Springboard- Data Science Course

Capstone Project 3

Classification of AI Generated and Real Images using Unsupervised and Supervised Models

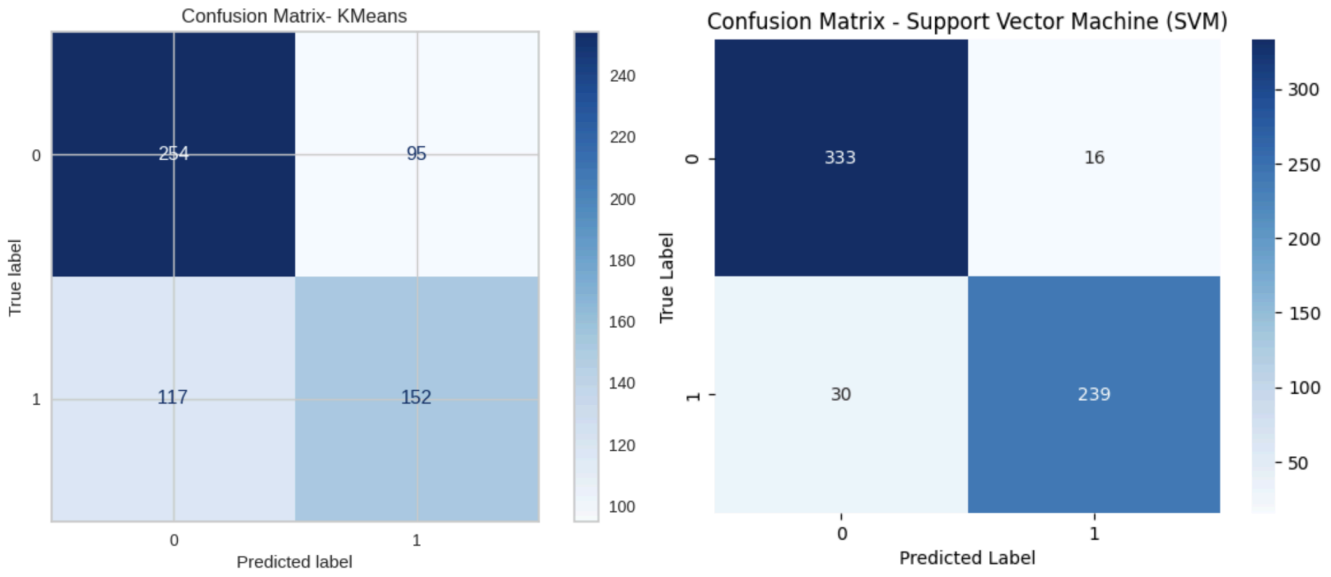
Priyanka Panhalkar
March, 2025

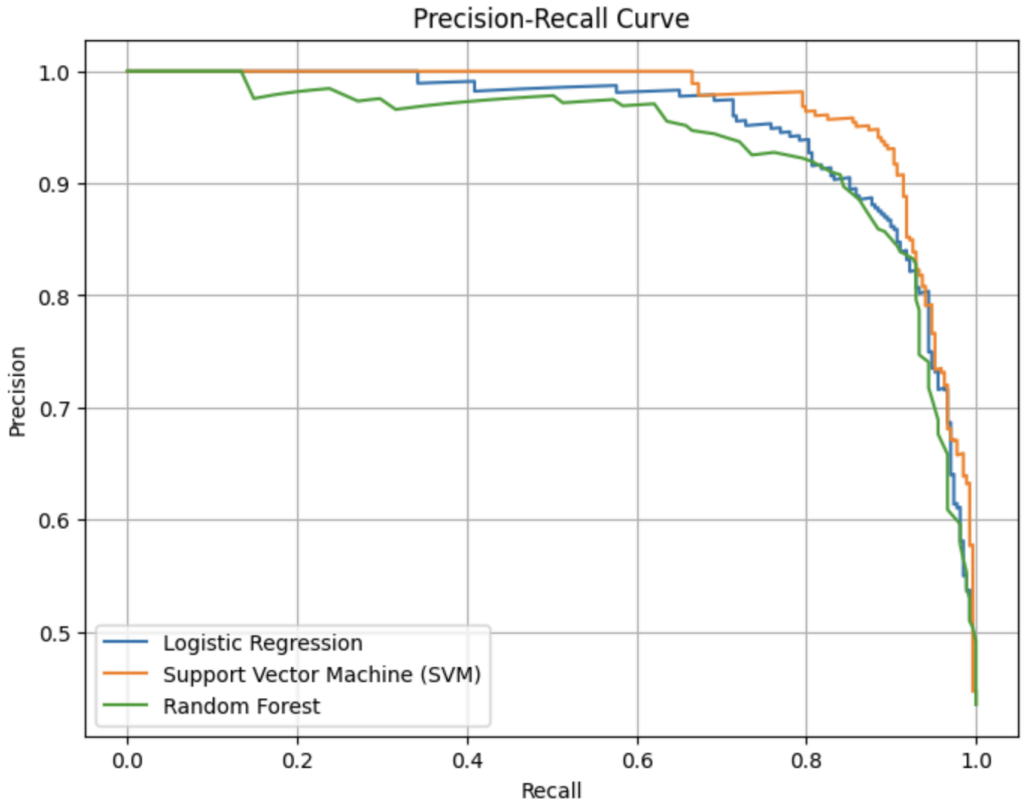
Introduction

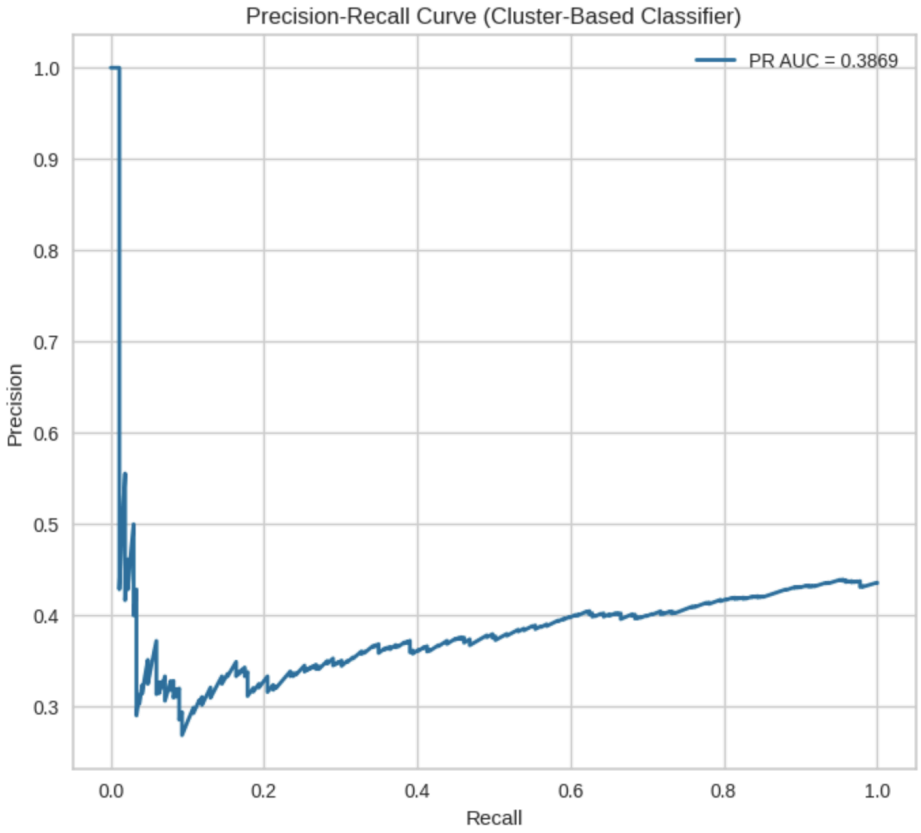
Business Problem: With the rise of Generative AI (GenAI), distinguishing between real and AI-generated images is challenging. This leads to issues like misinformation, fraud, and deep fake misuse. We aim to develop an AI model that automatically detects whether an image is real or AI-generated, improving trust, security, and transparency across industries.

Intended Stakeholders: Our intended stakeholders include tech companies who build AI detection tools,social media platforms for content filtering , researchers for media authenticating, e-commerce platforms for product images and many more.

Data Science Results: We compared Supervised and Unsupervised Models by using stratify method for train and test data. A **Support Vector Machine (SVM)** using CLIP embeddings emerged as the best model for this image classification task. It's reliable and generalizes well. **KMeans** with majority voting provided a decent unsupervised fallback, though performance lagged compared to supervised models. The implementation details for the project can be found here.




Precision-Recall Curve
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression
Support Vector Machine (SVM)
Random Forest


Precision-Recall Curve (Cluster-Based Classifier)
PR AUC = 0.3869
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

Approach

Data Acquisition and Wrangling: The image dataset was curated from an acceptable balanced set of **real** and **AI-generated** images. Each image was converted to a 512 dimensions embedding using the CLIP **transformer**, an OpenAI Convolutional Neural Network (CNN) Model.

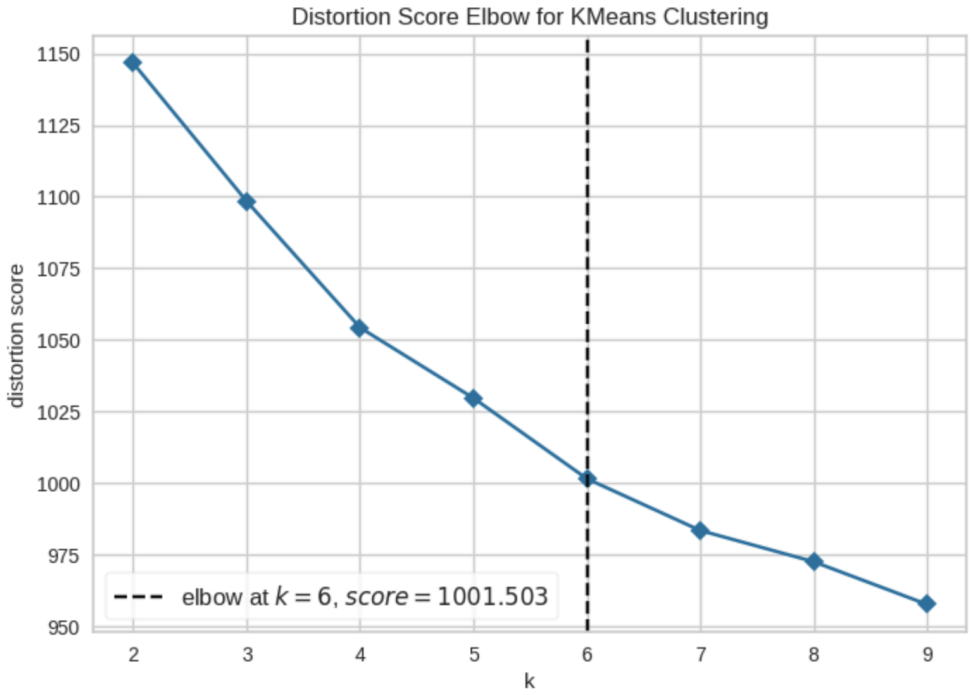
Steps performed for Data Exploring and Wrangling:

- We analyzed a curated dataset of 3,090 images from diverse categories like Mars, the Statue of Liberty, and the Taj Mahal. This included 1,748 real images and 1,342 AI-generated ones, forming the foundation for our classification analysis.
- Embeddings were parsed and normalized to unit L2 norm.
- Train/test sets were split using **stratified sampling** to preserve label balance.
- Embeddings were saved in `.csv` format for reproducibility.
- We generated multiple dot product heatmaps using normalized embeddings to explore the similarity patterns within AI-generated images, real images, and between the two groups.
- Dimensionality reduction using PCA (2D and 3D) revealed distinct clustering patterns in the embedding space, indicating separation between real and AI-generated images.
- 2D and 3D t-SNE visualizations further emphasized local structure and neighborhood similarities, reinforcing the suitability of these embeddings for downstream classification tasks.

Storytelling and Inferential Statistics:

Unsupervised Approach: We explored Unsupervised Clustering to better understand the structure and separability of real and AI generated images in embedding space.

1. Exploring Cluster Structure with the Elbow Method : We began by applying the **Elbow Method** on the training embeddings to determine the optimal number of clusters (K) for KMeans clustering. The plot of distortion scores showed a clear “elbow” at **K = 6**, suggesting that the dataset has underlying structure that can be grouped into six natural clusters.



This insight confirmed that our image embeddings, though high-dimensional, contain **compact regions of semantic similarity** — a promising signal for both clustering and classification.

2. Using Distance-to-Centroid for Label Inference: Building on this, we implemented a **post-clustering prediction strategy**:

- Each test embedding was assigned to the nearest cluster centroid using **Euclidean distance**.
- We then inferred the label of the test image by **majority voting** within that cluster based on the training labels.

This method served as an **unsupervised proxy classifier**, requiring no access to test labels during training — and yet producing meaningful predictions.

3. Inferential Visualization: Confusion Matrix & PR Curve: The **confusion matrix** revealed moderate classification ability, especially in distinguishing real images (True Negative = 254).

- The model struggled more with AI-generated samples, misclassifying a notable portion (117 false negatives).
- The **Precision-Recall Curve** had an AUC of **0.3869**, indicating weak—but non-random—separability using cluster-based inference.

These visualizations illustrated that:

- The embedding space **preserves semantic separability** between classes.
- Even **unsupervised** methods can capture meaningful relationships — reinforcing that the embeddings are highly informative.

4. Statistical Interpretation: The post-clustering prediction method provided evidence that **embedding-based similarity** can be leveraged for label inference without direct supervision. This supports the hypothesis that **AI-generated images tend to group together** in latent space, while real images exhibit broader variation — a pattern further supported by PCA and t-SNE in earlier analysis.

Supervised Approach: We began by reading the stratified training and testing datasets that were saved as CSV files to preserve class balance across splits. Each dataset contained image names, binary labels (0 = real, 1 = AI-generated), and high-dimensional CLIP-based embeddings. The current Models were trained with default settings and no hyperparameter tuning was performed.

Model Training: We trained and compared **three traditional supervised models**:

- **Logistic Regression**
- **Support Vector Machine (SVM)**
- **Random Forest Classifier**

Each model was trained on the same input space (L2-normalized embeddings) and evaluated on the test set.

Model Evaluation & Metrics**:** Each classifier was evaluated using the following:

- **Classification Report** (Precision, Recall, F1-Score)
- **ROC AUC Score**
- **Confusion Matrix for SVM Model**
- **Precision-Recall Curves**

Summary of Results:

Performance Summary of Supervised and Unsupervised Models. Below table shows Precision, Recall and F1 score for Real (0) and AI Gen (1) Images.

Model	Precision		Recall		F1 Score	
Logistic Regression	0.88	0.9	0.93	0.84	0.9	0.87
Support Vector Machine (SVM)	0.91	0.93	0.95	0.88	0.93	0.912
Random Forest	0.87	0.91	0.93	0.82	0.9	0.86
KMeans+ Majority Voting	0.68	0.61	0.72	0.56	0.7	0.58

Amongst all supervised models — especially **SVM with CLIP embeddings** — are highly effective for detecting AI-generated images. The high ROC AUC and F1-scores demonstrate **strong generalization and decision boundary clarity**, making this model suitable for real-world deployment.

Discussion of Results: We evaluated both **unsupervised** and **supervised** learning approaches to classify images as **real** or **AI-generated**, based on CLIP embeddings.

- Both **Logistic Regression** and **Random Forest** provided strong results, proving that even simple models can effectively leverage CLIP embeddings.
- **Support Vector Machine (SVM)** clearly outperformed all other models, achieving the **highest accuracy, F1 score, and ROC AUC** — making it the **best model** for this classification task.
- **KMeans clustering**, while useful for early exploration, underperformed due to its lack of label supervision — resulting in lower accuracy and poor PR AUC.

The original problem was to determine whether a given image is **AI-generated or real** — a concern with growing importance in various domains.
Our findings demonstrate that:

- **Embedding-based classification is highly effective**: The semantic richness captured in CLIP embeddings allows traditional models like SVM to make highly accurate distinctions.
- **Supervised learning is essential**: While unsupervised methods reveal useful structure, they fall short in prediction performance without labeled guidance.
- **The SVM model is deployment-ready**: With over **92% accuracy** and a **ROC AUC of 0.9685**, the SVM can be confidently used in real-world systems to flag suspicious or AI-generated content for further review.

Conclusions and Future Work

Conclusions: By combining semantic image embeddings with supervised modeling, we can accurately and efficiently detect AI-generated content providing a scalable and reliable solution to a timely and increasingly important business challenge.

Future Work:

1. Future iterations can explore **grid search**, **random search**, or **Bayesian optimization** to fine-tune parameters like:
   - C in SVM and Logistic Regression
   - n_estimators and max_depth in Random Forest
2. In Unsupervised Model training instead of Majority Voting (Post Clustering Labeling) we can also use weighted voting based on distance to centroid.
3. We can also fine-tune **vision transformers (ViTs)** or **EfficientNet** on top of the CLIP embeddings

Recommendations for the Clients

Based on the findings of our analysis, we propose the following **actionable recommendations .**

1. **Maintain an evolving Training Set:** We recommend periodic retraining of the model with newer data to maintain relevance and accuracy.
2. Since our SVM model has a good accuracy of 92.6% we can possibly deploy in pipeline as the primary content verification tool.

Consulted Resources

1. https://huggingface.co/docs/transformers/en/model_doc/clip
2. https://huggingface.co/openai/clip-vit-base-patch32
3. https://www.youtube.com/watch?v=EItlUEPCIzM&ab_channel=codebasics
4. https://www.youtube.com/watch?v=H99JRtDDnvk&ab_channel=KrishNaik