Aim & Objectives

The aim of this project is to analyze Udemy’s Finance & Accounting course dataset and build a predictive machine learning model that estimates the expected number of subscribers for a course based on its characteristics such as title, category, price, discount, ratings, and reviews.

Objectives:
1. Perform exploratory data analysis (EDA) to understand course trends and subscriber distribution.
2. Preprocess and clean the dataset (handle missing values, outliers, inconsistent values).
3. Engineer meaningful features from course metadata (e.g., title length, keyword flags, category mapping).
4. Build and train machine learning models to predict the number of subscribers.
5. Evaluate models using metrics such as MAE, RMSE, and R^2.
6. Deploy the trained model as:
- A Flask API for programmatic predictions.
- A Streamlit app for interactive user predictions.

7. Document insights into what factors influence course popularity.

Problem Statement

With the rise of online education platforms like Udemy, predicting the popularity of courses has become an important task. Popularity, measured by the number of subscribers, depends on various factors such as course title, category, price, discount strategy, number of reviews, and ratings.

However, it is not straightforward to estimate how many subscribers a new course might attract.

This project aims to solve the problem by building a machine learning model that can predict the number of subscribers for a course using its metadata. This will help:
- Instructors optimize their course design and pricing strategy.
- Udemy identify potentially popular courses.
- Learners gain insights into which course attributes drive popularity.

Methodology

1. Data Collection
- Dataset: Udemy Finance & Accounting courses (13K+ courses).

2. Data Preprocessing
- Cleaned course titles (lowercase, removed punctuation).
- Created course categories using keyword mapping.
- Handled missing values using median imputation.
- Engineered features like title length, word count, and discount percentage.

3. Exploratory Data Analysis (EDA)
- Distribution of subscribers, ratings, and reviews.
- Relationship between price, discount, and subscribers.
- Popular categories based on title keywords.

4. Feature Engineering
- Numeric features: ratings, reviews, lectures, practice tests, prices.
- Categorical feature: category (One-Hot Encoding).
- Text feature: course title (TF-IDF vectorization).

5. Model Building
- Baseline: RandomForestRegressor.
- Pipeline with preprocessing + model.
- Target transformed using log(1 + subscribers) to reduce skew.

6. Model Evaluation
- Metrics: MAE ≈ 2074, RMSE ≈ 6257, $R^2 \approx 0.67$.

7. Deployment
- Flask API and Streamlit app for predictions.

Results

Model Performance:
- MAE: ~2074 subscribers
- RMSE: ~6257 subscribers
- R^2: ~0.67

Interpretation:
- The model explains about 67% of the variation in subscribers.
- Predictions are, on average, off by around 2000 subscribers.
- Large errors occur for very popular courses (outliers).

Insights:
- Keywords like "excel", "sql", "data science" often appear in high-subscriber courses.
- Higher discounts generally correlate with more subscribers.
- Categories like "Finance", "Data Science", and "Spreadsheet" dominate the dataset.

Conclusion & Future Work

Conclusion:
This project successfully demonstrated how machine learning can be applied to predict course popularity on Udemy. Using metadata such as titles, ratings, reviews, and price details, we built a predictive model with $R^2 \approx 0.67$.

Future Work:
1. Experiment with advanced models (XGBoost, LightGBM, CatBoost).
2. Use deep learning embeddings (e.g., BERT) for titles.
3. Add more metadata (instructor reputation, course duration, promotions).
4. Try classification models (popular vs. not popular).
5. Deploy on cloud (Heroku, AWS, GCP) for wider access.