Accuracy

Understanding common challenges to model accuracy

Accuracy

Why should models be accurate?

The **value of models** relies on their **accuracy**, as this factor significantly influences their **efficacy** and **dependability**. Some reasons that highlight the significance of accuracy in models include:

01. Reliable predictions

Reliable predictions **align closely with the actual outcomes** of the system being modeled and are important in **fields where precise predictions are vital** for decision-making and risk assessment.

02. Resource optimisation

Accurate models help **optimize** the allocation of resources, whether it's allocating funds, manpower, or materials. This results in **cost savings**, **improved productivity**, and **enhanced efficiency**.

03. Risk management

Whether it's assessing credit risk, market volatility, or operational risks, accurate models provide **reliable estimates of potential risks**, enabling implementation of appropriate **risk mitigation strategies**.

04. Improved understanding

Accurate models contribute to a better understanding of the underlying data **processes**, **relationships**, and **patterns**. They **reveal insights** that might not be apparent, leading to **new discoveries**.

05. Enhanced competitiveness

Accurate models **allow organizations** to identify emerging trends or opportunities, optimize operations, and deliver improved products or services, giving them an **edge over their competitors**.

06. Trust and credibility

When models consistently demonstrate accuracy, they **instill confidence** in the organization's capabilities, leading to **stronger relationships** and **increased trust** with stakeholders.

Accuracy

Why should models be accurate?

For example, the primary goal of a **predictive model** is to make **accurate predictions for new, unseen data**.

Models make predictions by recognizing **complex patterns and relationships** between variables in the **training data**. The **more effectively** a model can learn these patterns, the **better its predictive performance**.

Prediction is **crucial** in **various fields** like finance, healthcare, marketing, and more. When done **accurately**, it allows us to anticipate **future outcomes** and make informed decisions.

A model's ability to predict accurately is **evaluated** using various performance metrics like **mean squared error** for **regression models** and **accuracy**, **precision**, **recall**, and **F1 score** for **classification models**. These metrics tell us how well our model is likely to perform on new data.

Why understand model accuracy challenges?

> Developing a comprehensive understanding of the challenges that can impact model accuracy is of utmost importance to ensure **effective and reliable data analysis**, enabling data analysts to make **informed decisions** and provide **accurate insights**.

Awareness and proactivity

By recognizing the **potential obstacles** that can impact the accuracy of models, data analysts can develop **robust and trustworthy models** by using **proactive measures**.

Being aware of **how a model works** is also a proactive way to produce accurate results.

Informed decision-making

Understanding model accuracy challenges enables analysts to make **informed decisions** based on the **model's outputs**.

It ensures that decision-makers are aware of the **limitations**, helping them avoid reliance on inaccurate predictions.

Building accurate models

By understanding accuracy challenges, analysts can employ **techniques to improve fairness** in model outcomes.

This understanding empowers data analysts to build more accurate models, enhance decision-making processes, and provide reliable insights to stakeholders.

Bias vs variance

When evaluating the accuracy of a model, two important components to consider are **bias** and **variance**.

Bias

- Bias refers to the **error** introduced by the model's **assumptions or simplifications**, causing it to **consistently miss the mark** when making predictions.
- A model with high bias **oversimplifies the underlying relationships** in the data and fails to capture the true patterns.
- This leads to **poor performance**.

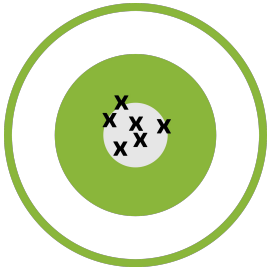
Variance

- Variance refers to the model's **sensitivity to fluctuations** in the training data.
- It measures how much the **predictions** of the model **vary** when fitted on **different subsets** of the data.
- A model with high variance tends to **overfit** the training data; that is, it **memorizes** the training data instead of **learning** the underlying patterns, causing **poor performance** when tested on new data.

Accuracy

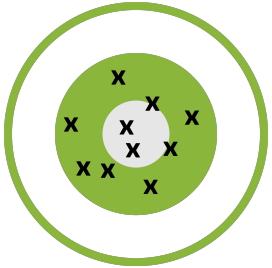
Bias vs variance

To illustrate bias and variance, we can use a **dartboard** analogy. Consider a scenario where there are four players in a darts game, each belonging to a different skill level.



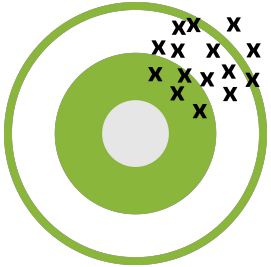
01. Ideal learner

The player represents an ideal learner who demonstrates exceptional precision in their performance.



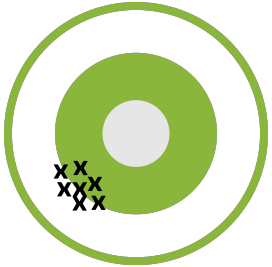
02. Good learner

The player adopts a specific method or strategy that leads to a satisfactory performance.



03. Terrible learner

The player lacks accuracy and stability, consistently missing the target due to their bad throws.



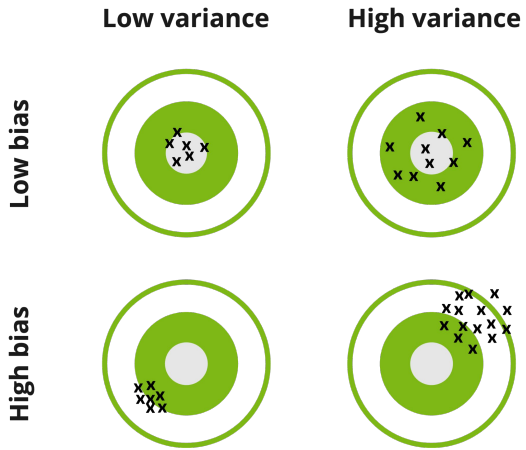
04. Naive learner

The player possesses good aim but lacks a proper approach, resulting in frequent misses or underperformance.

Bias vs variance

> We can **draw parallels** between the players' performance and the model's performance, examining the **influence of variance and bias** in shaping their respective capabilities.

- **Bias** can be understood as **how close** the dart throws **are to the center** of the target. If the darts consistently land around the same spot but not at the center, it indicates a high bias. This means the darts are consistently hitting a similar spot but off target.
- On the other hand, **variance** refers to the **spread** or **inconsistency** of the dart throws. If the darts land all over the target with no clear pattern or cluster, it indicates high variance. This means the darts are not consistently hitting the same spot and lack precision.



Our objective is to achieve a balance between minimizing bias and minimizing variance, but it's important to note that improving one aspect often comes at the expense of the other.

Bias vs variance

Causes of bias and variance

Causes of bias

- **Model simplicity:** A model that lacks the necessary complexity to capture underlying patterns in the data can result in high bias.
- **Under-representative features:** If important features or variables are not included in the model, it may overlook critical information, resulting in inaccurate predictions.
- **Incorrect assumptions:** Building a model on incorrect assumptions about the data or relationships can introduce bias.

Causes of variance

- **Model complexity:** Complex models with excessive parameters can result in high variance, leading to poor generalization on unseen data.
- **Over-reliance on noise:** Fitting the noise instead of the true patterns leads to poor performance on unseen data.
- **Insufficient training data:** The model struggles to learn the true underlying patterns effectively.

Accuracy

Bias vs variance

Reducing bias and variance

Reducing bias

- Use more **complex models** that can capture intricate patterns.
- **Incorporate** additional **relevant features or variables** into the model.
- **Increase** the **model's flexibility or capacity**, such as by using different types of models.

Reducing variance

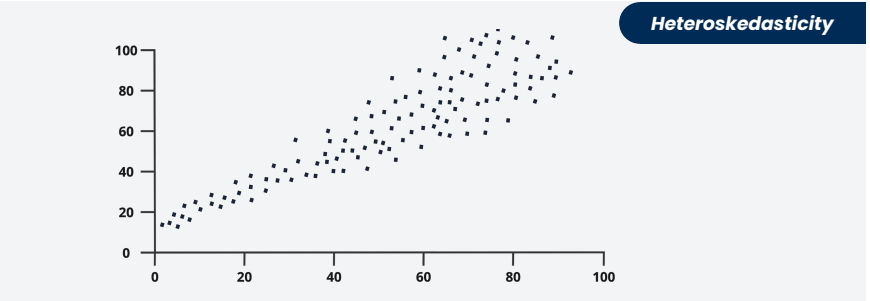
- **Increase** the amount of **training data** to provide the model with more diverse examples.
- Use a **dimensionality reduction** technique that reduces the number of features or variables in a dataset while preserving the most relevant information.
- **Simplify the model** by reducing its complexity, such as decreasing the number of features or using feature selection techniques.

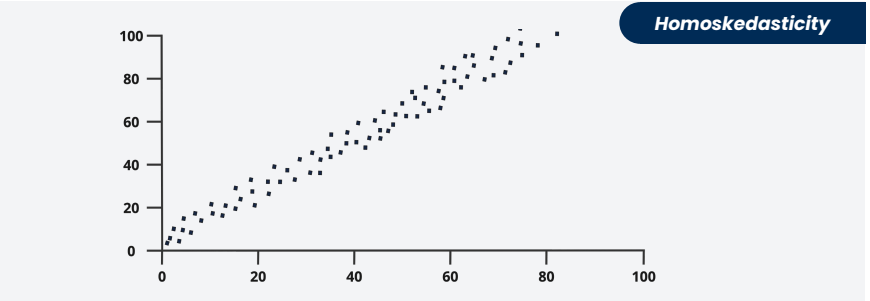
Other common challenges to model accuracy

Heteroskedasticity

- Heteroskedasticity refers to the situation where the **variability of the error terms** in a statistical model is not constant across the range of independent variables.
- In simpler terms, the **spread or dispersion of the errors is not consistent**. This can have implications for the accuracy and reliability of the model's predictions.

- Heteroskedasticity can be **detected** through visual inspection of a **residual plot**.
- If the resulting graph has a **fan or cone shape,** it suggests the presence of heteroskedasticity.
- **We aim for** our model to demonstrate **homoskedasticity**, which implies that our model's **errors have a consistent level of variability** and do not exhibit patterns or trends that could lead to biased or inefficient estimates.





Other common challenges to model accuracy

To address **heteroskedasticity**, there are a few common solutions:

01. Transforming the data

The **variables involved** in the model are transformed using common transformations like taking the **logarithm**, **square root**, or **reciprocal** of the variables. These transformations help stabilize the variance and make it more constant.

02. Weighted least squares regression

Observations with **larger variances** are given **less weight**, while those with **smaller variances** are given **more weight.** This gives **more importance** to the observations with **less variability**, which can help **mitigate** the impact of heteroskedasticity.

03. Robust standard errors

Standard errors help us understand how much the **estimated values** of sample data might **vary** if the sampling process is **repeated multiple times**. Robust standard errors provide **valid statistical inference** even when heteroskedasticity is present.

04. Heteroskedasticity-consistent standard errors

Using specific estimation techniques, **consistent standard errors** in the presence of heteroskedasticity can be provided. These techniques **adjust the standard errors** to account for heteroskedasticity, allowing for **reliable statistical inference**.

Other common challenges to model accuracy

Endogeneity

- Endogeneity refers to a situation in which there is a **two-way relationship between** a **variable of interest** and **other variables** in a statistical model **(correlation)**.
- In simpler terms, it means that the **variable** you are trying to study is **influenced by other factors**, and at the same time, it also **influences those factors**.
- This **intertwined relationship** makes it challenging to disentangle **which factor truly leads to changes** in the variable under study.

- **Detecting** endogeneity can be **challenging**, but there are a few methods that can help identify its presence and assess its impact on model accuracy.
- The most common approach is **correlation analysis**.
- Correlation analysis is a statistical method used to measure the **strength** and **direction** of the **relationship between two variables**. It helps us understand how changes in one variable are associated with changes in another variable.
- However, correlation alone **does not prove causation** or establish which variable is causing the changes. Therefore, while correlation analysis can raise suspicions of endogeneity, it **cannot definitively confirm** it.

Other common challenges to model accuracy

To address **endogeneity**, there are a few common solutions:

01. Instrumental Variables (IV) analysis

An instrumental variable is a **variable that is correlated with** the **endogenous variable** but not directly related to the outcome variable. It acts as a "tool" to help us **estimate the causal effect** of the endogenous variable on the outcome variable.

02. Control variables

Including additional control variables helps to account for **other factors** that may be related to both the independent and the outcome variable, thus **reducing the influence** of endogeneity. However, the **control variables should not be endogenous**.

03. Difference-in-Differences (DID)

In **observational studies**, this technique **compares the changes** in an outcome variable **over time** between a **treatment group** and a **control group**. It estimates the **causal effect** while accounting for confounding factors.