So, I’ve been working with a dataset from a Portuguese bank, trying to figure out if customers would sign up for a term deposit after receiving a phone call. The goal was to predict their response using machine learning models like KNN and SVM, and I’ve got some pretty interesting results.

KNN

I started off with KNN and played around with different K values (from 1 to 20) to see how the model would behave. It turns out, K=20 gave me the best performance. The accuracy was decent—about 89.5%, but precision was a little low, hovering around 66% for K=20.

Here’s the thing with KNN though: recall was super low across all the values I tried. For instance, with K=2, recall was just 15.7%. That’s not great when you want to predict people who are more likely to subscribe to the term deposit. The model was mostly catching the non-subscribers, but it was missing a lot of the subscribers, which is kinda the point, right?

SVM

I then tried out SVM with three different kernels: linear, polynomial, and Gaussian. Each one gave pretty decent accuracy, all hovering around 89% or higher. But, man, recall was still the problem. It seemed like the SVM was more focused on predicting people who wouldn’t subscribe to the deposit.

* SVM Linear had good precision (63.7%), but recall was way too low at 20.1%.
* The polynomial kernel bumped up the precision to 69%, but recall didn’t improve much, still stuck around 18.3%.
* Even with the Gaussian kernel, accuracy was slightly higher at 89.5%, but recall stayed low at 17.8%.

What I Learned

The models are doing a good job of predicting people who won’t subscribe, but they’re really bad at catching the ones who will. The accuracy looks nice, but the real issue is that the models are biased toward predicting non-subscribers because the data is so imbalanced—there are way more people who didn’t sign up than those who did.

Precision is fine, but recall is the real issue here. I need the models to get better at finding the customers who are likely to sign up. So, it’s clear that there’s some serious work to do around balancing the data or tweaking the models to catch more of the positive cases (subscribers).

What to Do Next

* I think the data imbalance is the main thing dragging down recall, so I’d try out something like SMOTE to balance the dataset and see if it helps.
* Maybe I could fine-tune the hyperparameters a bit more, especially with SVM. There’s also the option of using cross-validation to make sure the models are actually generalizing well.
* It might be worth exploring some extra features or transformations—more info on the customers could help improve how the model is making predictions.

Final Thoughts

Overall, the models are decent, but they need a lot of improvement, especially when it comes to finding the subscribers. I’m pretty happy with the Gaussian SVM for accuracy, but I need to figure out how to get the models to be more balanced. With some adjustments, I’m confident the results can be a lot better.