**Top machine learning project for final year with dataset and source code**

[[](https://naemazam.medium.com/?source=post_page-----f149017e794f--------------------------------)](https://naemazam.medium.com/?source=post_page-----f149017e794f--------------------------------)

[Naem Azam](https://naemazam.medium.com/?source=post_page-----f149017e794f--------------------------------)

·

Follow

12 min read

·

Jan 1

274

5



Top machine learning project for final year with source code

Machine Learning Projects are the key to understanding the real-world implementation of machine learning algorithms in the industry. These machine learning projects for students will also help them understand the applications of machine learning across industries and give them an edge in getting hired at one of the top tech companies. A resume with one or some ML projects (listed below) will boost students’ opportunities and make their resumes stand out from the pile of resumes. Every final-year student interested in pursuing a career in data science or machine learning must work on a hands-on project to experience a practical approach to how machine learning models are implemented and deployed in production.

**1. Recommender System Projects**

Have you ever seen movies or web series on online streaming platforms? Once you watch one or two of them, you will notice that apps like Netflix and Amazon Prime recommend new web series and movies. It is because these apps render machine learning models that try to understand the customer’s taste. Modern e-commerce sites like Flipkart, Amazon, Alibaba, etc., also have the same feature. Recommendation engines are popular in media, entertainment, and shopping. All modern apps come with a recommendation engine that suggests users for more engagement.

You can use libraries like recommenderlab for testing and developing the recommendation models for your ML project. Apart from that, libraries like ggplot, reshape2, data.table will complement your machine learning project. Datasets like Google Local, Amazon product reviews, MovieLens, Goodreads, NES, Librarything are preferable for creating recommendation engines using machine learning models. They have a well-researched collection of data such as ratings, reviews, timestamps, price, category information, customer likes, and dislikes.

GitHub Link => [Here](https://github.com/microsoft/recommenders)

**2. Sales Forecasting Project**

Big B2C marts and retailers want to predict the sales demand for each product present in their inventory. Sales forecasting helps business owners get a clear idea of what products are in demand. Accurate sales forecasting will reduce wastage to a significant level and determine the incremental impact on future budgets. Retailers like Walmart, IKEA, Big Basket, Big Bazaar leverage sales forecasting for sale predictions of product requirements.

To build such ML projects, you must know different approaches to cleaning raw data. Also, must have a thorough understanding of regression analysis especially, simple linear regression. You have to use libraries like Dora, Scrubadub, Pandas, NumPy, etc., for developing these kinds of projects. Dummy datasets like univariate time-series datasets, shampoo sales datasets, etc., can help you model such machine learning projects.

GitHub Link => [Here](https://github.com/mohan-aditya05/walmart_sales_forecasting)

**3. Stock Price Prediction Project**

Creating a stock price prediction system using machine learning libraries is an excellent idea to test your hands-on skills in machine learning. Students who are inclined to work in finance or fintech sectors must have this on their resume. Nowadays, many organizations and firms lookout for systems that can monitor, analyze and predict the performance and stock price. There is a broad spectrum of data available on finance and the stock market. For this reason, the final year students find it a hotbed of opportunities. Before you start working on this project, you must have proficiency in the following areas:

a. Statistical modeling: Here, the student has to prepare a real-world model of the mathematical representation and the statistics uncertainties within the analyzing and prediction process.

b. Regression analysis: This technique talks about the predictive methods that your system will execute while interacting between dependent variables (target data) and independent variables (predictor data).

c. Predictive Analysis: This analysis will utilize data mining, web scraping, and data exploration techniques for better prediction and accurate analysis.

For analyzing statistical data and handling data clusters in the stock price prediction project, you must use libraries like Sklearn, SciPy, Pandas, and SciPy. If you want to visualize the data, you can use Seaborn or Matplotlib. For monitoring and visualizing analyzed stock price and stock market data, you can use Tableau. For training the machine learning model, you can use the NSE-TATA-GLOBAL dataset. It contains all the attributes you need to build your stock price prediction system.

GitHub Link => [Here](https://github.com/venky14/Stock-Market-Analysis-and-Prediction)

**4. Build a Sorting, Categorizing, and Tagging System**

You can create crowd-sourced software systems that allow categorizing, sorting, and tagging various forms of data (structured or unstructured). Such data will contain photos, reviews, rating, location information covering everything from hotels, restaurants, gyms, clinics, offices, clubs, sports-hub, etc. Many organizations like Yelp.com collect and provide crowd-sourced data of all local businesses. In such projects, machine learning models are used to sort, categorize and tag millions of such local-business datasets so that they can sell those data to other corporate giants. Creating such a system will make it easy for consumers and other firms to find relevant spots instead of riffling over each of them.

Developing such ML projects requires an in-depth understanding of image clustering, classification, computer graphics, and data analysis. You have to use libraries like OpenCV, Scikit-Image, PIL (Python Imaging Library), NumPy, Pandas, Mahotas, etc. Machine Learning frameworks like Scikit-learn and TensorFlow can help you in this project. This type of application is beneficial in searching, mobile application development, and social media apps. You can download the Yelp dataset that has around 8,635,403 reviews from 160,585 businesses with 200,000 pictures.

GitHub Link => Here

**5. Patient’s Sickness Prediction System**

Machine learning has been proven effective in the field of healthcare also. Traditional healthcare systems became increasingly challenging to cater to the needs of millions of patients. But, with the advent of ML, the paradigm shifted towards value-based treatment. Every modern healthcare equipment and gadget comes with internal apps that can store patient’s data. You can leverage these data to create a system that can predict the patient’s ailment and forecast the admission. KenSci is an AI-based solution that can analyze clinical data and predict sickness along with more intelligent resource allocation.

You can use open-source medical datasets like CHDS (Child Health and Development Studies), HCUP, Medicare to test your machine learning algorithm. You can incorporate such projects in healthcare wearables, telemedicine, remote monitoring, etc. To develop such algorithms, you need to have a thorough understanding of the following:

a. Classification & Clustering model: While classification determines the categorization of data, clustering in ML looks for distinctive patterns in the data when the data available does not have a definite outcome.

b. Regression analysis: Its principal purpose is to find value. This technique talks about the predictive methods that your system will administer while interacting between dependent variables (target data) and independent variables (predictor data).

Use libraries like NumPy, Pandas, Matplotlib, Theano, etc., and frameworks like Keras and Hugging face to implement this ML project.

GitHub Link => [Here](https://github.com/NYUMedML/GNN_for_EHR)

**6. AI-driven Sentiment Analyzer**

Social media is swinging with millions of user-generated posts and content. Most users go to social media to convey their personal views, opinions and share feelings through images and words. But there lies the biggest challenge in accurately understanding the feelings or sentiments behind such user-generated posts. Modern social media firms like Snapchat, Facebook, Linked In, Twitter, etc., and online food delivery services are spending large budgets on projects to understand human sentiments and feelings. Analyzing the texts and images to understand the sentiments would make these firms recognize user behavior easily. Based on such analysis, companies can cater to improved customer service, hence increasing customer satisfaction.

Sentiment analysis projects require an in-depth understanding of topics like text analysis, NLP, and computational linguistics. Machine learning frameworks like Huggingface and TensorFlow become handy and supervised algorithms like neural networks, Random Forest, decision trees, Support Vector Machines (SVM), and logistical regression come into use. Libraries like OpenCV and SimpleITK help in image segmentation, image registration, and object detection.

GitHub Link => [Here](https://github.com/imjoseangel/sentiment-analysis-es)

**7. Email Spam-Filtering System**

Mining text is one of the popular computation techniques widely applied in applications like text summarization, topic classification, machine translation, sentiment analysis, etc. Modern cybersecurity systems are utilizing machine learning methods a lot. Spam email detecting systems are one of them. Spam filtering also leverages text mining and document classification to segregate legitimate mails and spam emails. All modernized email services come with this segregation system that runs machine learning algorithms behind. Such a project comes under the text classification problems. Building this kind of a ML project involves the following important steps -

a. Text Processing

b. Text Sequencing

c. Model Selection

d. Implementation

Use libraries like Sklearn, NumPy, Counter, Scrubadub, Beautifier, Seaborn, and machine learning frameworks like TensorFlow and Keras. For training such machine learning models, a Spambase dataset can help you. Spambase dataset is an open-source UCI machine learning repository comprising around 5569 emails, of which nearly 745 are spam emails.

GitHub Link => [Here](https://github.com/nishi1612/Email-Spam-Classification-using-SVM)

**8. Digit Classification Project using MNIST Dataset**

The digit classification project is a remarkable machine learning project that employs neural network and machine learning concepts. From the outset of machine learning, it was challenging to work with unstructured data (image dataset) and transform it into structured data (texts). In this project, you will use Convolutional Neural Networks (CNNs) to train machine learning algorithms. The MNIST dataset will make the training of ML models seamless so that your system can recognize handwritten digits easily.

This type of project comes under the domain of computer vision. To build this project, you have to use libraries like NumPy, PIL, Pillow, Scikit-image, Tkinter, etc. Tkinter will help in developing a GUI interface for your application. All the neural network algorithms will get managed by Keras. You have to use frameworks like TensorFlow and Keras. You can train your model with the MNIST dataset having sixty thousand images containing handwritten digits from zero to nine for training and around ten thousand images for testing purposes.

GitHub Link => Here

**9. Credit Card Fraud Detection Project**

Finding anomalies in credit cards and detecting fake credit card transactions was challenging before the advent of machine learning. This project will approve and differentiate swindling credit card transactions from legitimate ones. This project will make you discover and learn how to perform the classification of data. Also, you have to be thorough with the concepts like decision trees, artificial neural networks (ANN), logistic regression, and gradient-boosting classifiers. You can implement the credit card fraud detection project using libraries like NumPy, Pandas, Matplotlib, Seaborn, XGBClassifier, and frameworks like Scikit-Learn. The[credit card dataset](https://datahub.io/machine-learning/creditcard#readme) and[credit-card fraud detection dataset](https://www.kaggle.com/mlg-ulb/creditcardfraud) are preferable to train your ML model for such a project. These datasets have credit card details along with dummy data of fraudulent and non-fraudulent transactions.

GitHub Link => [Here](https://github.com/benedekrozemberczki/awesome-fraud-detection-papers)

**10. Fake News Detection Project**

It is another innovative machine learning project for final year students. As you all know, fake news is speeding like a bushfire. Right from connecting people to reading the daily news, everything is available on social media. Hence, it has become more skeptical these days to detect fake news. Many popular social media platforms like Facebook and Twitter already have fake news detection algorithms running behind the scenes on posts and feeds. Implementing this kind of ML project requires good know-how of [various NLP techniques](https://www.projectpro.io/article/10-nlp-techniques-every-data-scientist-should-know/415) and classification techniques (PassiveAggressiveClassifier or Naive Bayes classifier) to detect fake news. PassiveAggressiveClassifier is an online learning algorithm that remains passive while discovering correct classification outcomes. You can prefer supervised learning models to develop such projects.

Libraries like NumPy, Pandas, Itertools, and spaCy (for NLP tasks) will become handy for such a project. Apart from that, frameworks like Scikit-Learn and Streamlit will help a lot. Scikit-Learn has different machine learning approaches and statistical modeling for clustering, classification, and regression. Streamlit helps in building web applications more efficiently and quickly and includes the deployment of ML models. Also, the[Great Fake News dataset](https://github.com/derevirn/gfn-detector/tree/main/data) and[ISOT Fake News Dataset](https://www.uvic.ca/engineering/ece/isot/datasets/fake-news/index.php) are the best for training this ML model.

GitHub Link => [Here](https://github.com/nishitpatel01/Fake_News_Detection)

**11. Sign language Recognizer**

A lot of research and advancement is going on in technology to help individuals who are deaf and dumb. The progress in this particular domain is using machine learning together with neural networks and computer vision. Communicating in sign language is not a common language or expression. Hence, choosing this as your final year project will make you stand out from the rest. The most basic research you need to do while developing this project is the concepts of sign language and understanding these signs. Your project will detect the signs and extract out the meaning for others. You will be using the various concepts of NLP, computer vision (CV), and data prediction. Libraries like NumPy, OpenCV, SimpleITK, etc., and frameworks like Keras and TensorFlow will come in handy. Since different countries have different ways of representing signs, the dataset varies from country to country. The[sign language recognition dataset](http://facundoq.github.io/guides/sign_language_datasets/slr) has a sign language collection from 18 separate countries. This dataset of images can help your ML algorithm train over a large set of signs.

[GitHub Link](https://github.com/hthuwal/sign-language-gesture-recognition)

**12. Speech Emotion Recognizer**

It is another advanced-level ML project where you have to deal mostly with audio data. Speech emotion recognition attempts to identify and interpret user emotions and deduce the emotional state from speech. To train such an algorithm, you have to use most of the training data as audio data. This system will take user speech as input. Apart from general Python libraries like NumPy, Pyaudio, and Soundfile, this project will also utilize Librosa. Librosa is a Python library that helps to analyze audio data and music files. You also need Scikit-learn to build the project model implementing the MLPClassifier (Multi-layer Perceptron classifier). You can use the JupyterLab web-based UI to develop this project.

This project requires the[RAVDESS dataset](https://zenodo.org/record/1188976#.YQ5EvogzaMo) to train the machine learning model with different audio. The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset contains 7356 files with voice samples of 24 professional actors (12 female, 12 male). It comprises different speech intensities like happy, calm, angry, sad, surprise, fearful, and disgust expressions, along with songs with diverse emotions like sad, happy, frightful, peaceful, and fierce.

GitHub Link => [Here](https://github.com/MiteshPuthran/Speech-Emotion-Analyzer)

**13. Music Genre Classification System**

It is an advanced machine learning project that leverages deep learning concepts to classify music files based on different genres. Many music companies like Spotify, Gaana use such algorithms in their app to understand the listening preferences of the user. You can also select a specific genre for your customized playlist through this project. It uses deep learning algorithms to classify the list of songs to the smartphone user. To develop this project, you need to convert the audio signals into compatible formats. You need to have an understanding of two important concepts — Spectrogram generation and Wavelet generation.

This project also requires K-Nearest Neighbor Algorithm (KNN), Convolution Neural Network (CNN), and Support Vector Machines (SVM) to develop the CNN model. Then we have to pass the spectrogram & wavelet data in that CNN model for multimodal training. This project will demand an audio dataset. So, you can use the[GITZAN dataset](http://opihi.cs.uvic.ca/sound/genres.tar.gz) that comprises 1000 music files. Also, this dataset has ten different varieties of genres (blues, hip-hop, metal, classical, disco, country, pop, rock, jazz, and reggae) included with uniform distribution. Each of these files has a length of thirty seconds.

GitHub Link => Here

**14. Intelligent Chatbots**

Chatbots are intelligent systems or a program that can communicate and assist users similar to that of humans. All modern companies develop chatbots and integrate them into their web apps or smartphone apps to solve user queries and FAQs. Developing chatbots have become the need of the hour with almost 9 out of 10 companies using them for enhanced customer experience. So, creating an ML-based chatbot from scratch will be an excellent project for the final year students. There are two basic kinds of chatbot models:

· Retrieval Based models and

· Generative Based models

This project will require concepts like Natural Language Processing and artificial neural networks. In this project, you will need libraries like JSON, Natural Language Toolkit (NLTK), Pickle, and frameworks like TensorFlow and Keras. Chatbots require different types of datasets like question-answer datasets, conversation datasets, logical reply-based datasets, etc. Some well-known datasets you can use to train your chatbot are:

Among them, NPS Chat Corpus comes as a component of Natural Language Toolkit (NLTK) distribution. Stanford Question Answering Dataset (SQuAD) is another comprehension-based dataset. It has a collection of well-researched questions professed by crowd workers that can help build an effective training model for your chatbot project.

GitHub Link => Here

**15. Image Caption Generator**

It is one of the trending projects for final year students that harnesses deep learning algorithms and techniques. It uses Convolutional Neural Networks (CNN) and Long Short-term Memory (LSTM) networks to build the algorithm that can produce captions depending on the given image. This type of project demands proficiency in concepts like Natural Language Processing and Computer Vision. This project will use computer vision to understand the delivered image and identify its context. Then, based on the context, it will describe that image leveraging the NLP. This type of project is helpful for understanding user motive, detecting image-based conversations, and understanding what users want to convey via images. Companies like Snapchat use such algorithms to understand user moods and sentiments.

To implement this project, you need to use libraries like NumPy, Matplotlib, Scipy, OpenCV, Scikit-Image, Python Imaging Library (PIL), and Pgmagick. It also uses ML frameworks such as Keras and Scikit-learn to develop a convolutional neural network (CNN) as an encoder and recurrent neural network (RNN) as a decoder. There are three popular datasets:[Flickr8k](https://www.kaggle.com/adityajn105/flickr8k/activity),[Flickr30k](https://metatext.io/datasets/flickr30k-entities), and[MS COCO Dataset](https://cocodataset.org/#home) to train such algorithms. Flickr8k & Flickr30k datasets serve the purpose of image captioning, while MS COCO improves object detection, identification, and segmentation.

GitHub Link => [Here](https://github.com/dabasajay/Image-Caption-Generator)

???? Let’s be friends! Follow me on [Twitter](https://twitter.com/naemazamankon) and [FaceBook](https://www.facebook.com/naemazamchowdhury) and connect with me on LinkedIn. You can visit My [website](http://www.naemazam.com/) Too . Don’t forget to follow me here on [Medium](https://naemazam.medium.com/top-machine-learning-project-for-final-year-with-dataset-and-source-code-38cf9f986ec0) as well for more technophile content.

*Originally published at*[*http://www.naemazam.com*](http://www.naemazam.com/post/54-top-machine-learning-project-for-final-year-with-dataset-and-source-code)*.*