NO TIME: Rest of the data story

Neural Networks: The Data Brainiac

Neural networks are the overachievers of machine learning. Inspired by how our brains work, they process data through layers of "neurons." Each neuron takes in inputs (features like runtime, budget, or actor count), applies weights and biases, and passes the result to the next layer.

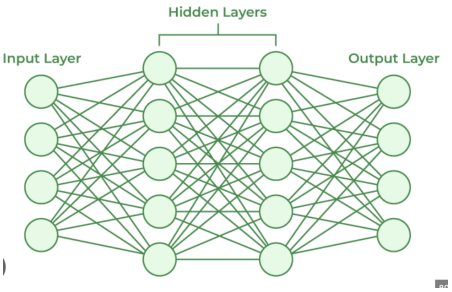
- **Input Layer:** This is where all your movie data comes in.
- **Hidden Layers:** These are the magic zones where the network learns complex patterns. The more layers, the more abstract patterns it can detect.
- **Output Layer:** This gives probabilities of the movie to be for each genre. For example, "60% sure is comedy, 30% sure is drama, 10% sure is action..."

Why is this powerful? Neural networks can find relationships in the data that simpler models might miss. For instance, they might learn that movies with "high budgets + long runtimes + low vote counts" are often historical dramas or something even more complex!.

However, they are not that easy to train, neither to interpret — so we think of them like a black box of genius. We are going to use two kind of Neural Networks. One on a box-like shape and one with a U-like shape.

The fist one has more neurons and so is able to get more complex relations between the features. Meanwhile, the U-shape one “compresses” the data in fewer neurons so the information can be mixed. Then, both rescale the layers into the right size to make the predicition.

Why Use These Models Together?



Each model has its strengths:

- Decision trees are easy to interpret and explain.
- Random forests are robust and accurate.
- KNN is simple and doesn’t assume much about the data.
- Regression is efficient and works well for simpler patterns.
- Neural networks are powerful and capture complex relationships.

By trying multiple models, we get a better sense of what works best for predicting genres. Plus, it’s fun to watch them compete!

How Do We Compare Their Performance?

The **F1 Score** is our referee. It balances precision (*how many predictions are correct) and recall (how many relevant genres are identified*). A high F1 score means the model is both accurate and comprehensive — perfect for genre classification.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

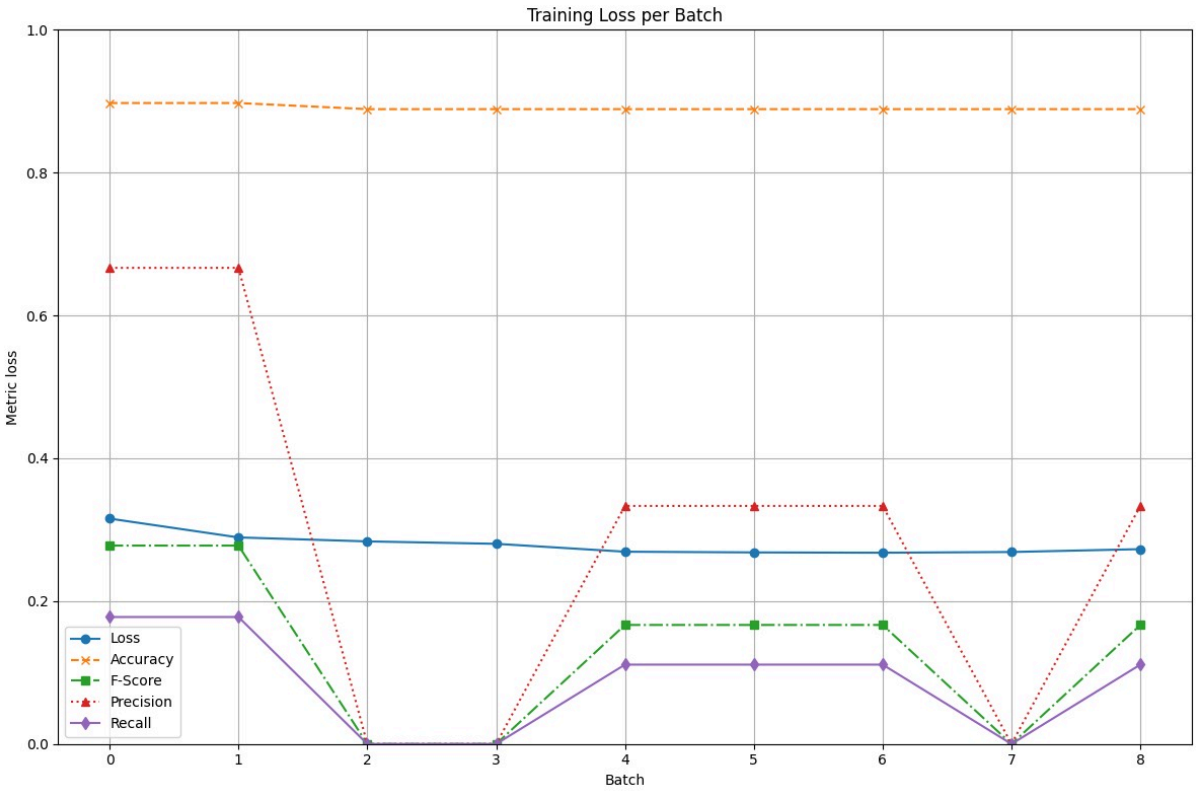
Fine tuning the models

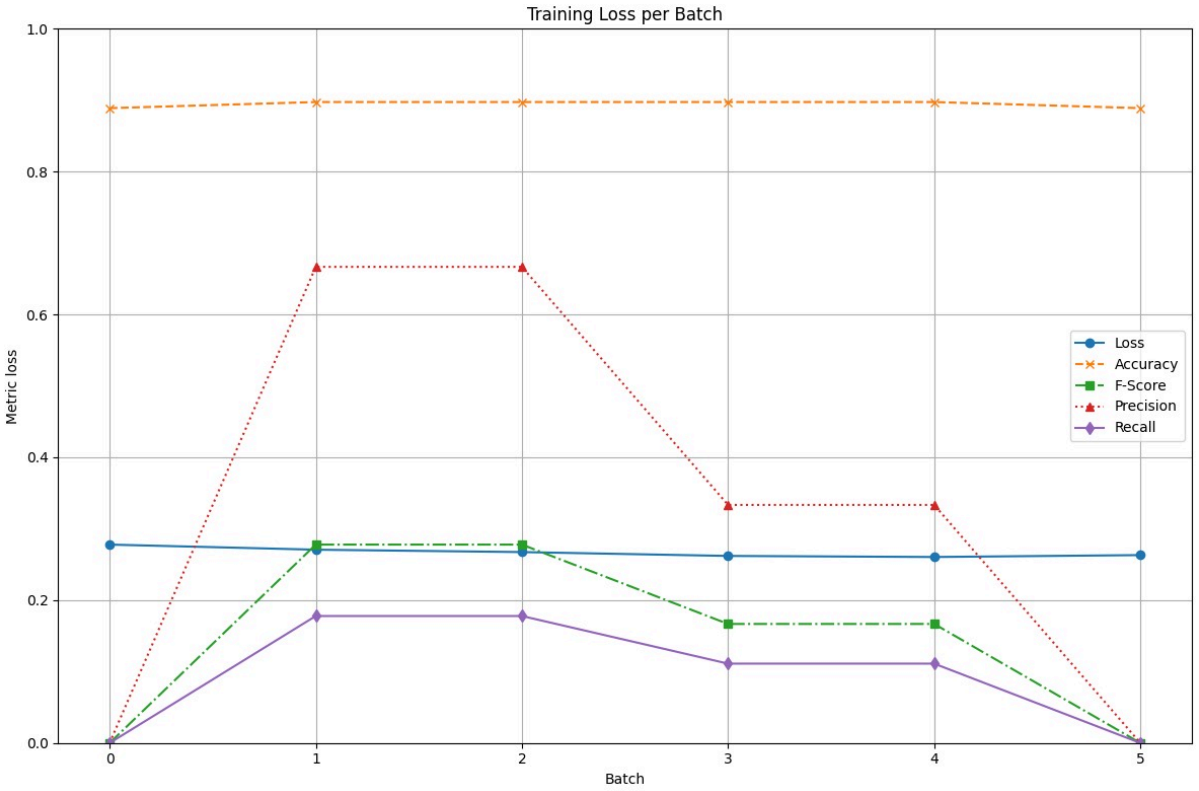
Default settings may be good, but sometimes we need to look closer to see which ones make our models work better.
The first thing we do is grid search of all the possible features of the models. One by one we try which option is better with the other settings fixed. At the end we choose the best combinations.
Also, we do a feature selection process where, sadly, we discard some before training the models. Why? Well, even if our big boy look big, they may get overwhelmed by all the data, overfit, or not focus on the important parts. We remove some data that may not be that useful and in the process we make our models fasters (less computations, less time thinking).

At the end, we removed features like 'tragedy', 'betrayal', or 'fear' from the plot topics. Why? as we seen before some of those share a words that may be just useless for the classification. Also 'N/A actor count' is a feature that comes from the lack of information in the dataset. Like those, we end up finding more. It is sad but we have to say bye to them in order to achieve our goal :,(.

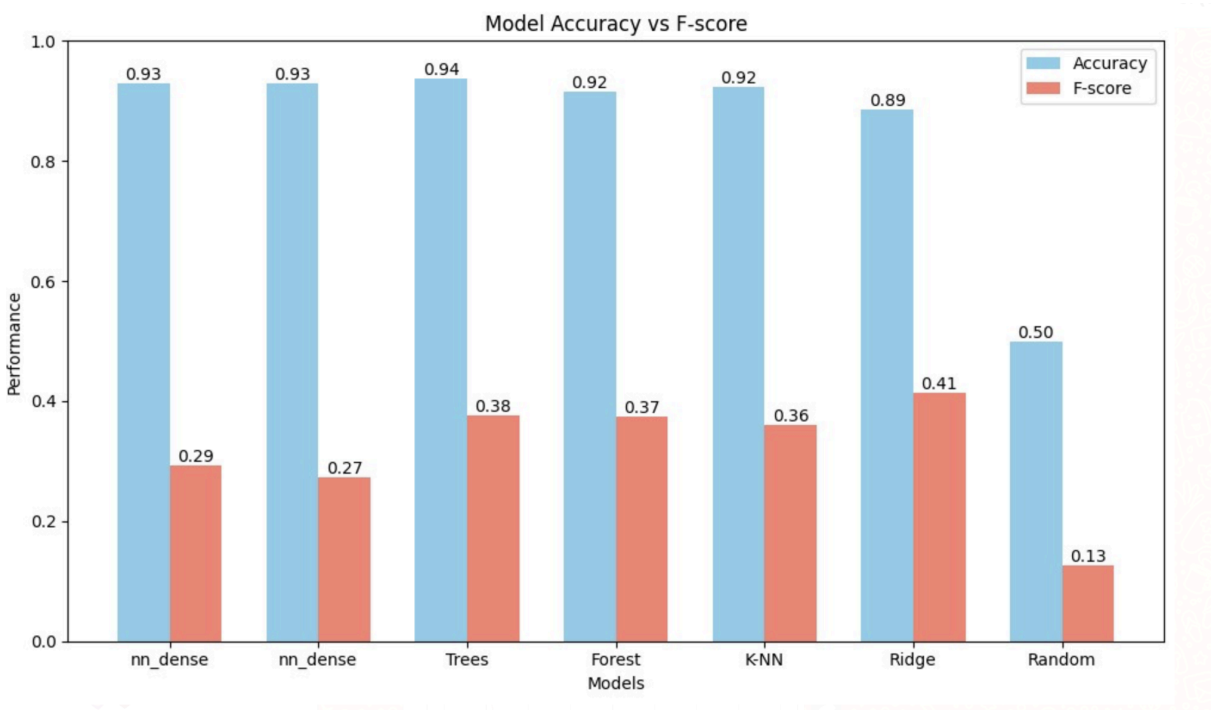
Training the beast

Training a standard model is rather simple: Get the data, get the model, feed the model the data :). Yet, not all is happiness, neural networks require more work. Usually we will have to feed them the data more than ones (have some epochs). Therefore, we can see how it evolves over time.





Unlikely, our big boys are not very strong, and they had a bad time training :(. The lack of computational power does not allowed us to make more validation on this models, and therefore train longer and with better settings. We leave this as a learning taks for the reader. So now, lets compare the performance of the models.



As we can see all of them have a high accuracy and decent f-score. We also tried doing the predictions at random and as we can see, our model are better! Now, let's see some of the predictions.

As we can see our ridge guy get some a lot of the genres right but then, add some other for you to be aware of any potential topics hahaha!