being deemed important and retained in the final model for hyperparameter tuning.

Using classification to determine 1 of 3 categorical outcomes resulted in better performance than ROI predictions, with GB model accuracy of ~0.55 and similar precision/recall scores.



Feature importance was assessed based on Gini Importance, with all features ultimately being deemed important and retained in the final model for hyperparameter tuning.

In order to improve model explainability for revenue and ROI, Shapley values are incorporated to better understand the contribution of features leading to individual predictions.

Shapley Value Analysis for Revenue Prediction for Prometheus (2012)

The above plot shows that:

- Belonging to a collection, having a top director, having a mystery genre and having certain runtime and vote_count values increased Prometheus' revenue prediction to above the base value.
- Meanwhile, being classified as the science fiction genre drove the revenue prediction lower.