**How to Avoid Mistakes in Data Science**

**Understanding the occurrence of various mistakes in data science, especially when building machine learning code is useful for practitioners of data science and artificial intelligence. In this article, we would be exploring steps to avoid these mistakes and improve productivity.**



Photo by [Jan Antonin Kolar](https://unsplash.com/@jankolar?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

Machine learning and data science are being used in a wide variety of applications. Some of the cool applications of machine learning are in self-driving cars and also in banking industries of whether a person is going to default on a loan based on a set of features. **Machine learning** is also used in a vast array of other industries starting with pharmaceutical, retail, manufacturing, and agricultural industries.

A lot of data scientists and machine learning engineers are hired to make use of **huge amounts** of data and generate valuable predictions based on the business use cases respectively.

However, there are often times when practitioners run into **issues** along the way when trying to build these applications in the field of artificial intelligence. We will now go over a list of ways in which mistakes in data science can occur when doing the task of fulfilling business requirements with AI and also taking the steps to **avoid** them to a large extent. We explore some of the simple ways in which data science can fail along with taking the right practical steps to ensure that these failures are avoided later when building the applications.

**Mistakes in Data Science and Machine Learning**



Photo by [Sarah Kilian](https://unsplash.com/@rojekilian?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

When building interesting machine learning applications, there is sometimes a possibility for practitioners to be making mistakes in the field. As a result, the **quality** of work done to impress the team is minimized. Therefore, taking a look at various mistakes in data science and looking for ways to reduce them improves productivity to a large extent. Below are some of the mistakes that occur mostly in the data science field.

**Failing to Understand Bias in ML Models**



Photo by [Will Myers](https://unsplash.com/@will_myers?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

There can often be times when the models might have a good capacity to perform quite well with the test set. **Bias** in models can occur when we find that we have not sufficiently trained the models to harness their full potential when making machine learning predictions. It could be mainly due to not tuning the hyperparameters, not giving enough data, and not adding features that would make a significant impact on the model.

In order to avoid this situation, one must properly train the models with the available data along with ensuring that the model reaches the global minimum in terms of the error produced. This would ensure that we are getting the best machine learning models.

Considering the task of predicting whether a customer is going to **churn** (leave the service) the internet service based on the set of features such as the age, type of internet service, and other factors, it would be seen that we would have to use complex machine learning models for predictions in this case. Using less complex models such as logistic regression might not always capture the **trend** and insights from this data because there is a good amount of complex relationships. If we were to use the logistic regression model, it would mostly be suffering from high bias as it has failed to capture the trends. One of the best ways to get around this would be to add more complex models and improve the machine learning predictions.

**Not Understanding Business Requirements**



Photo by [Timon Studler](https://unsplash.com/@derstudi?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

The technology that is often being used in machine learning and data science is quite impressive and **fascinating**. The opportunity to take a look at how these models can extract data and understand and gain useful insights from it seems like an impressive feat. A lot of teams in your organization might be pushing efforts to implement machine learning: they want to jump into a bandwagon like others in order to produce quality work. However, it is always a good step to **ask** quality questions during the data science journey and whether machine learning is a feasible solution to a particular business problem at hand.

Considering the example of using data science for predicting whether a person would be buying a house or not. This is a scenario where data science could be most useful because predicting it accurately would **save** companies millions of dollars and revenue. They could better plan their budgets based on the predictions and ensure that data science is gathering good value overall in the process. Hence, it is an important step to understand the business requirements before trying to apply machine learning to a large set of problems.

**Failing to Remove Outliers in Data**



Photo by [Rupert Britton](https://unsplash.com/es/@rupert_britton?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

There can be times when you have discussed well with your team the requirements of the business and applied machine learning and generated results and good predictions. However, the data that was used to train the ML models might have a large number of **outliers**. This is a scenario where a large number of values lie in a certain range of values while the others have significantly higher or lower values than the **mean** or average of the data. This is the case where models perform well on the training set and fail to generalize well on the data that they have not seen before.

Having the presence of these outliers would impact the performance of a large number of machine learning models. Therefore, efforts should be taken to remove them and ensure that there is proper functioning of these models in real time. Steps can be taken in identifying them well. Some of the steps involve finding the standard deviation and taking a look at values that lie between 2 **deviations** away from the mean. This would ensure that we get the best predictions on the data at hand.

Examples of outliers in datasets can include predicting the price of cars with a set of other variables. When we try to predict the prices of various cars taking into account features such as mileage, horsepower, and other factors, there are situations where outliers can be encountered due to **human error**. In these situations, the models when trained with the dataset that contains outliers would be performing far worse than the one that does not contain them. Therefore, removing those outlier values from various features could be a step forward in building an effective solution.

**Failing to Use the Right Feature Engineering Techniques**

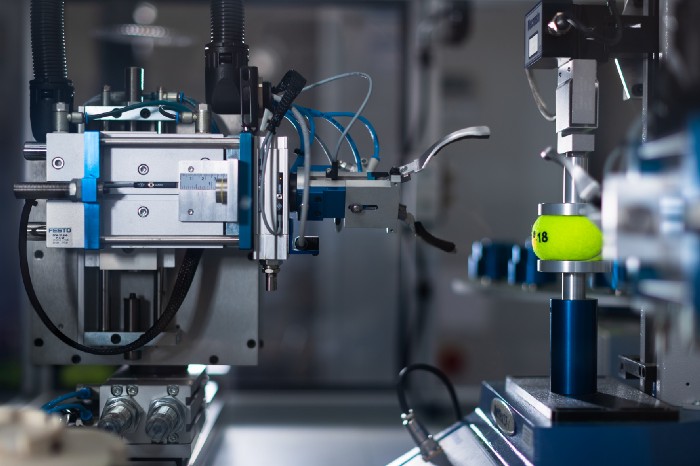


Photo by [ThisisEngineering RAEng](https://unsplash.com/@thisisengineering?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

The features that are used in ML model predictions determine how well they perform on **unseen** data. Therefore, giving our models the right access to various features and implementing strategies to create new ones boosts the performance of these models. Furthermore, using featurization and creating new features helps in exploring the data well with the use of various plots. These plots can oftentimes help in gaining valuable insights and can be handed over to the business so that they take data-driven decisions.

Oftentimes in machine learning, there are a large number of missing values in the data that we are going to feed to our ML models. Some examples of real-world problems with missing data include loan default prediction, heart disease prediction, and cancer diagnosis prediction. All of these examples contain features that contain **missing** values. This leads to the models not performing well on this data. Performing feature engineering can be handy for problems such as credit fraud detection where the data might contain missing salary information. Imputing the values with either the mean of the entire salary or the mode does the trick.

If you are interested to know more about various featurization techniques, you can take a look at my earlier article where I mention them in great detail along with practical examples. Below is the link.

[Which Feature Engineering Techniques improve Machine Learning Predictions? | by Suhas Maddali | Nov, 2022 | Towards Data Science (medium.com)](https://medium.com/towards-data-science/which-feature-engineering-techniques-improve-machine-learning-predictions-227d732068f5)

**Assuming that Deep Learning can Solve any problem**



Photo by [Sai Kiran Anagani](https://unsplash.com/@anagani_saikiran?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

With the rise in technological innovations and newer strategies implemented by various companies, it is becoming easier to get access to **large volumes** of data which can be extracted and made available to a large number of teams to perform machine learning related tasks. It was also later revealed that with the increase in the quantity of data, it is becoming more appropriate to use deep learning.

While it is true that using deep neural networks (deep learning) can improve the performance with the data, there are oftentimes expectations from the team to **explain** why the models gave predictions in the first place. It is during this time that the approach of deep learning can fail especially when explainability is one of the most important requirements for a specific ML application.

Consider the case of diagnosing whether a patient would be suffering from cancer based on a set of factors such as weight, blood pressure and BMI. As we have a large amount of data from cancer patients, it is easier to come to a conclusion to use **deep learning** to predict the chances. In the case of cancer diagnosis, however, it is equally important to explain the predictions from the deep learning models. Due to the nature of deep learning models being more complex with their capability of extracting intricate relationships, it becomes harder for them to explain why exactly they have given predictions in the first place. In this case, therefore, it would be a good approach to use simple machine learning models that are highly interpretable to the practitioner, doctor and the patient respectively.

**Conclusion**

After going through this article, hope you have understood some of the mistakes that can occur as a result of using machine learning and deep learning to build interesting AI applications for products. Taking the steps that were mentioned in the article can help some of these challenges to a large extent while also increasing efficiency. Thanks for taking the time to read this article.

If you like to get more updates about my latest articles and also have unlimited access to the medium articles for just **5 dollars per month**, feel free to use the **link** below to add your support for my work. Thanks.

<https://suhas-maddali007.medium.com/membership>

Below are the ways where you could contact me or take a look at my work.