Machine Learning Framework for Company Bankruptcy Prediction

Kuber Budhija
Palak Bhardwaj
Rishi Pendyala
Yashovardhan Singhal
Group number 66
ML-mid sem project

INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY
DELHI

Motivation

Motivation

- Increasing number of companies face bankruptcy, impacting investors, employees, and stakeholders.
- Early prediction of bankruptcy helps in minimizing economic impact.
- Aim: Use machine learning models to predict corporate bankruptcy.

Literature review

[Reference 1](#)

SVM and comparing with traditional methods.

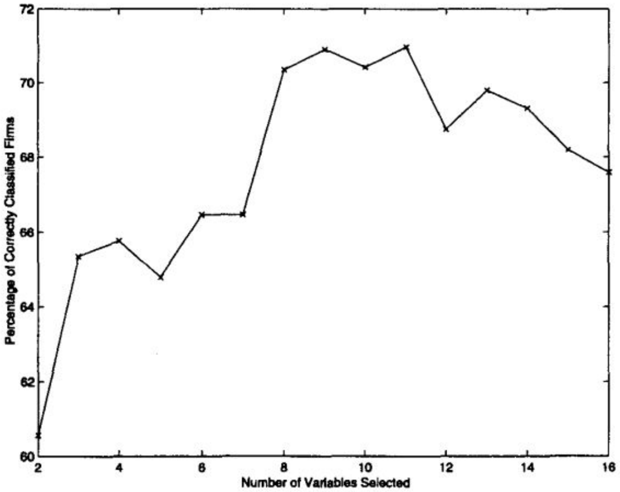
	Altman		Lincoln		Ohlson	
	Training	*Testing*	*Training*	*Testing*	*Training*	*Testing*
LDA	65.76	64.31	68.78	62.15	70.87	64.72
MLP	68.07	62.85	81.59	64.90	72.23	68.61
LVQ	69.59	62.71	72.55	66.25	72.55	66.25
SVM	74.05	65.14	87.24	67.22	81.15	69.17

[Reference 2](#)

Use of Boosting, bagging, and random forest models.

[Reference 3](#)

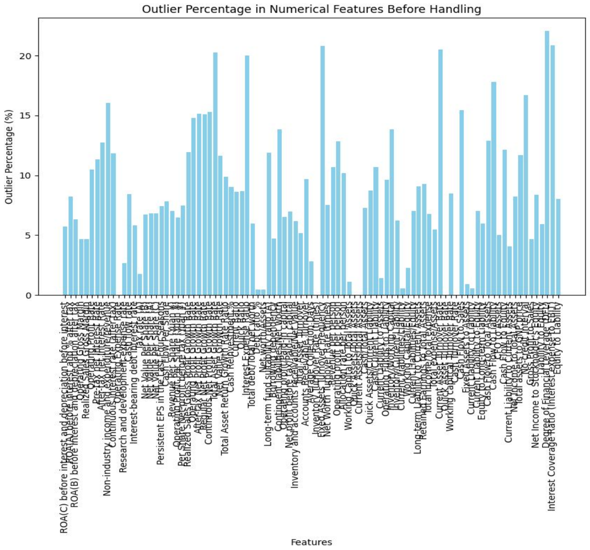
Deep learning models.



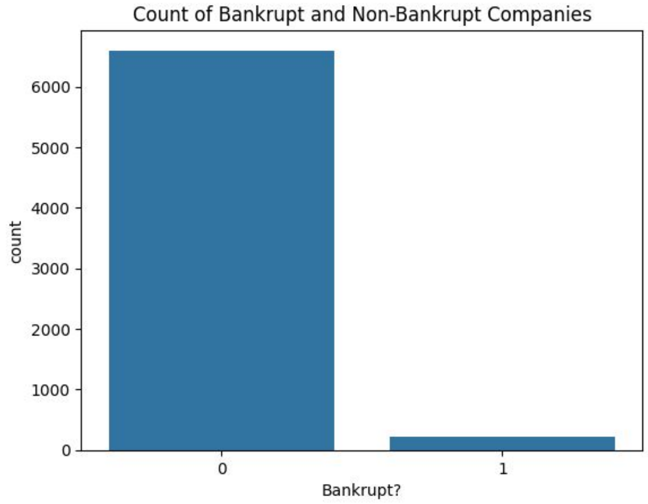
Dataset description

Dataset Overview

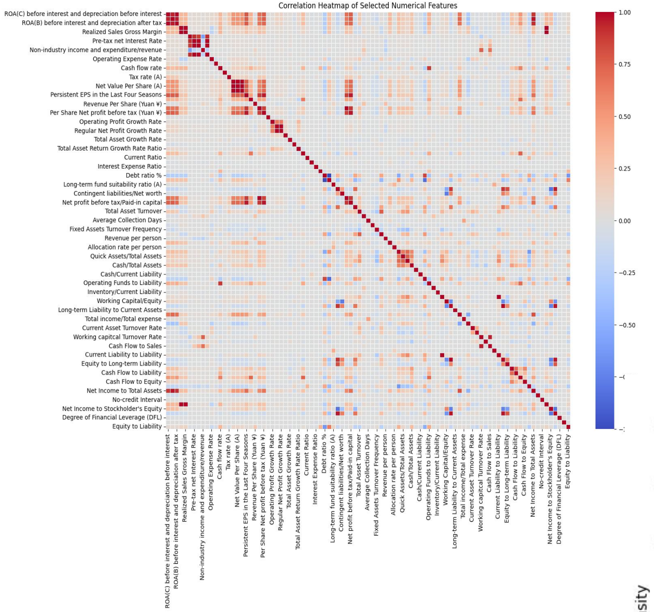
- Source: Taiwan Economic Journal (1999-2009).
- Features: 96 continuous parameters, 6819 rows.
- Imbalance: Majority of companies are non-bankrupt.
  [Dataset]

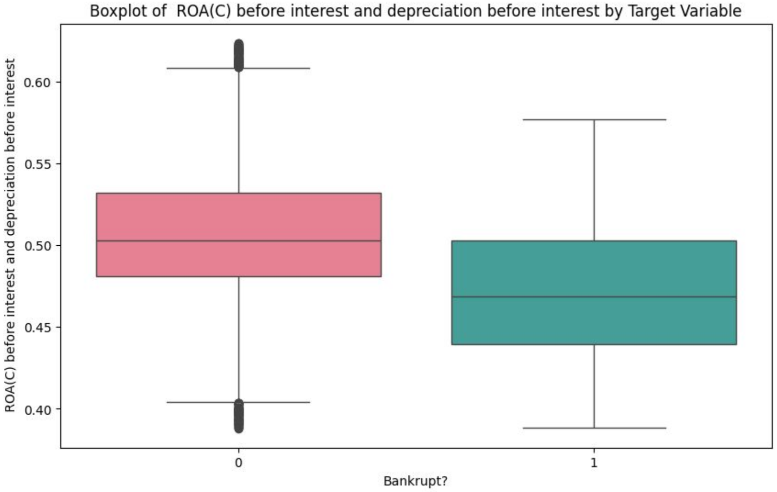


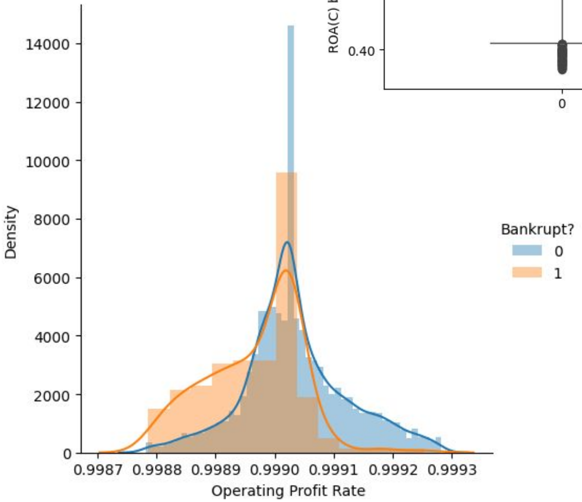
TEJ Ratings · Risk · Index · Databank



Dataset Description: EDA



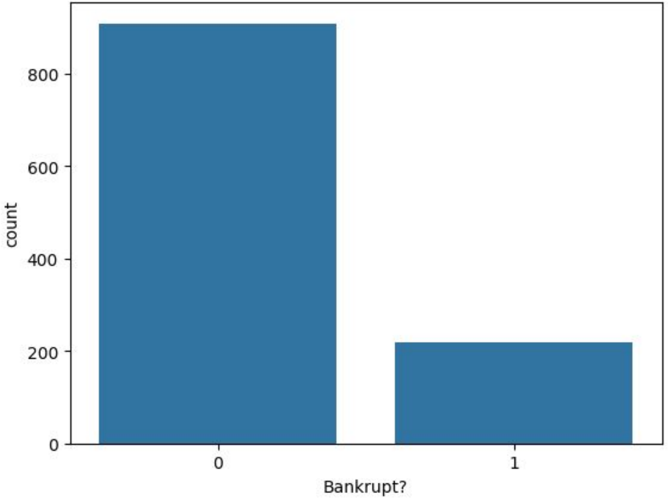




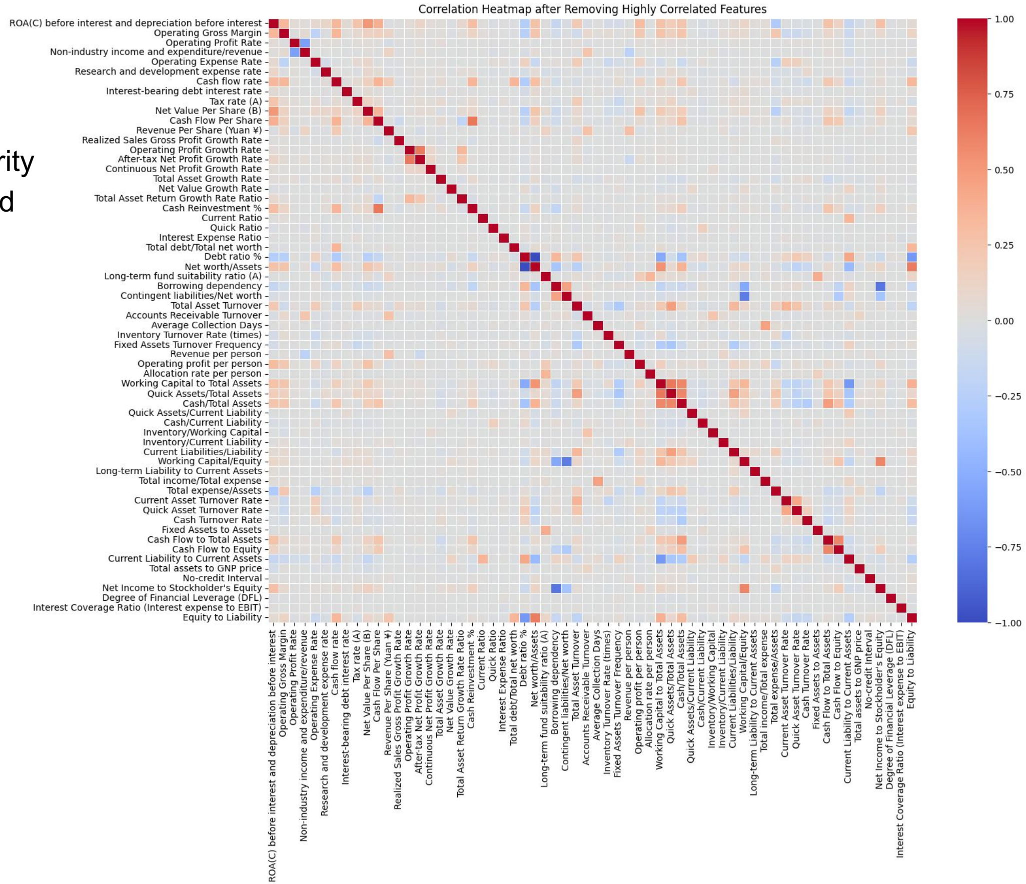
Methodology : Preprocessing

Data Preprocessing

- Handled imbalance using:
  - RandomOverSampler, RandomUnderSampler, SMOTE, ADASYN, Condensed Nearest Neighbor, Tomek Link



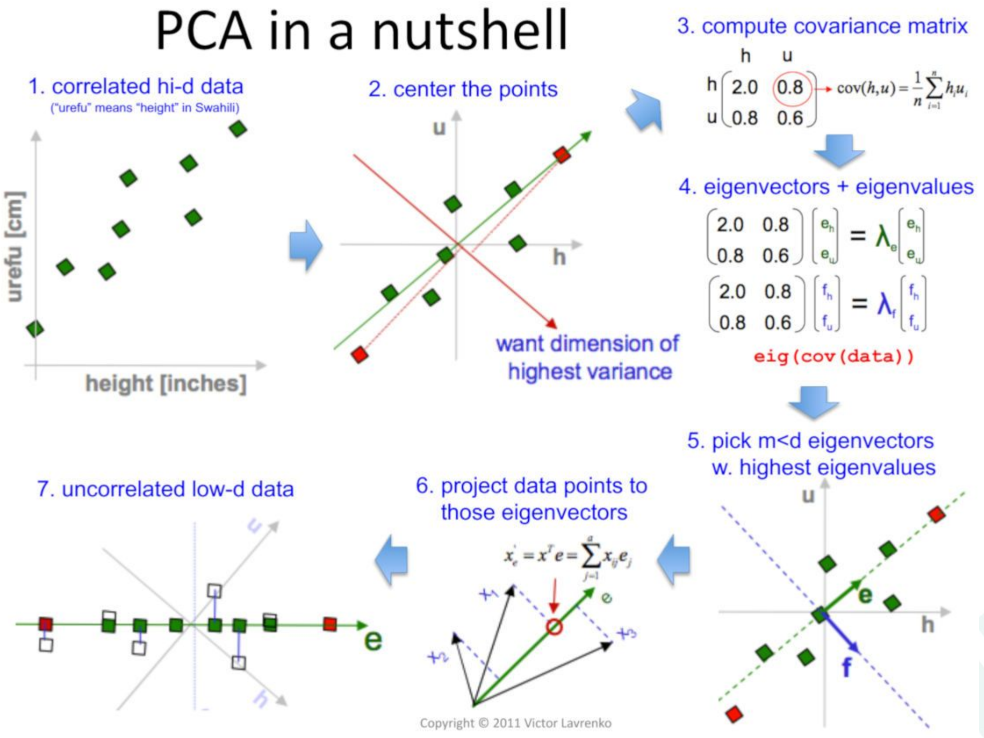
- Handled multicollinearity with threshold 0.70



Methodology : Feature Selection

Feature Selection

- Used ANOVA for numerical features and Chi-square tests for categorical features for importance
- Used PCA for dimensionality reduction but was not very useful
- Used SelectKBest to choose best 30 features



https://www.youtube.com/watch?v=UW9U0bYJ-Ys

https://devopedia.org/images/article/139/4543.1548137789.jpg

Methodology : Models

- Random Forest
- Logistic Regression
- Naive Bayes
- SVM
- MLP
- XGBoost

Results and Analysis

Table 1. Training Metrics for Different Models

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	Log Loss
Logistic Regression	0.7763	0.4620	0.8202	0.5911	0.8613	0.4994
Random Forest	1.0000	1.0000	1.0000	1.0000	1.0000	0.0839
Gaussian Naive Bayes	0.7663	0.4489	0.8146	0.5788	0.8711	2.3367
Decision Tree	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
SVM (Linear Kernel)	0.7741	0.4601	0.8427	0.5952	0.8619	0.3468
SVM (RBF Kernel)	0.7785	0.4656	0.8371	0.5984	0.8629	0.3454
SVM (Poly Kernel)	0.7885	0.4792	0.8427	0.6110	0.8728	0.3368
MLP Classifier	0.8472	0.6493	0.4888	0.5577	0.8712	0.3379
ANN (Custom)	0.8427	0.6475	0.4438	0.5267	0.8722	0.3365
XGBoost	1.0000	1.0000	1.0000	1.0000	1.0000	0.0107

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	Log Loss
Logistic Regression	0.8053	0.4868	0.8810	0.6271	0.8763	0.4893
Random Forest	0.8938	0.7368	0.6667	0.7000	0.9109	0.2816
Gaussian Naive Bayes	0.7920	0.4684	0.8810	0.6116	0.8480	2.3010
Decision Tree	0.8097	0.4909	0.6429	0.5567	0.7453	6.8579
SVM (Linear Kernel)	0.8009	0.4810	0.9048	0.6281	0.8813	0.3250
SVM (RBF Kernel)	0.8009	0.4810	0.9048	0.6281	0.8815	0.3239
SVM (Poly Kernel)	0.8097	0.4935	0.9048	0.6387	0.8863	0.3179
MLP Classifier	0.8540	0.6364	0.5000	0.5600	0.8888	0.3177
ANN (Custom)	0.8407	0.6250	0.3571	0.4545	0.8868	0.3186
XGBoost	0.8850	0.6905	0.6905	0.6905	0.9076	0.3853

Table 2. Test Metrics for Different Models

Results and Analysis

Preprocessing:

- Condensed Nearest Neighbors handled class imbalance better than other methods.
- SelectKBest was more effective than PCA for feature selection.

– **XGBoost:**

* The test accuracy of 88.50% is among the highest across models.
* Balanced precision and recall scores (0.6905 each) reflect consistent identification of positive cases and low false positives.
* The F1 score (0.6905) shows good performance

- **Random Forest:**

  – The model exhibits perfect training accuracy (100%), but there was a drop to 89.82% on the test set, which suggests slight overfitting; this can possibly be solved by better feature selection and sampling methods.

  – It achieved a precision score of 0.7879 on the test set, indicating a strong ability to classify positive samples correctly.

  – The F1 score (0.6933) indicates a reasonable balance between precision and recall

- **MLP Classifier**

– **MLP Classifier:**

* The model achieved a test accuracy of 85.40%, higher than many other models.
* Precision (0.5000) and recall (0.5600) are moderate

Results and Analysis

MLP Classifier	Predicted: No	Predicted: Yes
Actual: No	172	12
Actual: Yes	21	21

Table 4. Confusion Matrix for MLP Classifier

XGBoost	Predicted: No	Predicted: Yes
Actual: No	171	13
Actual: Yes	13	29

Table 5. Confusion Matrix for XGBoost

Random Forest	Predicted: No	Predicted: Yes
Actual: No	174	10
Actual: Yes	14	28

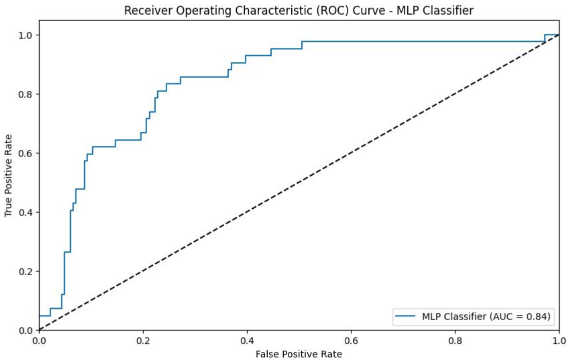
Table 6. Confusion Matrix for Random Forest

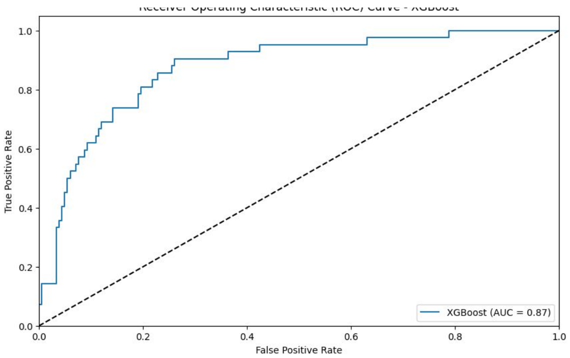
Predicting whether a company will go bankrupt or not

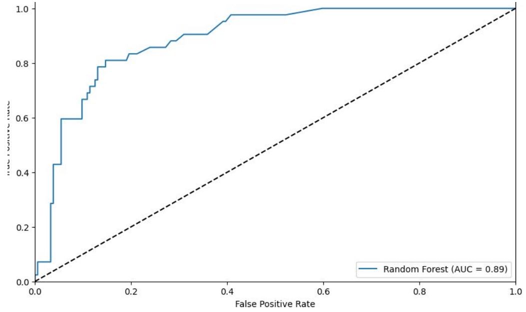
Models with higher sensitivity were preferred over models with higher specificity.

False negatives (failing to predict a bankruptcy) are more critical than false positives (incorrectly predicting a company will go bankrupt).

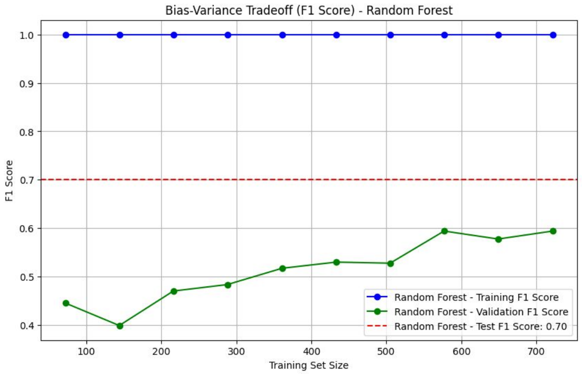
Thus, it is acceptable to have some false positives as long as the false negatives are minimized.

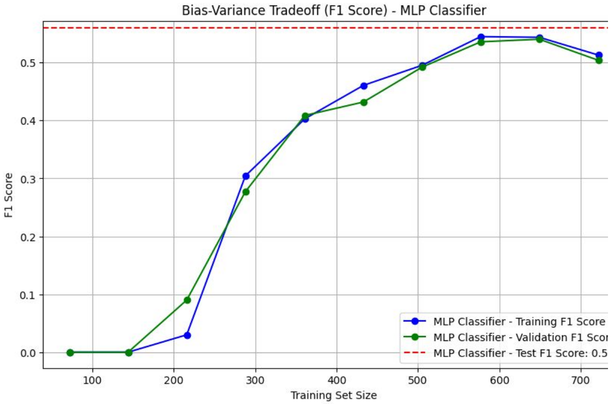


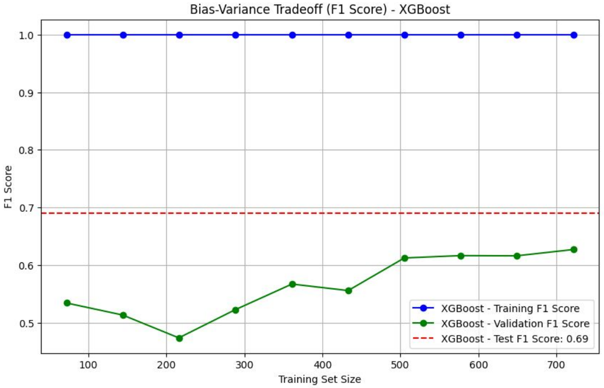


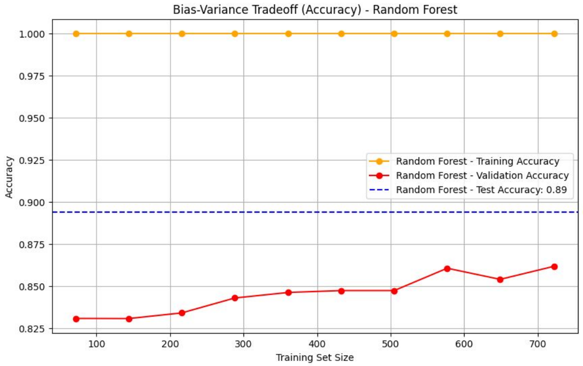


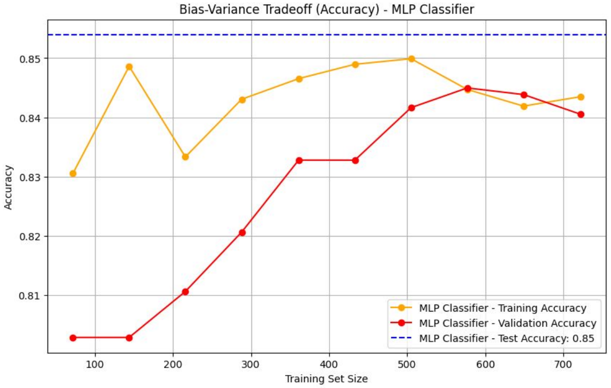
Results and Analysis

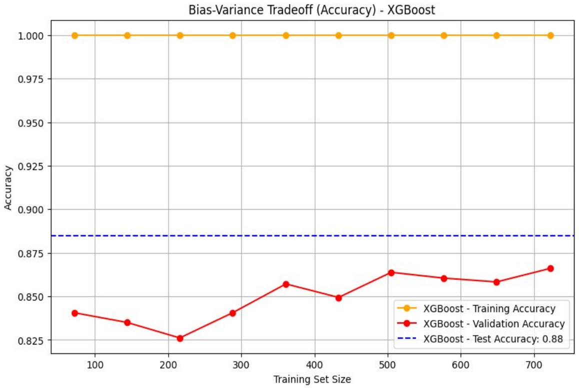












Individual team members' contributions

- Rishi: EDA, Preprocessing, Feature Selection, Model Training, Documentation
- Palak: EDA, Feature Selection, Documentation
- Yashovardhan: Literature Survey, Model Training, Presentation.
- Kuber: EDA, Preprocessing, Model Training, Presentation

Thank You