

JANE STREET GROUP RESPONDER 6 PREDICTION

IST 718 – Big Data Analytics

Nikolay Yurashku, Benjamin Tisinger, Kéli
Davis



Introduction

Financial markets are notoriously complex and fiercely competitive. According to economic theory, any profit opportunity created by arbitrage is likely to be discovered quickly and eroded to zero as more market participants rush in to exploit it. In an era of unprecedented access to information, it can be argued that finding true alpha—an investing “edge”—has become nearly impossible without industrial-scale data mining and sophisticated predictive models. Wall Street quant shops such as Jane Street, Millennium, and Citadel dominate the CTA space by employing top PhDs and math Olympiad winners to sift through massive datasets in search of signals that yield elusive market advantages.

In this project, despite not having access to the same level of cutting-edge infrastructure, we aim to develop a machine learning model to forecast *responder_6*, predicting its behavior up to six months into the future. Our goal is to demonstrate that even with more modest resources, thoughtful methodology and careful modeling can still provide valuable insights.

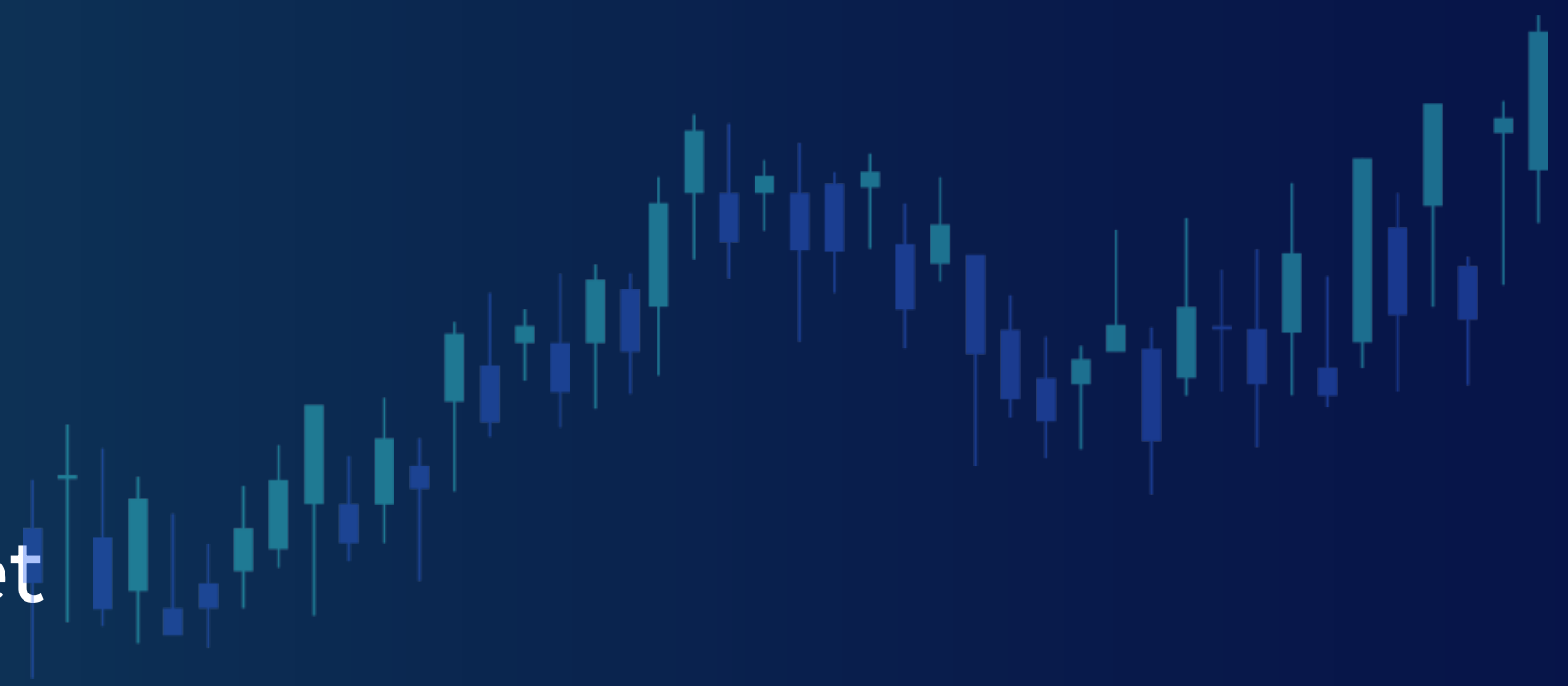
Dataset Description

- Dataset from an ongoing Kaggle competition hosted by Jane Street Group
- Consists of time series financial data
- Over 47 million rows with 92 columns (including 83 features and 9 responders)
- All features and responders are anonymized



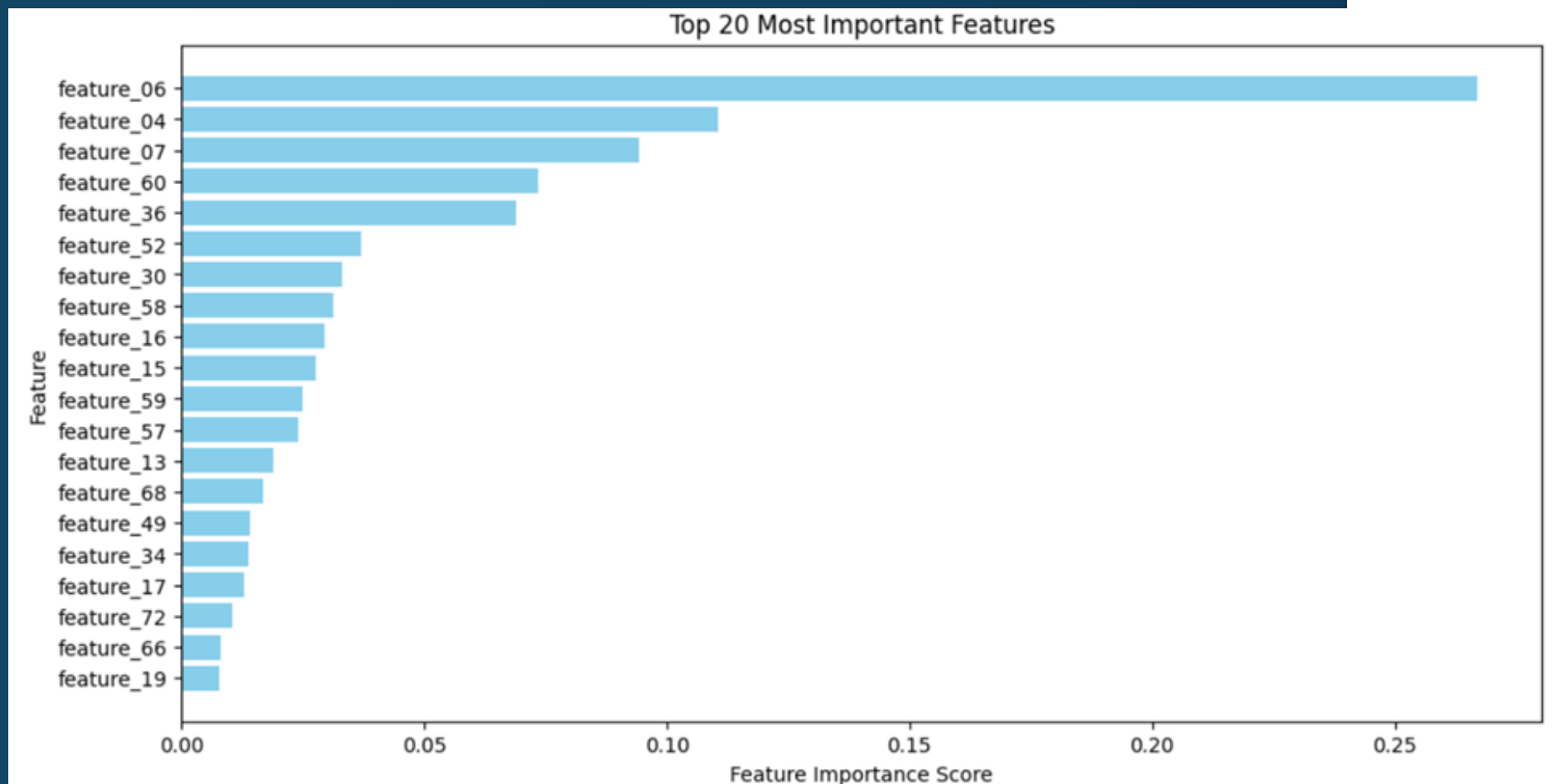
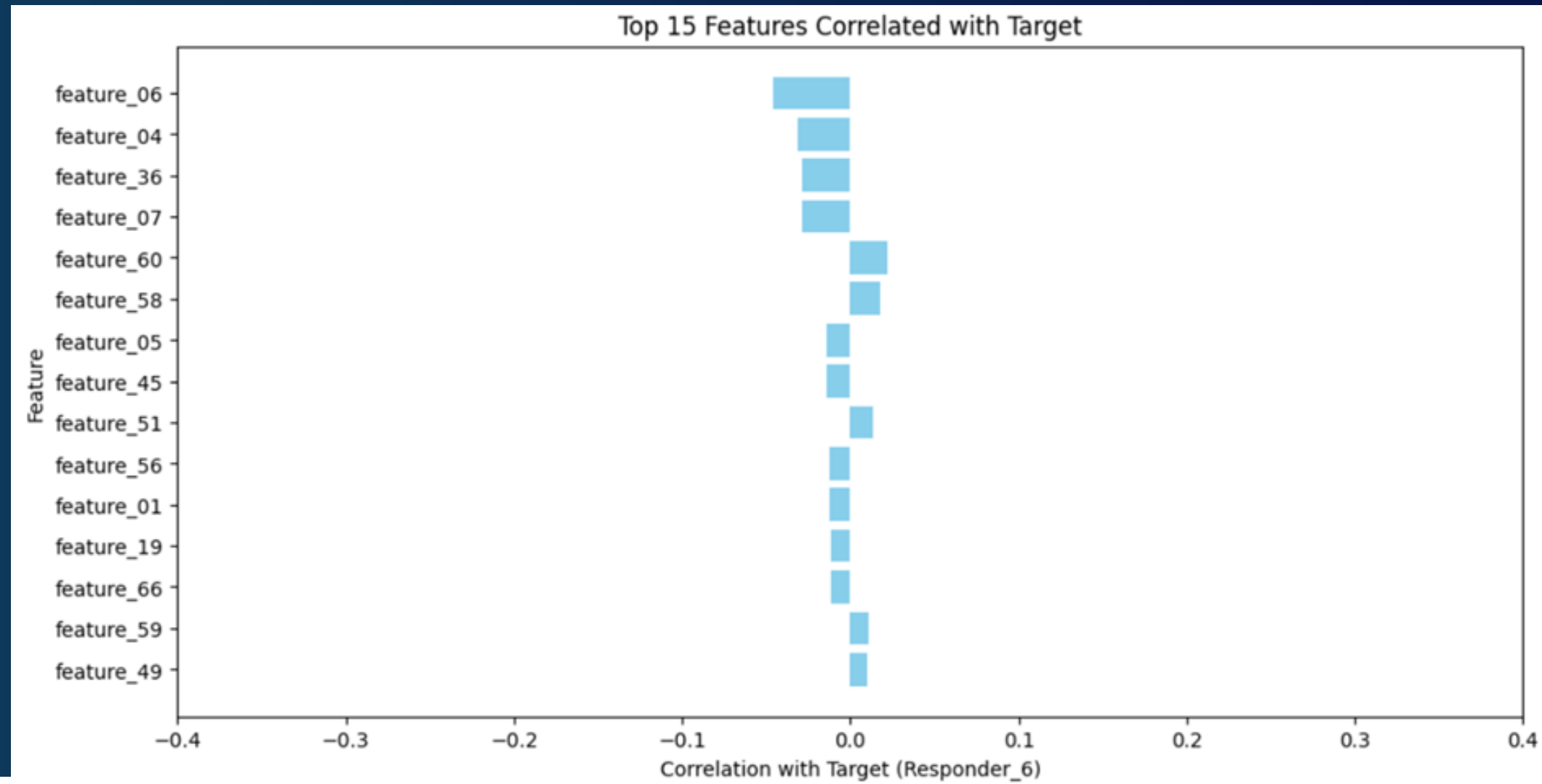
EDA

- 5% random sample of full dataset
- Dropped responders except for target
- Check for categorical data
- Check for missing data
- Dropped rows with missing data in features with $> 10\%$ missing data
- Imputed missing values in features with $< 10\%$ missing data
- Dropped low variance features



Data Preprocessing

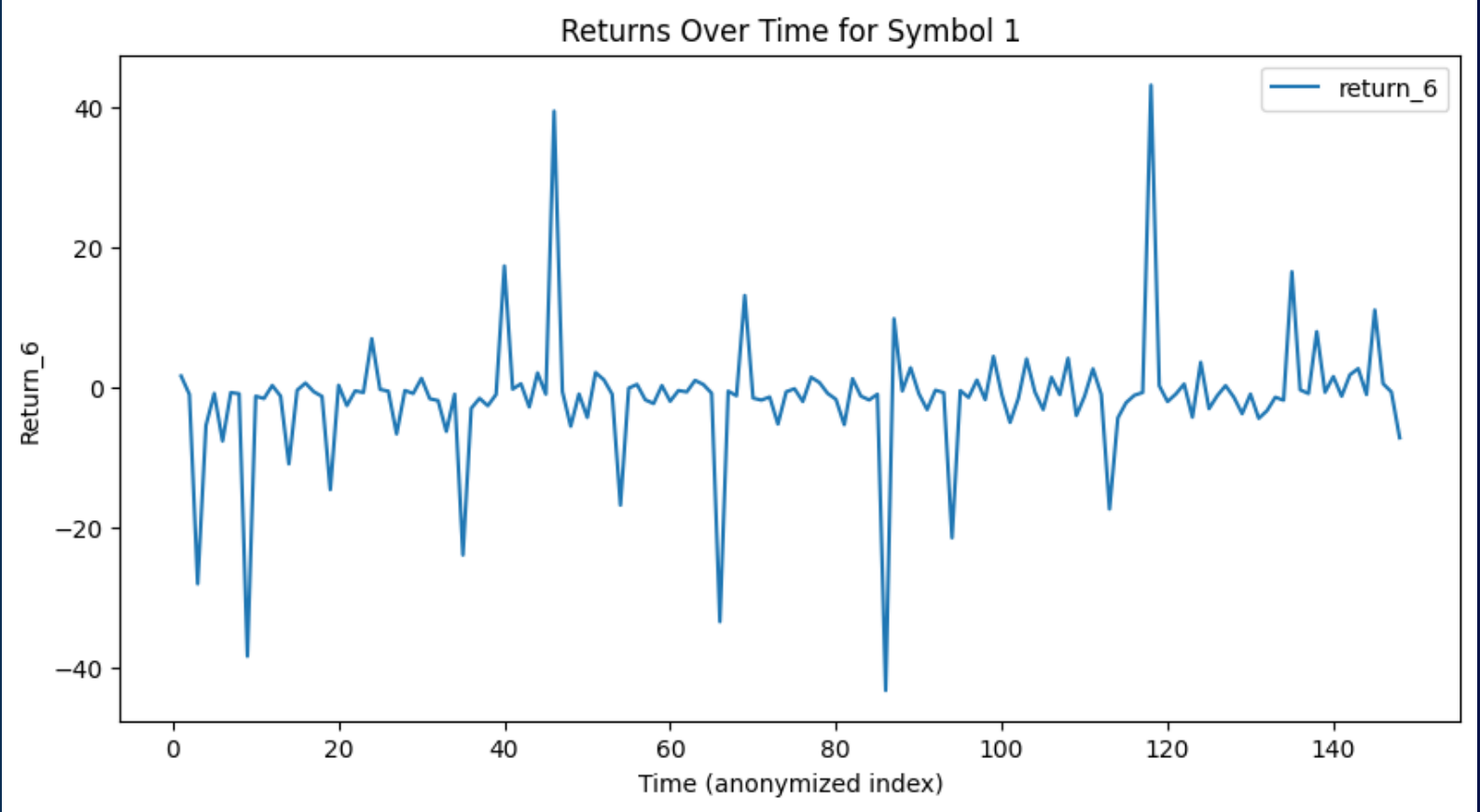
- Feature 6 is most predictive and has strongest correlation to Responder 6



Approach 1 to Modeling

- Models:
- XGBoost and LightGBM
 - Tested on Entire Dataset and Sample

Model	R2
XGBoost (Entire Dataset)	0.0079
LightGBM (Entire Dataset)	-0.11
LightGBM (Curated Dataset)	0.015



Method 2 Modeling

Method 2 process

- created the 5 % random sample
- extracted only features with highest correlation and in the top 15 most important features

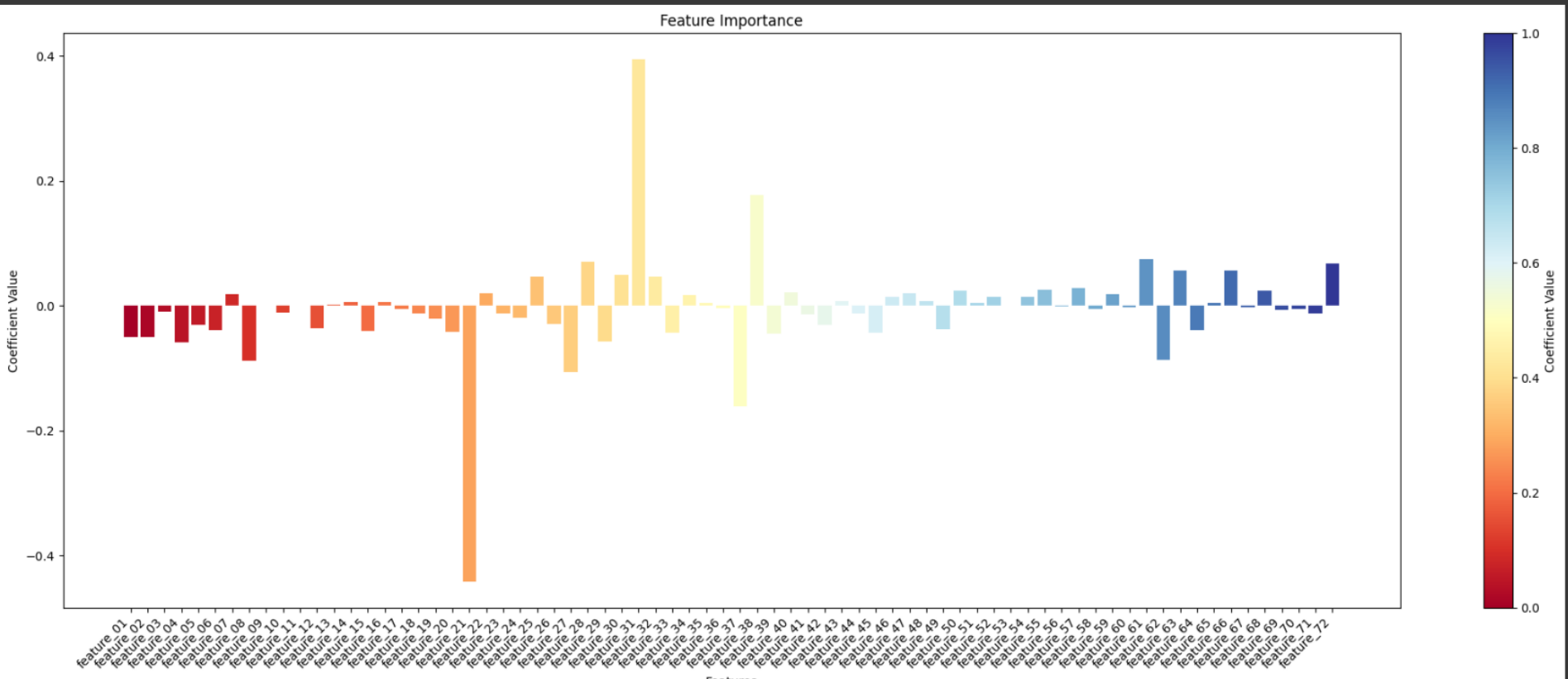
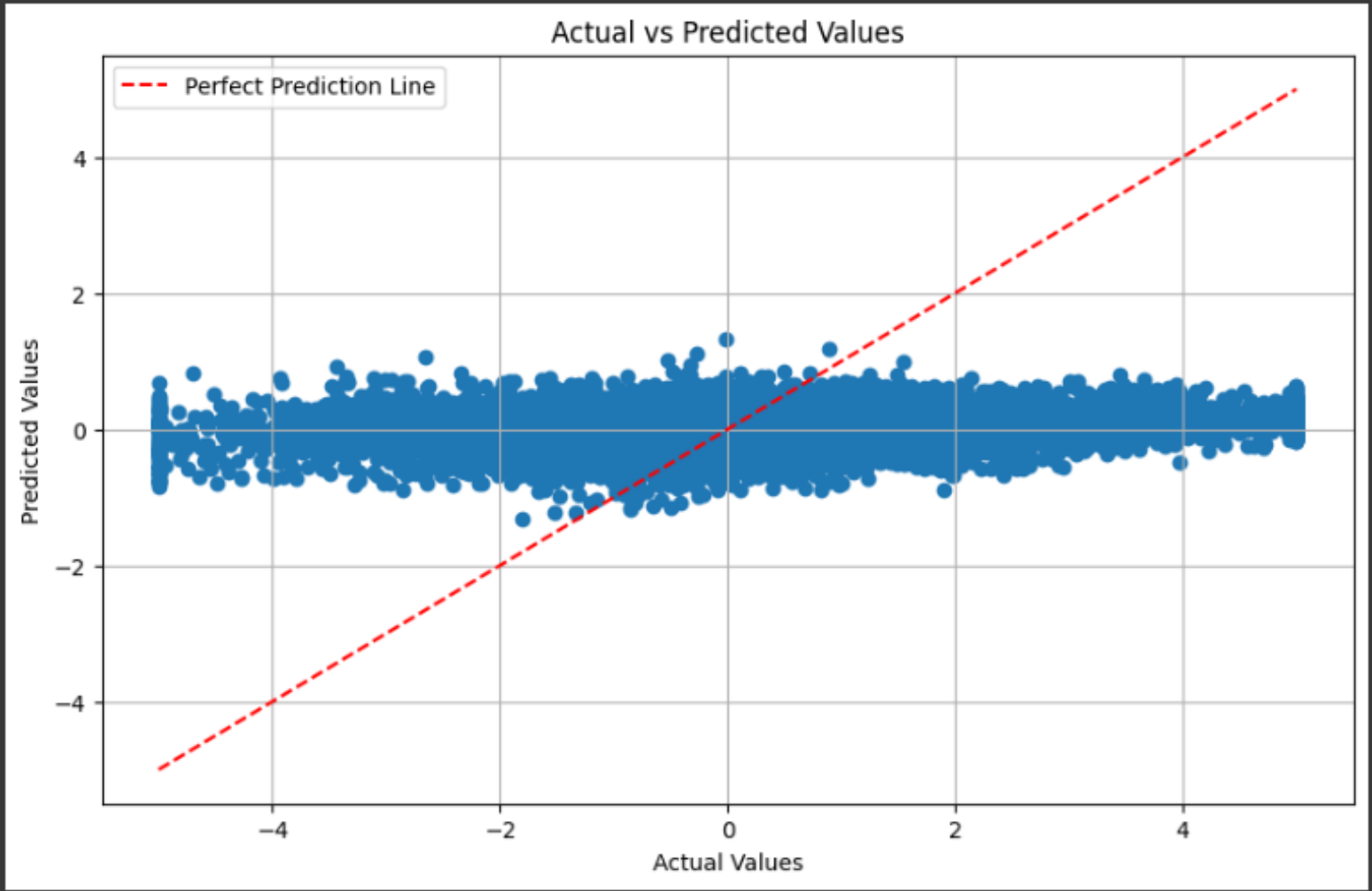
Models:

- XGBoost Regression with cross validation
 - best performance
- Random Forest Regression with manual tuning
 - poor performance may suggest issues with model stability or data preprocessing

Model	XGB	Random Forest
Best Model	Max Depth = 5 Learning Rate = 0.1	Max Depth = 3 Num of Trees = 20 Subsampling = 0.5
R ²	0.0129	0.0036

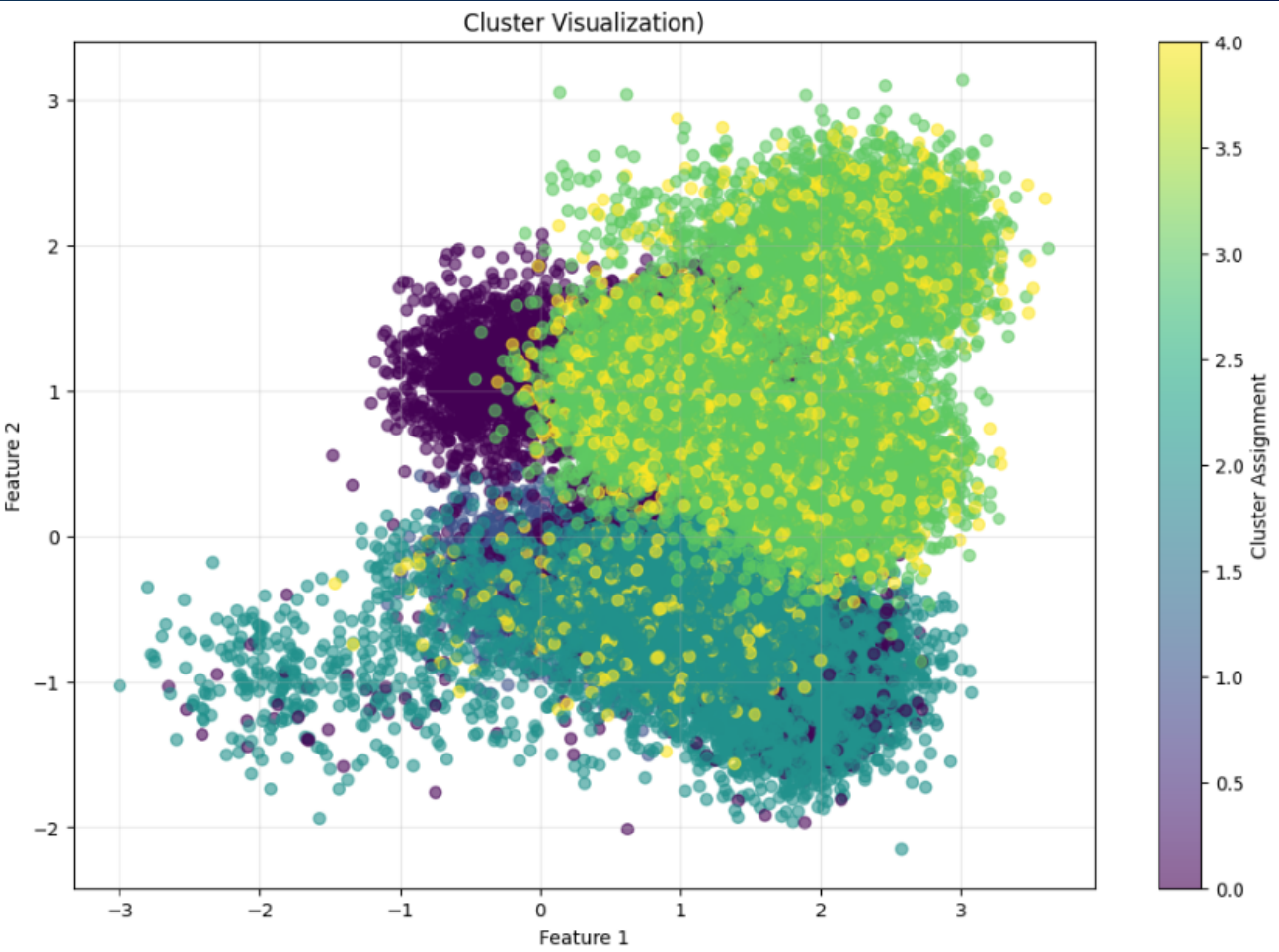
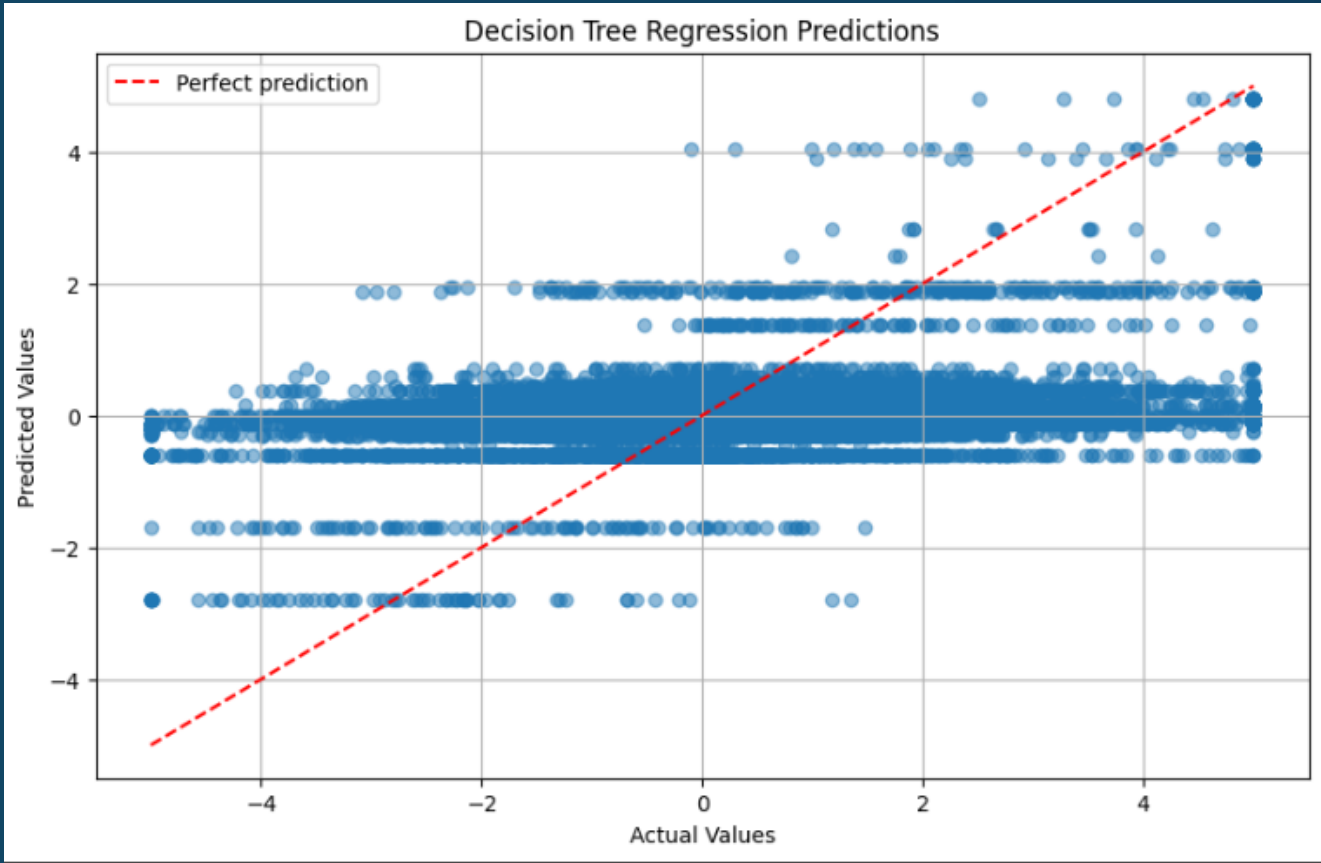
Method 3 LR and LR with PCA

Model	Linear Regression	Linear Regression (PCA)
R^2	0.02966	0.00636



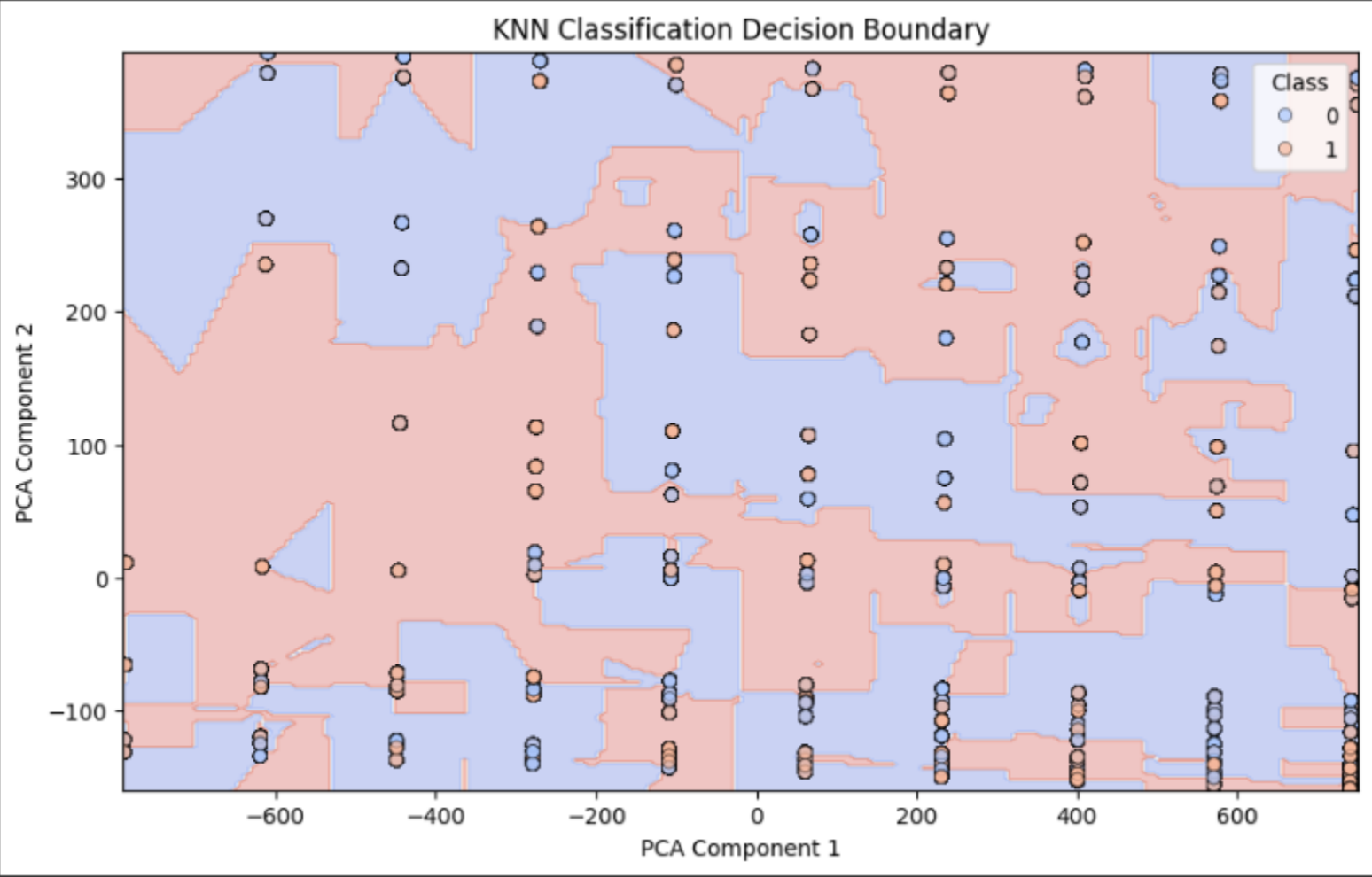
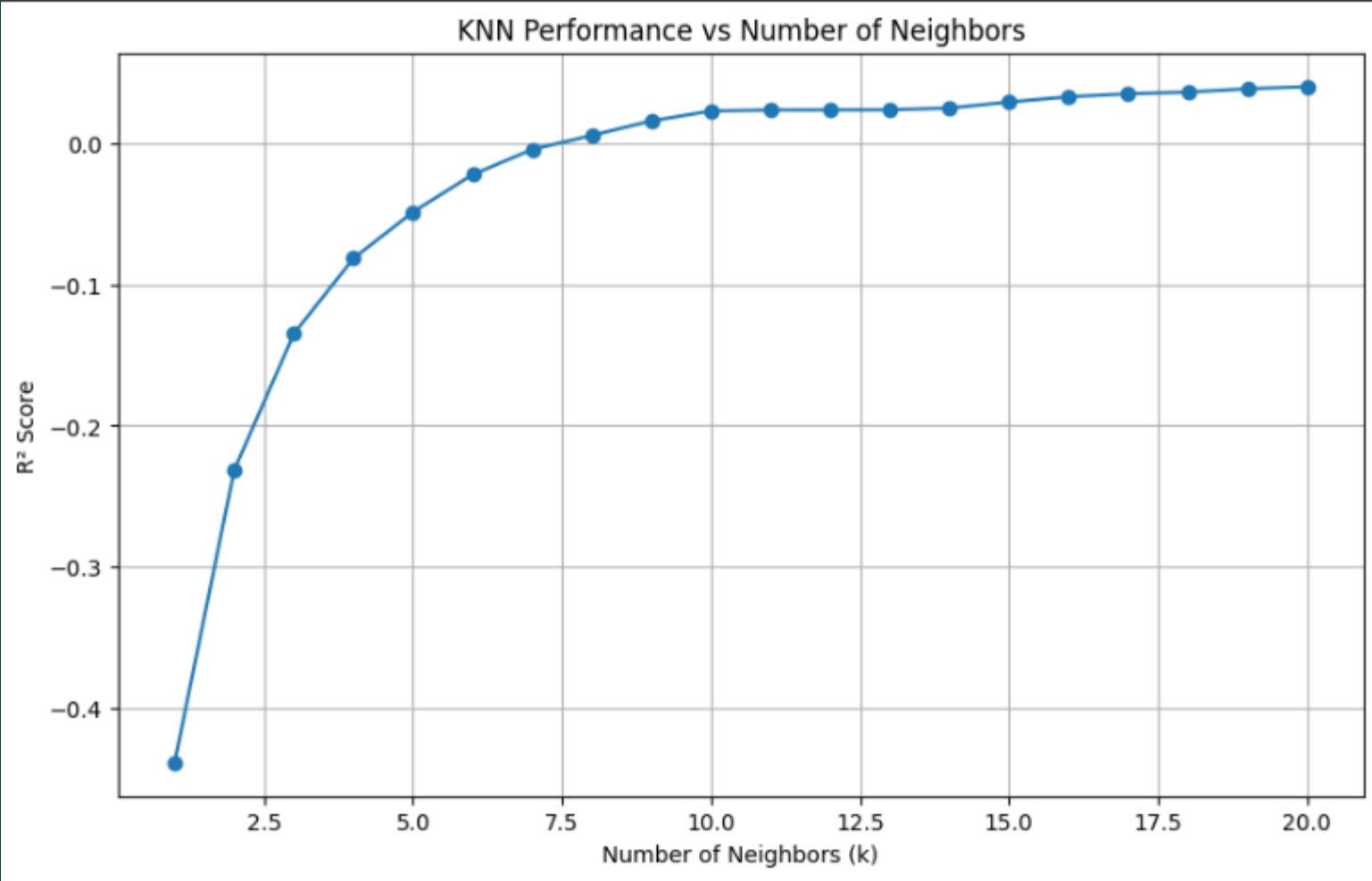
Method 4 Decision Trees

Model	Decision Trees(Small Sample)	Decision Trees(Large Sample)	Decision Tree (With Clustering)
R ²	0.00753	0.08499	0.09723



Method 5 KNN

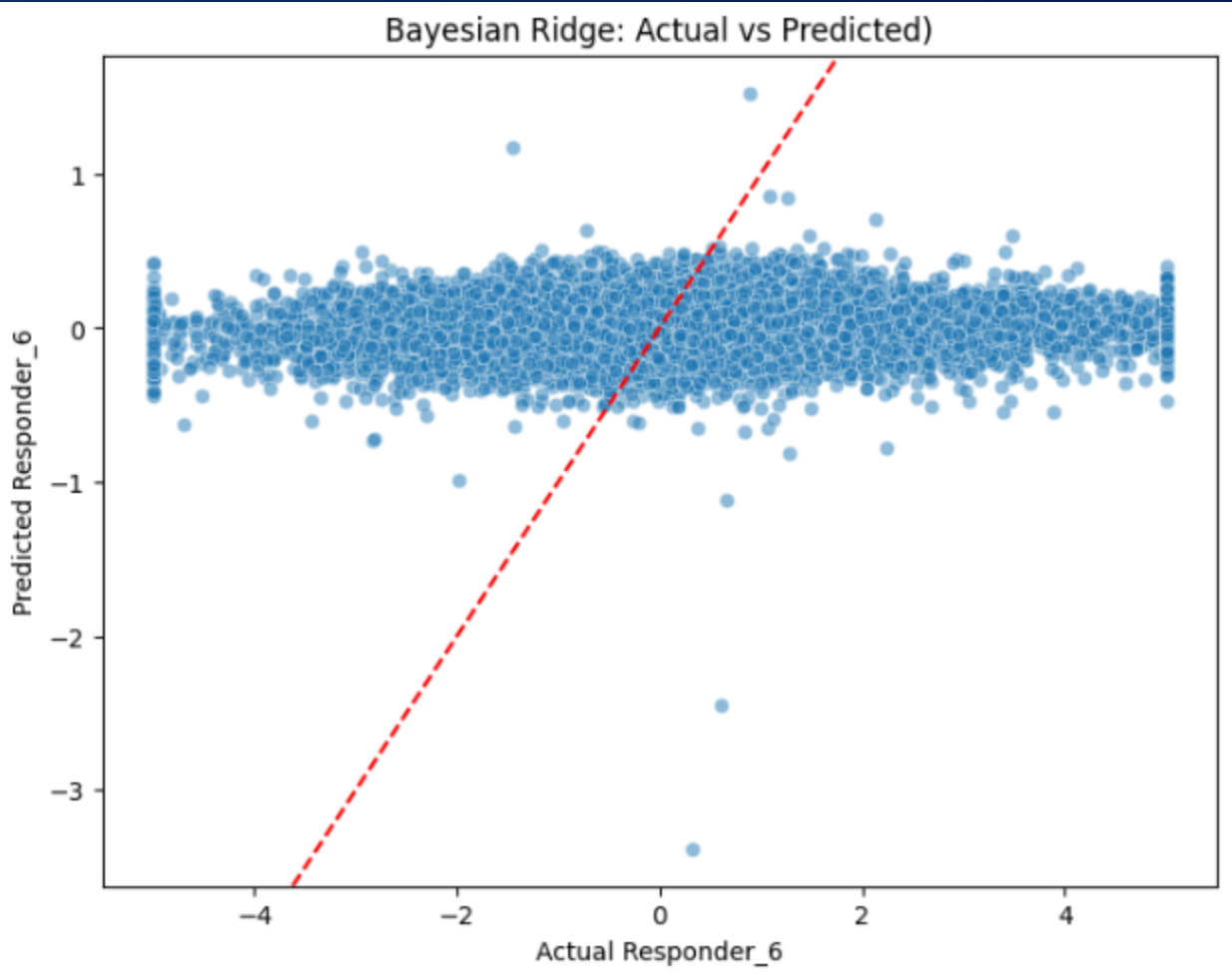
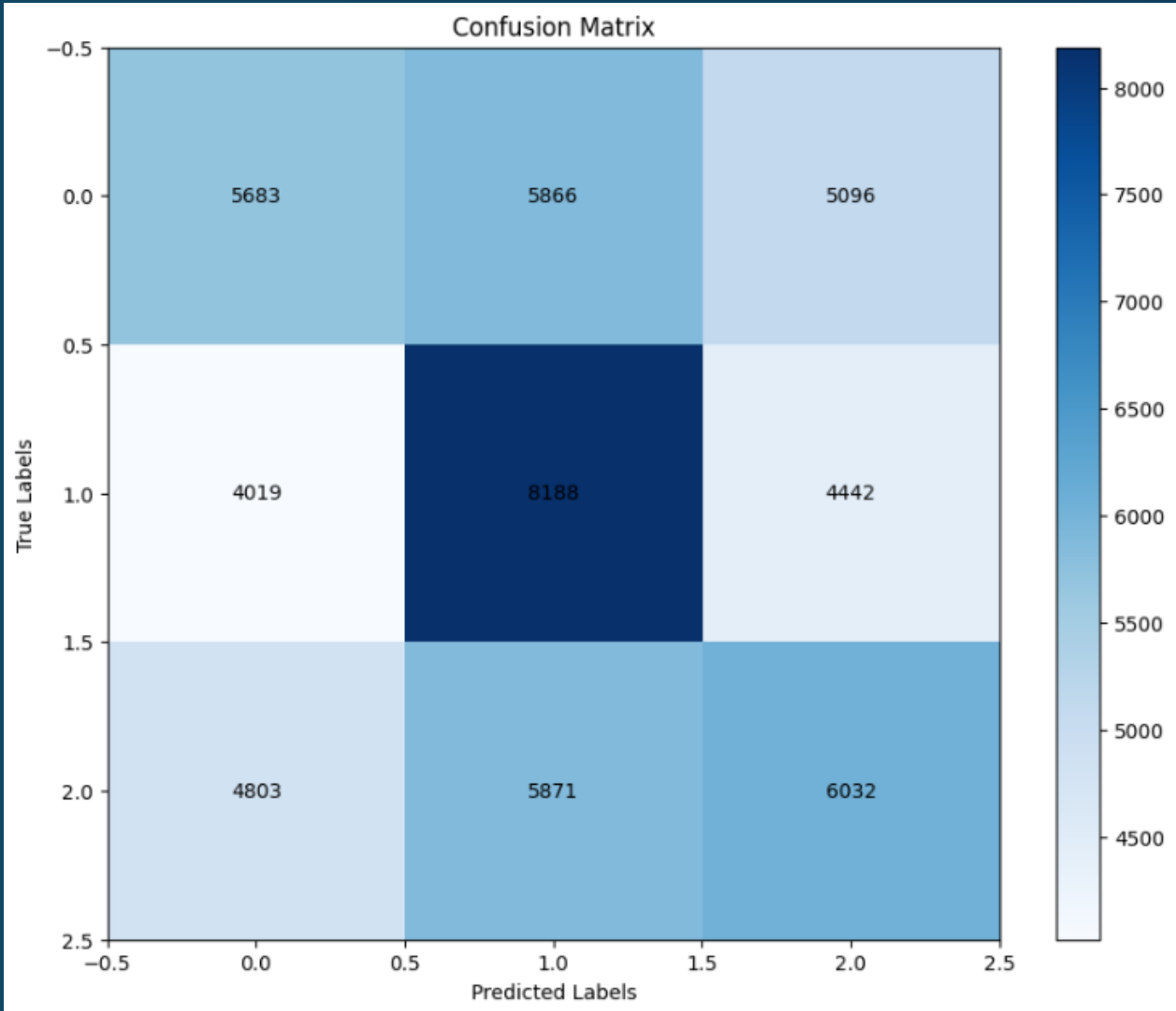
Model	KNN	KNN Classification Accuracy	KNN Best
R ²	0.02239	0.57%	0.040



Method 6 Naive Bayes

Model	Naive Bayes Accuracy	Bayesian Ridge
R ²	0.398%	0.00449

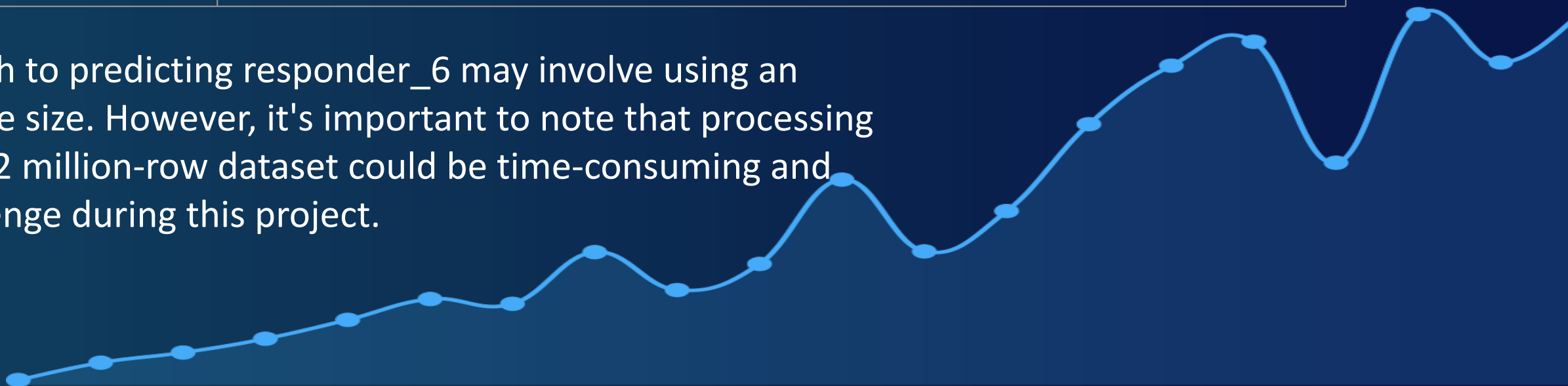
Using the Naive Bayes and Bayesian Ridge Models proved Ineffective



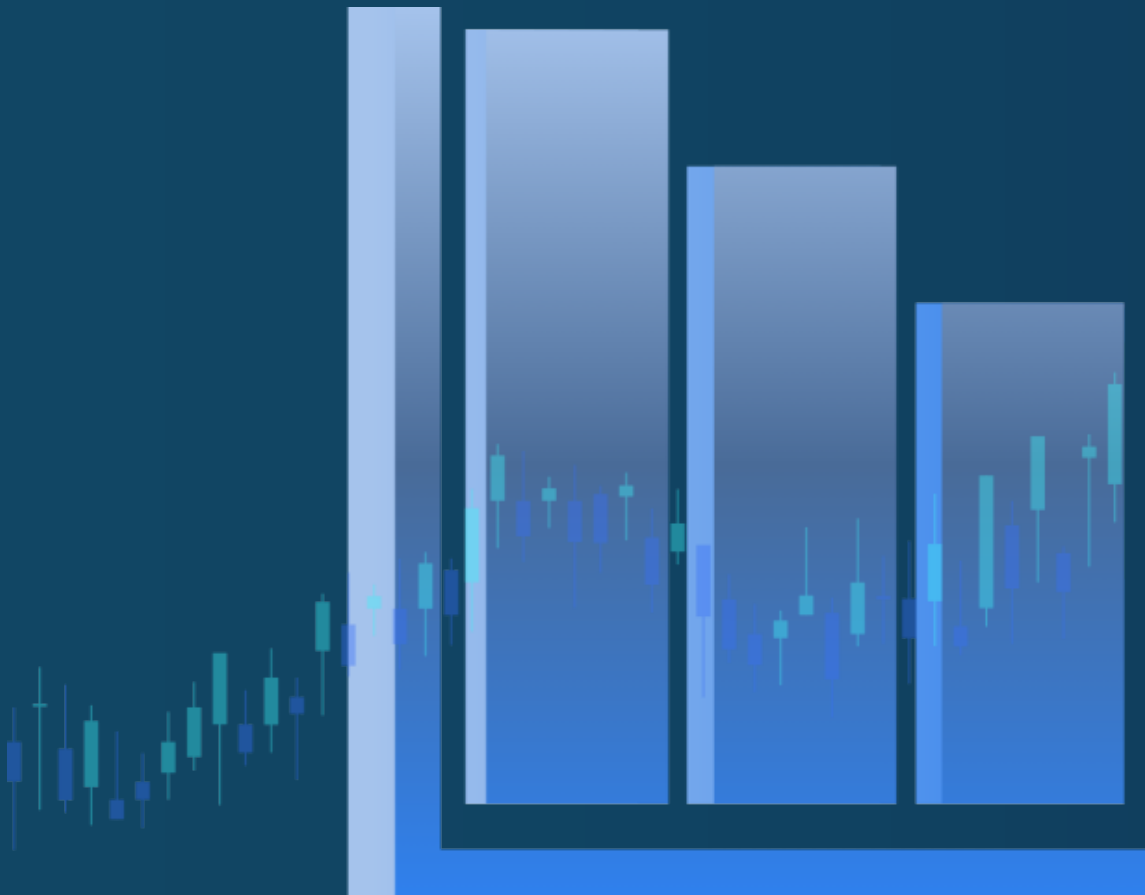
Results Summary - All Methods
























Model	R2
XGBoost	0.0079
LightGBM (Full Dataset)	-0.11
LightGBM (Curated Dataset)	0.015
XGBoost (with Feature Selection)	0.0129
Random Forest (with Feature Selection)	0.0036
Linear Regression (Small Sample)	0.0296
Linear Regression (PCA)	0.00636
Decision Trees (Small Sample)	0.00753
Decision Trees (Large Sample)	0.08499
Decision Trees (Clustering)	0.09723
KNN	0.02239
Naive Bayes / Bayesian Ridge	0.00449

The summary suggests that the most effective approach to predicting responder_6 may involve using an enhanced version of Decision Trees with a larger sample size. However, it's important to note that processing the mathematical calculations for decision trees on a 12 million-row dataset could be time-consuming and resource-intensive, which presented a significant challenge during this project.



Submission Results



Jane Street Real-Time Market Data Forecasting							...
OverviewDataCodeModelsDiscussionLeaderboardRulesTeam							
PublicPrivate							
This leaderboard is calculated with only the public test data. This is a forecasting competition; private scores will use data gathered after the submission deadline.							
Prize Contenders							
#	Team	Members		Score	Entries	Last	
1	ms capital	  		0.014276	2	1mo	
2	Patrick Yam			0.014069	2	1mo	
3	friend of volatility			0.013848	2	1mo	
4	hyd			0.011935	2	1mo	
5	Haoze Hou			0.011896	2	1mo	
6	ponythewhite			0.011098	2	1mo	
7	Thomas Dueholm Hansen			0.010924	2	1mo	
8	Maaax	 		0.010608	2	1mo	
9	leo			0.010520	2	1mo	
10	Victor Shlepov			0.010438	2	1mo	

Thank You!

