Predicting Bankruptcy & Detecting Money Laundering using Machine Learning

PRESENTED BY:

AMAN SINGH

VISHAL SINGH

Paper: Cognitive Modelling of Bankruptcy Risk

- Objective: Predict company bankruptcy using machine learning
- Dataset: US Company Bankruptcy Prediction Dataset (20 years of US company data from Kaggle)
- **Size**: 78,682 instances, 21 attributes
- Target Variable: status_label (alive: 1, failed: 0)
- Models used:
  - Logistic Regression
  - K-Nearest Neighbors
  - Decision Tree
  - Support Vector Machine (Linear & RBF)
  - Neural Network (Model Excluded)
  - Random Forest

Results of the Original Study

Best Performing Models

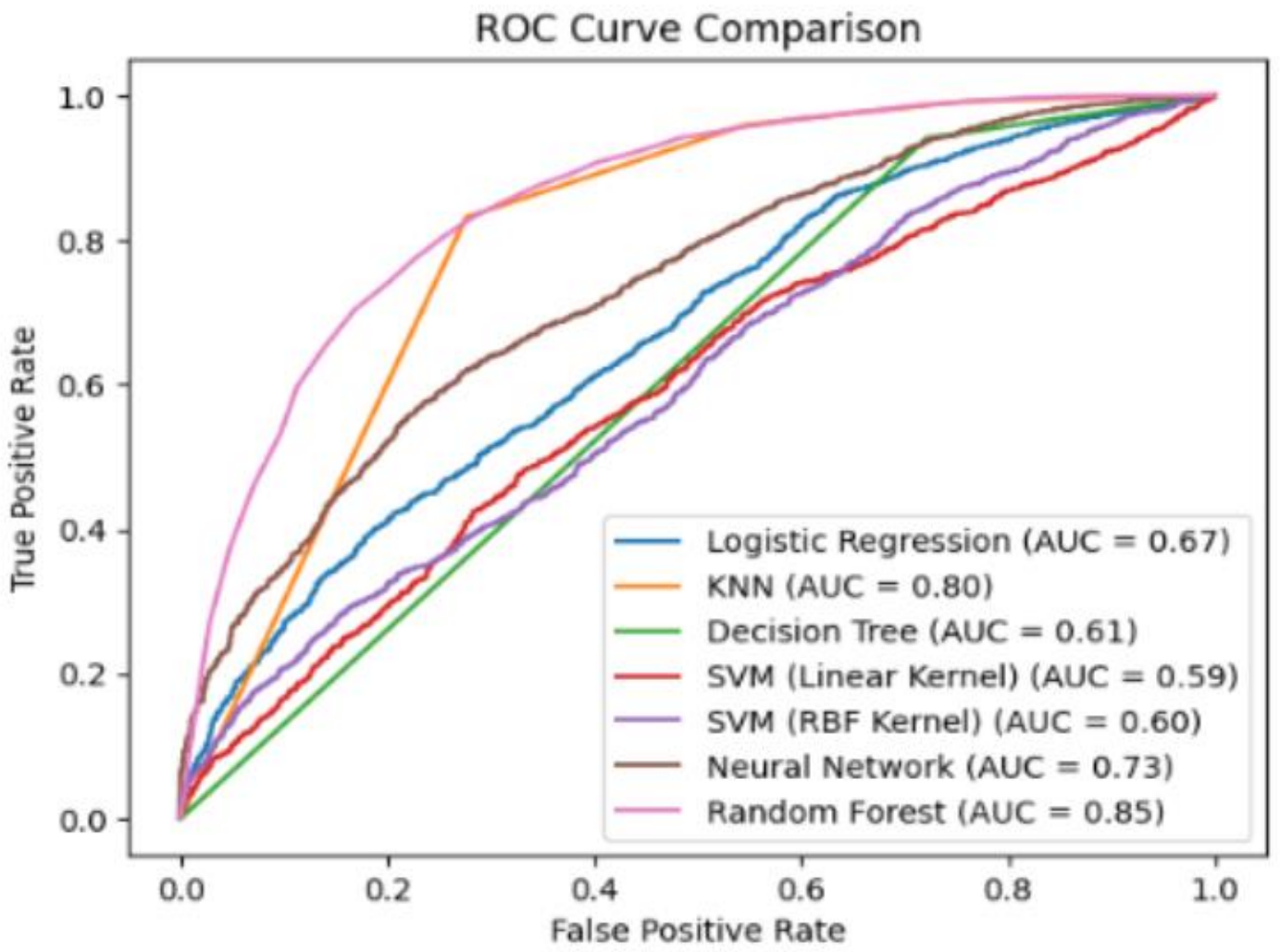
- **K-Nearest Neighbors (KNN):** Highest accuracy and strong balance across all metrics
- **Random Forest:** Best ROC-AUC, highly precise and stable

TABLE I. COMPARISON OF SEVEN MACHINE LEARNING ALGORITHMS

Models	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.936011	0.936641	0.999253	0.966935	0.667623
KNN	0.944144	0.95069	0.991788	0.970804	0.804537
Decision Tree	0.899028	0.950452	0.941228	0.945818	0.609836
SVM (Linear Kernel)	0.936328	0.936328	1.0	0.967117	0.591862
SVM (RBF Kernel)	0.936583	0.936622	0.999932	0.967242	0.599162
Neural Network	0.936519	0.937787	0.998439	0.967163	0.733672
Random Forest	0.941158	0.94121	0.999593	0.969523	0.853076

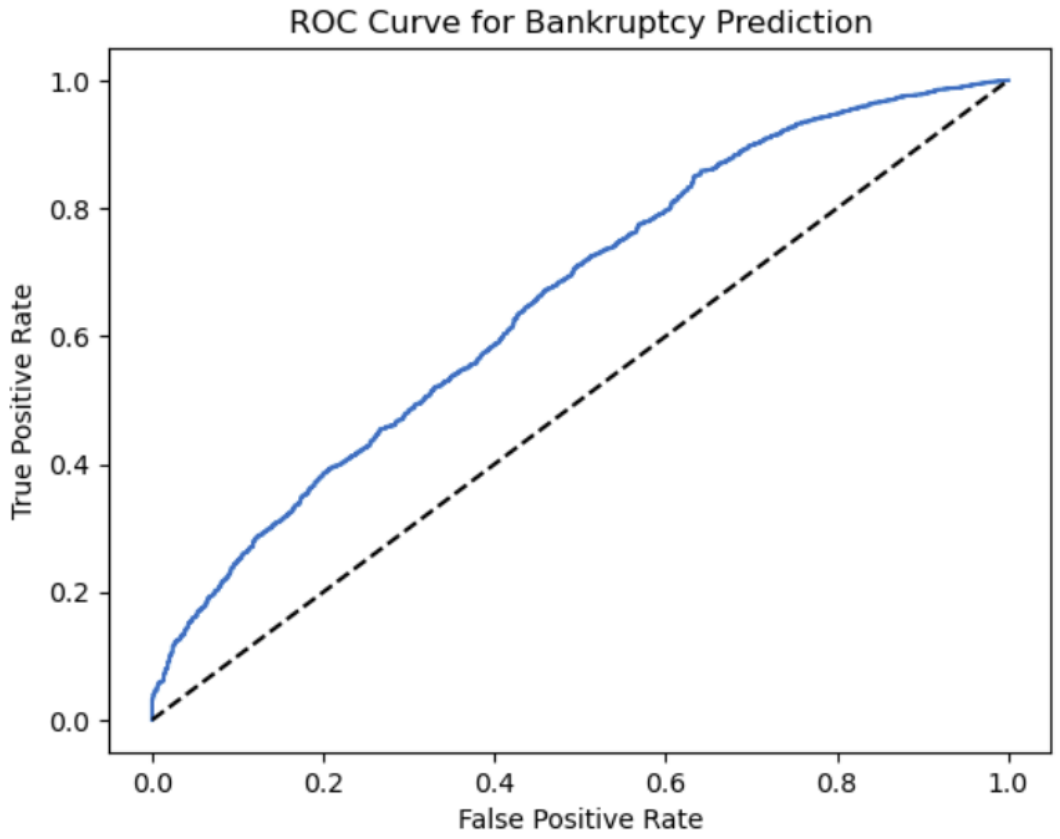
Results of the Original Study

- Random Forest: Highest AUC → 0.853
- KNN: AUC of 0.804
- Logistic Regression & SVMs: Lower AUCs (~0.59–0.66)



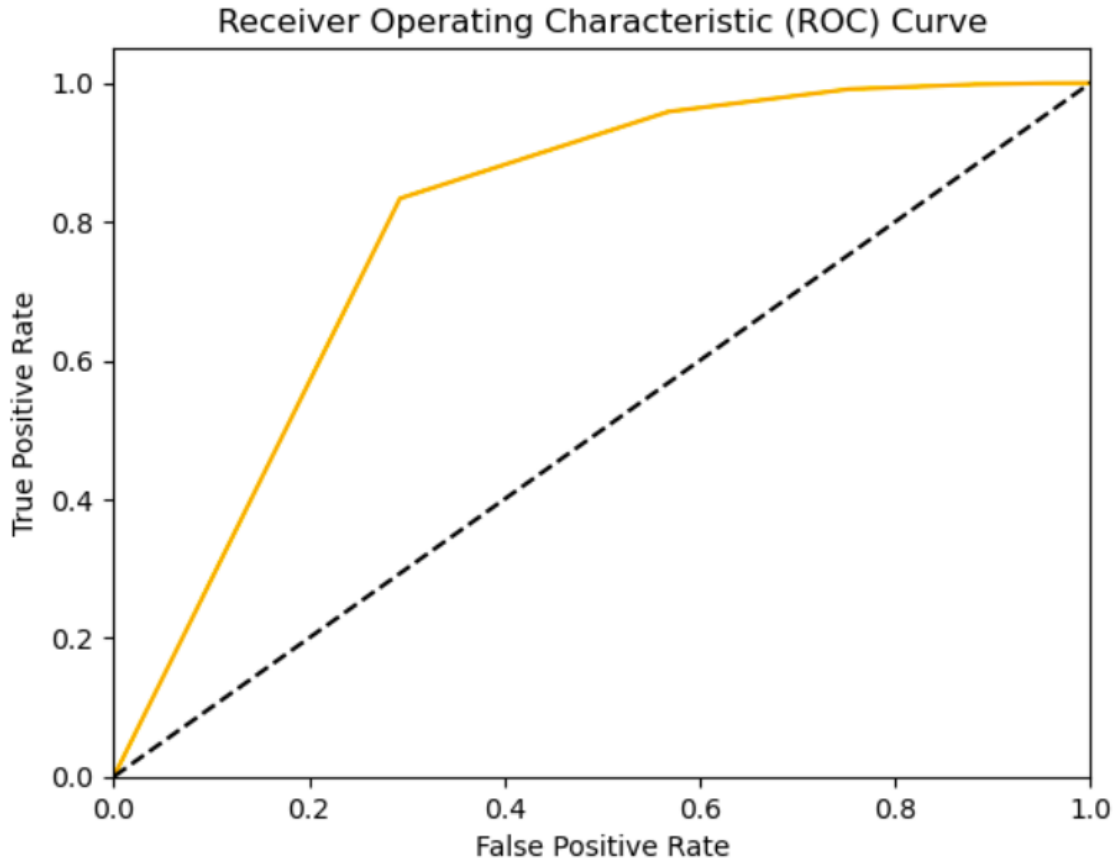
Logistic Regression – Model Evaluation

Model	Linear Regression
Accuracy	0.9335
Precision	0.9343
Recall	0.9992
F1 Score	0.9656
ROC-AUC	0.6578



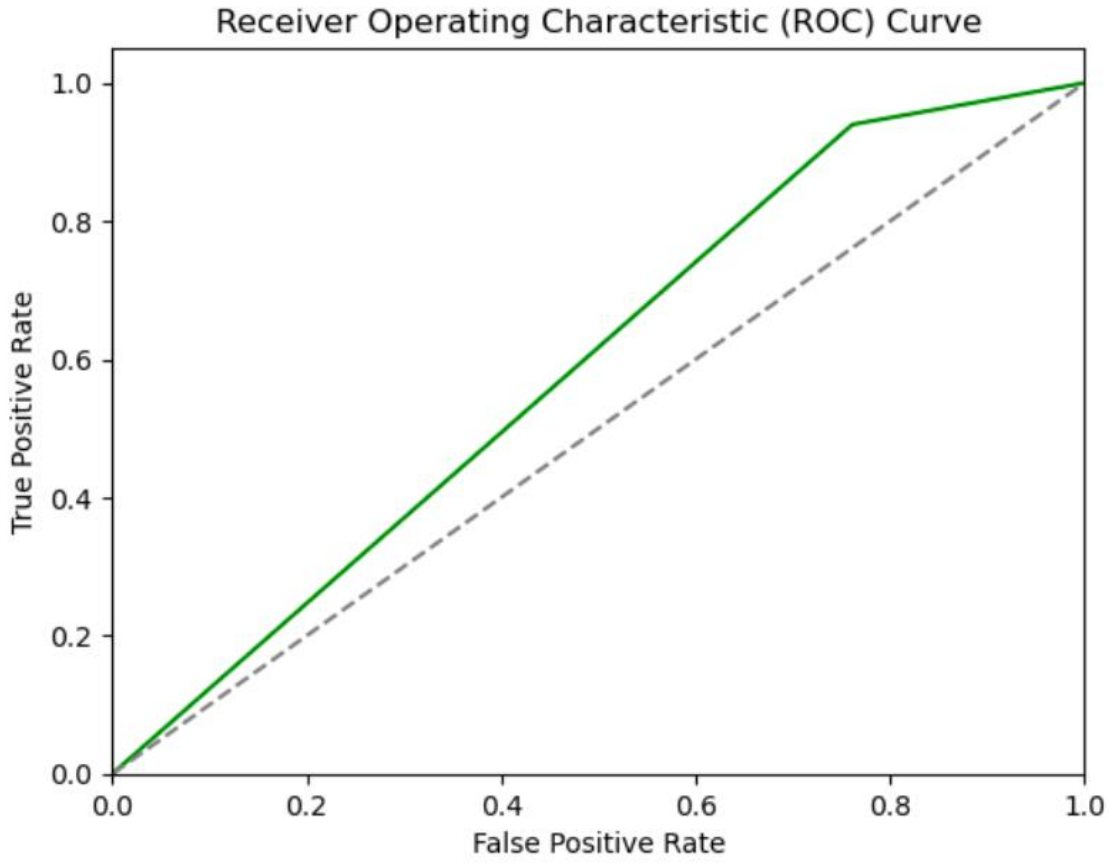
K-Nearest Neighbors (KNN) – Model Evaluation

Model	KNN
Accuracy	0.9418
Precision	0.9488
Recall	0.9912
F1 Score	0.9695
ROC-AUC	0.7951



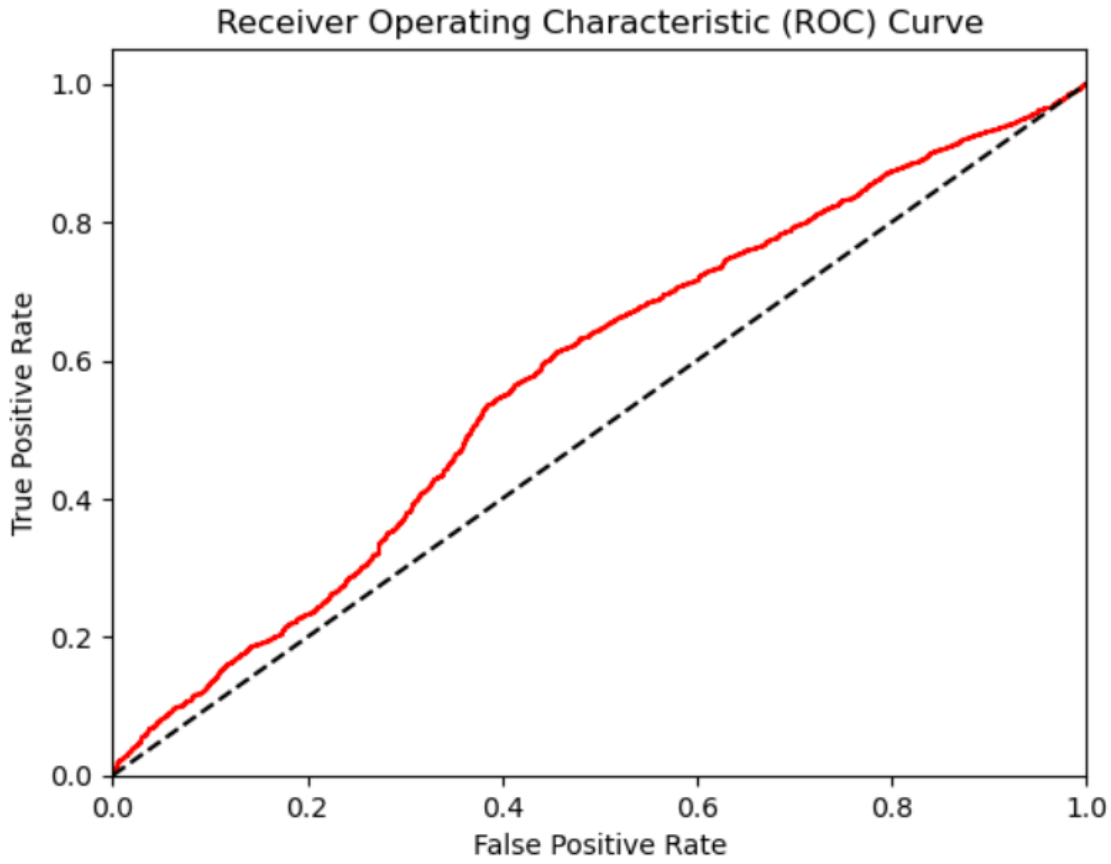
Decision Tree – Model Evaluation

Model	Decision Tree
Accuracy	0.8923
Precision	0.9443
Recall	0.9399
F1 Score	0.9421
ROC-AUC	0.5893



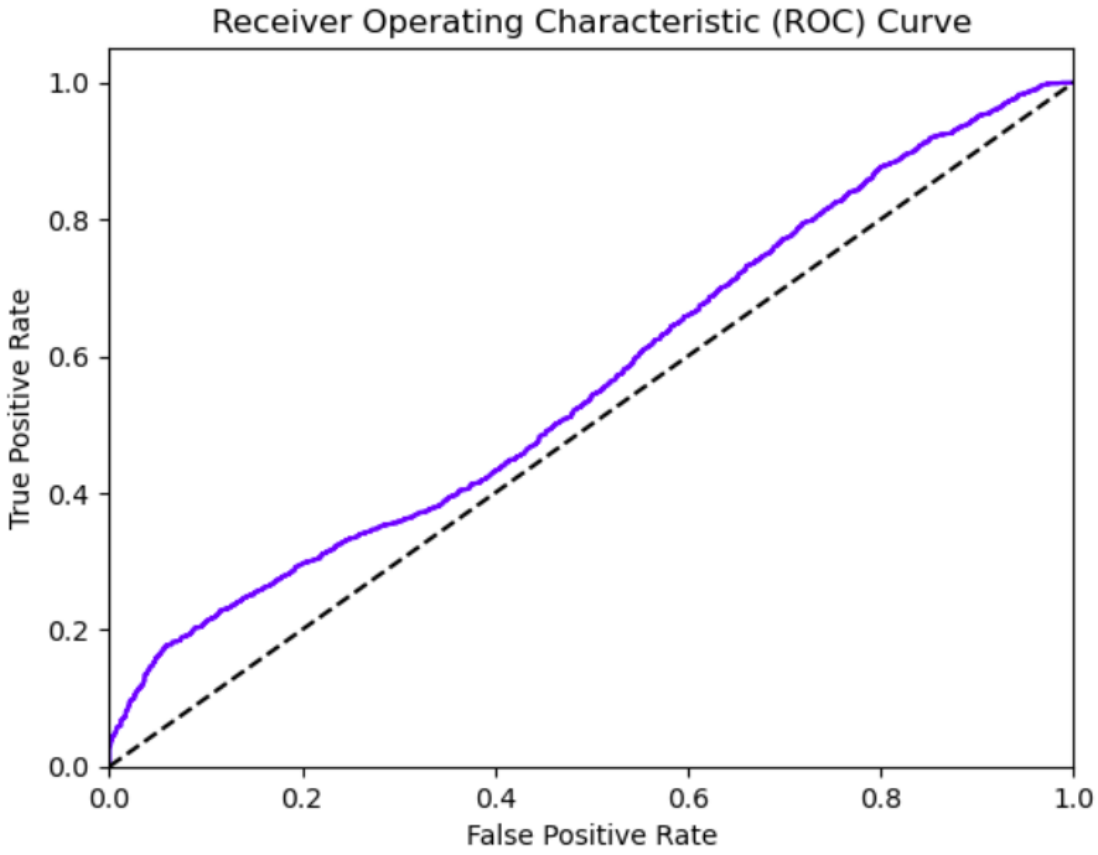
SVM (Linear Kernel) – Model Evaluation

Model	SVM(Linear Kernel)
Accuracy	0.9321
Precision	0.9321
Recall	1.0
F1 Score	0.9649
ROC-AUC	0.5748



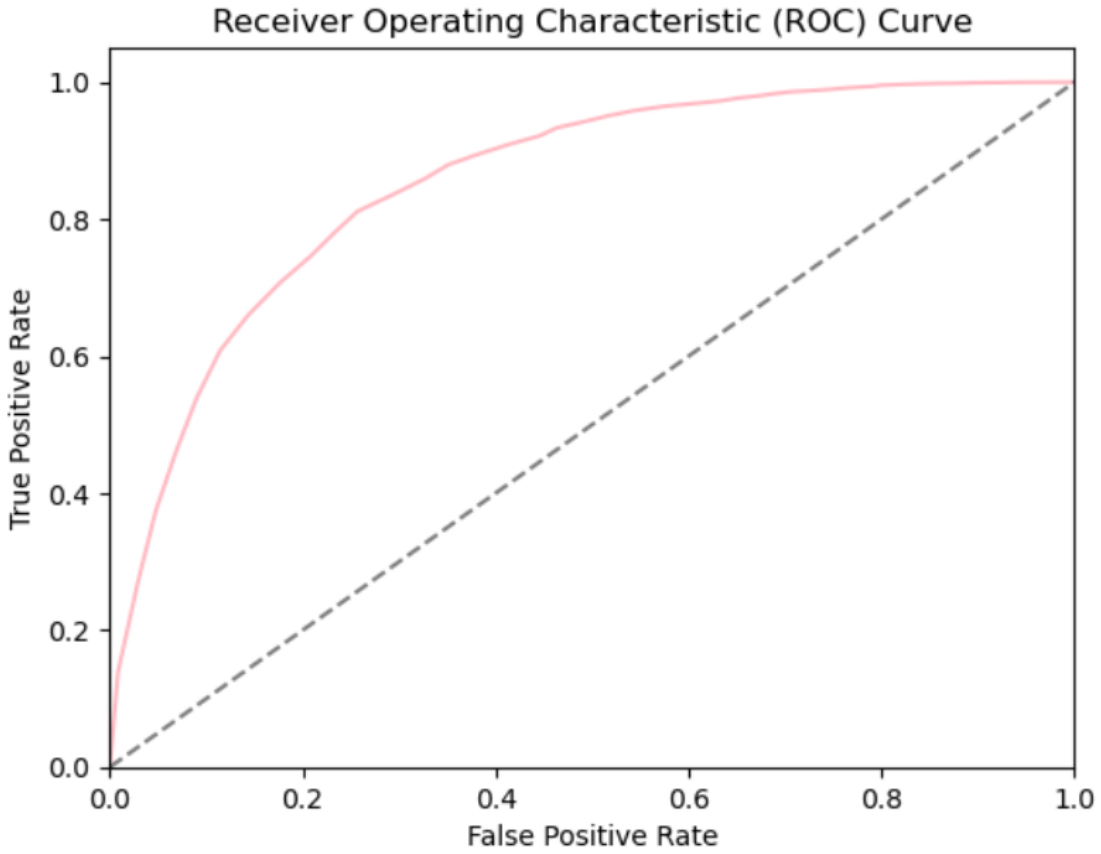
SVM (RBF Kernel) – Model Evaluation

Model	SVM (RBF Kernel)
Accuracy	0.9325
Precision	0.9324
Recall	1.0
F1 Score	0.9650
ROC-AUC	0.5630



Random Forest – Model Evaluation

Model	Random Forest
Accuracy	0.9364
Precision	0.9364
Recall	0.9996
F1 Score	0.9669
ROC-AUC	0.8541



Comparative Conclusion

Original Research Findings (IEEE 2024)

- **Best Model**: K-Nearest Neighbors (KNN)
- **Top Accuracy**: 94.41%
- **Random Forest** showed strong performance with the highest ROC-AUC (85.31%)
- **SVMs and Logistic Regression** had high recall but relatively lower ROC-AUC scores

Our Replication Findings

- **KNN** replicated as the most accurate model, matching the research outcome
- **Random Forest** consistently delivered the highest ROC-AUC, reinforcing its reliability.
- **Logistic Regression and SVMs** maintained similar trends: high recall, weaker AUC
- **Consistency Across Metrics**: Accuracy, recall, and AUC values followed the same patterns

Paper: Machine Learning Model for Detecting Money Laundering in Bitcoin Blockchain Transactions

- Objective: Identify suspicious (money laundering) transactions
- Dataset used in Paper: 2,906 instances from Kaggle, 24 attributes each.
- Dataset adopted for our study: Elliptic Data Set (203768 instances, 167 attributes each)
- Labels in Elliptic dataset: Illicit: 1, licit: 2, unknown: 0
- Models used:
  - K-Nearest Neighbors
  - Random Forest (Implemented an Additional Model)

Results of the Original Study

- KNN Model demonstrates high precision in flagging money laundering activity.
- Low error rate of only 2%.

Techniques Used:

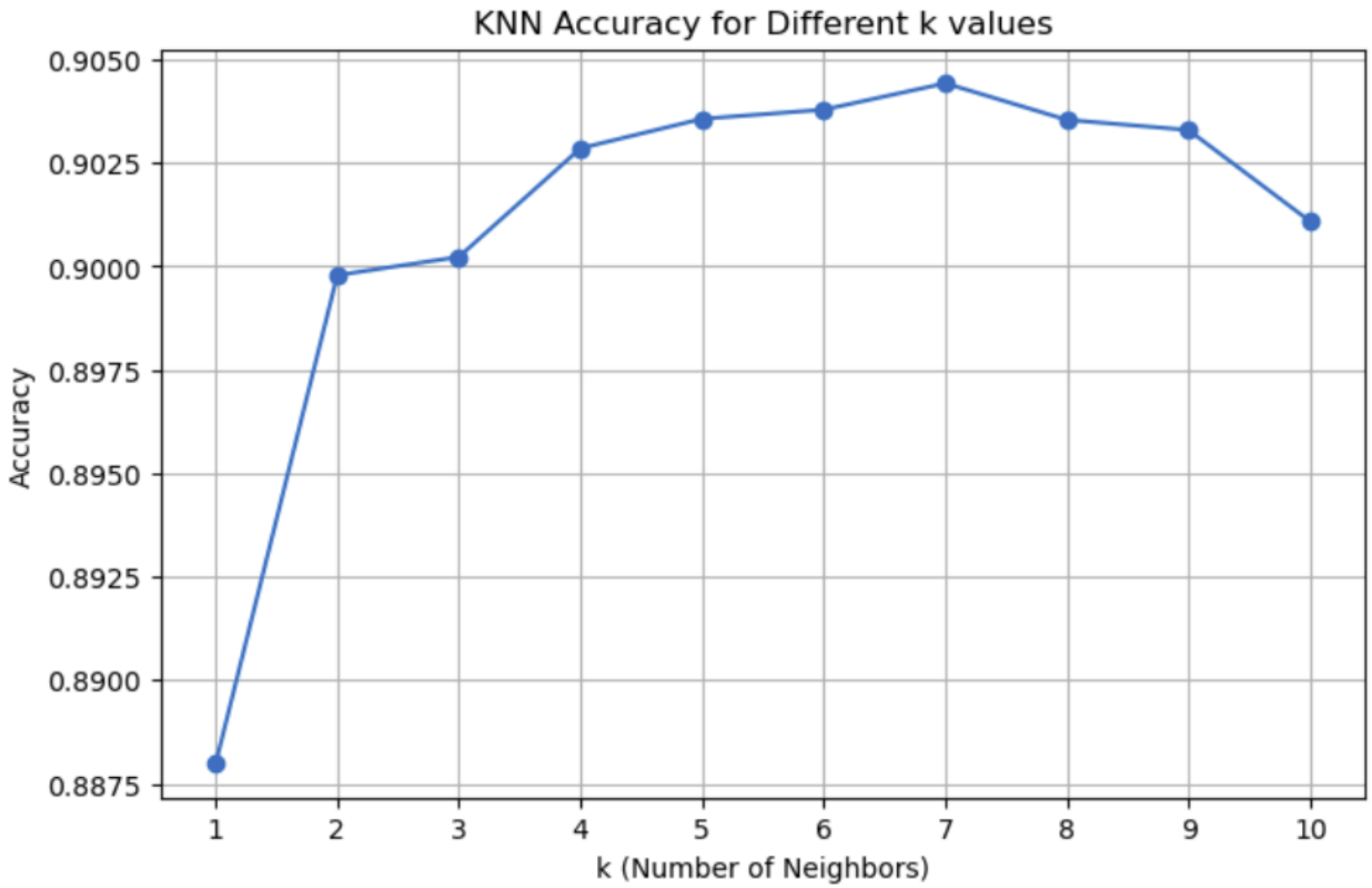
- *K-Means Clustering* → Unsupervised grouping of transactions
- *K-Nearest Neighbors (KNN)* → Classification based on clustering results

Model	KNN
Accuracy	0.99
Precision	0.97
Recall	1.0
F-Score	1.0
Error Rate	0.020

KNN Accuracy Evaluation

- ✓ The accuracy steadily increased up to k = 7, where it peaked at 0.9044, indicating the best performance.
- ✓ Beyond k = 7, the accuracy slightly declines, indicating potential over-smoothing.

```
Best k: 7 with Accuracy: 0.9044
```



Our Findings - Model Evaluation

Model	KNN	Random Forest
Accuracy	0.9044	0.9509
Precision	0.8202	0.9429
Recall	0.7431	0.8434
F1 Score	0.7764	0.8864
Error Rate	0.0955	0.0490

- Random Forest model outperformed KNN, achieving the highest accuracy at 95.09% and the lowest error rate at 4.90%.
- KNN also demonstrated strong performance with an accuracy of 90.44%.
- Random Forest is highly effective for high-dimensional classification tasks

Comparative Conclusion

Original Study (JAAI 2024 - Ogunleye et al.):

- **Models Used:**
  - **Unsupervised:** K-Means (to label data)
  - **Supervised:** K-Nearest Neighbors (KNN) only
- KNN was reported to be extremely accurate and robust.

Our Replication Findings:

- Using a much larger dataset, our replication suggests the superior performance of **Random Forest**, with the highest accuracy (0.9509) and lowest error rate (0.0490).
- **KNN** also performed well with 0.9044 accuracy closely aligning with the original study.

References

1. Ogunleye, O. S., Oyelade, O. N., & Adebayo, S. O. (2024). *Machine Learning Model for Detecting Money Laundering in Bitcoin Blockchain Transactions*. Journal of Advanced Artificial Intelligence (JAAI), 12(3), 144–159.

2. Vandana. (2012). *Machine Learning Techniques for Financial Fraud Detection: A Review*. International Journal of Computer Applications, 45(21), 25–30.

3. IEEE (2024). *Cognitive Modelling of Bankruptcy Risk Using Machine Learning Algorithms*. Proceedings of the IEEE Conference on Cognitive Computing, pp. 101–110.

4. Kaggle. (n.d.). *US Company Bankruptcy Prediction Dataset*. Retrieved from: https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset4

5. Elliptic. (2020). *Elliptic Dataset for Bitcoin Money Laundering Detection*. Retrieved from: https://www.kaggle.com/code/smlopezza/elliptic-data-set-eda-graphs-random-forest/input?select=elliptic_bitcoin_dataset