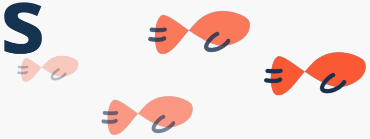
neue fische

School and Pool for Digital Talent

Designing Data Products

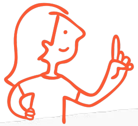


Why would we build a model?

Exploratory analysis - understand what happened in the past

Predictive - predict what will happen

Predict what, for whom and for what purpose?



you do not always need an ML model

Product = Customer x Business x Technology

Usability
Business viability
Feasibility

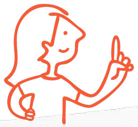
Value = product of the three.. If one is zero then the value too

Measuring Success

The first model you build should be the simplest model that could address the product needs

Business performance: measured usually by one KPI (key performance indicator)

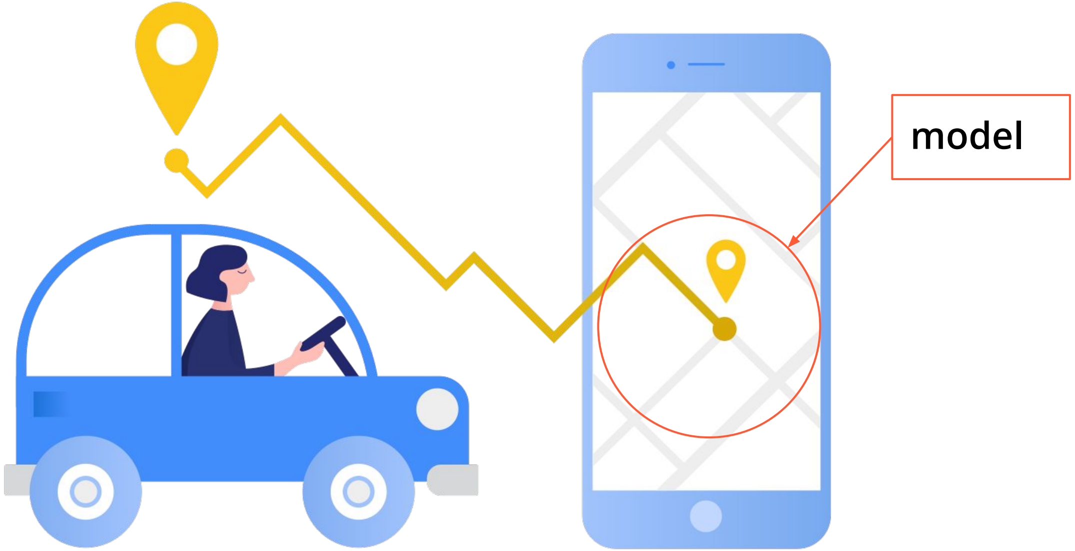
Model performance: an offline metric that captures how well the model will fit the business need



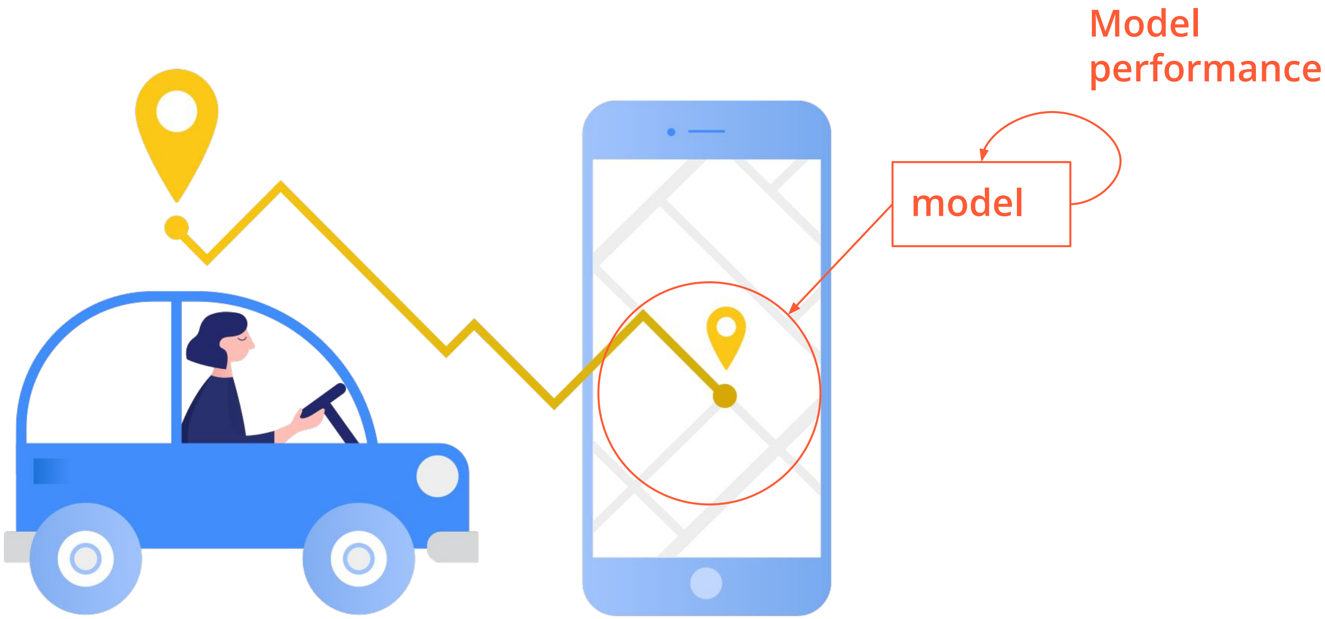
The business metric is independent from the model metric.. It is a measure of the product success

Business Performance vs Model Performance

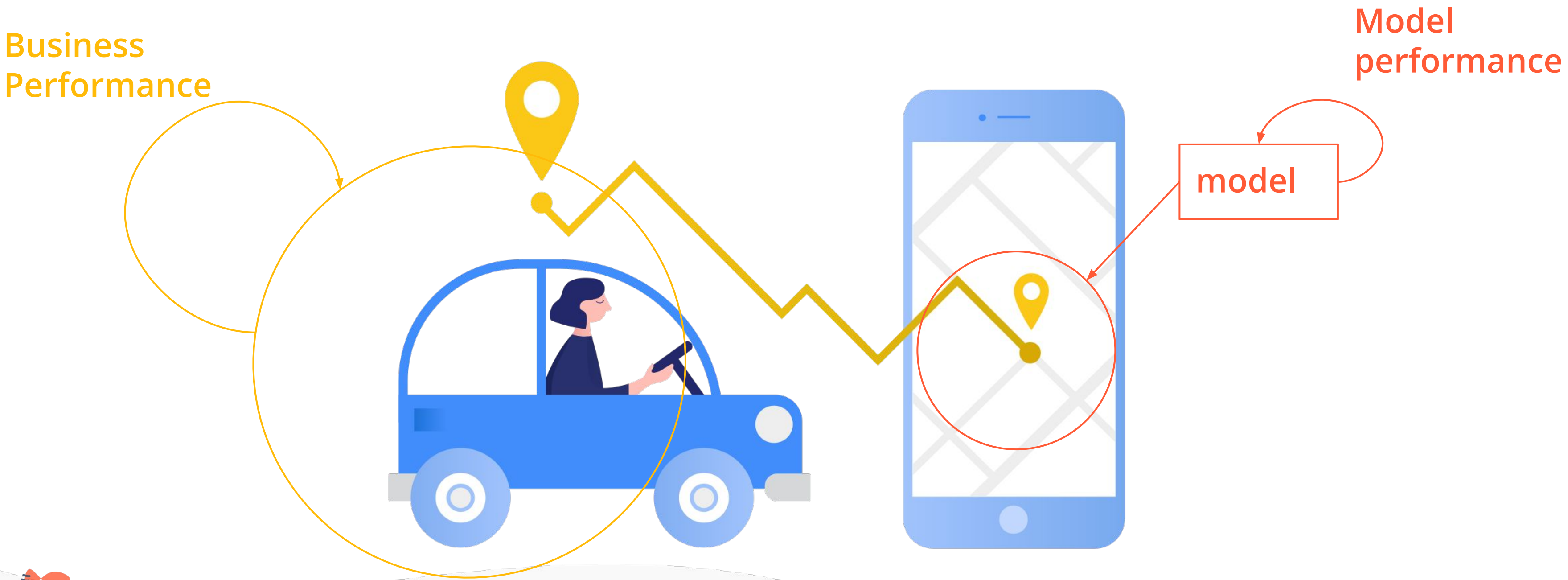
Business Performance vs Model Performance



Business Performance vs Model Performance



Business Performance vs Model Performance



examples

Model Performance

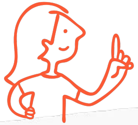
Regression:

- RMSE, RMSLE
- MAPE (mean absolute percentage error) - accuracy as a ratio

Classification

- Accuracy
- Precision
- Recall

Custom metric: based on the worst case scenarios of your product.



If you need to present to stakeholders you need a simple metric.. rmse , precision, recall.. Are too complex to explain

Relationship

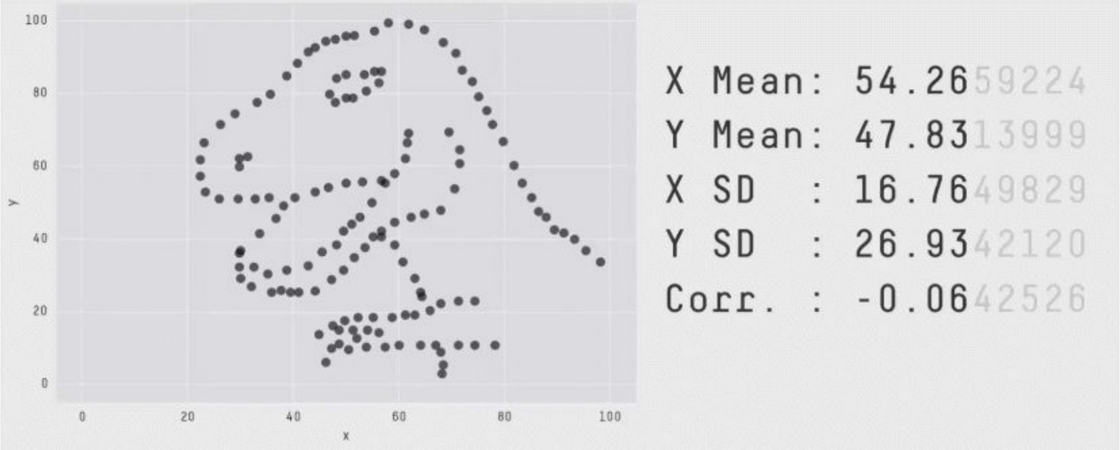
Business Performance & Model Performance

Thinking of the business value of your model and the cost of being wrong can help you choose the right model metric

Always start from the value!

ERROR ANALYSIS

Remember the Summary vs details?



X Mean: 54.2659224

Y Mean: 47.8313999

X SD : 16.7649829

Y SD : 26.9342120

Corr. : -0.0642526

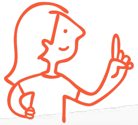
Going beyond aggregated metrics

All the performance metrics we've seen are aggregated metrics

They help determine whether a model has learned well from a dataset or needs improvement

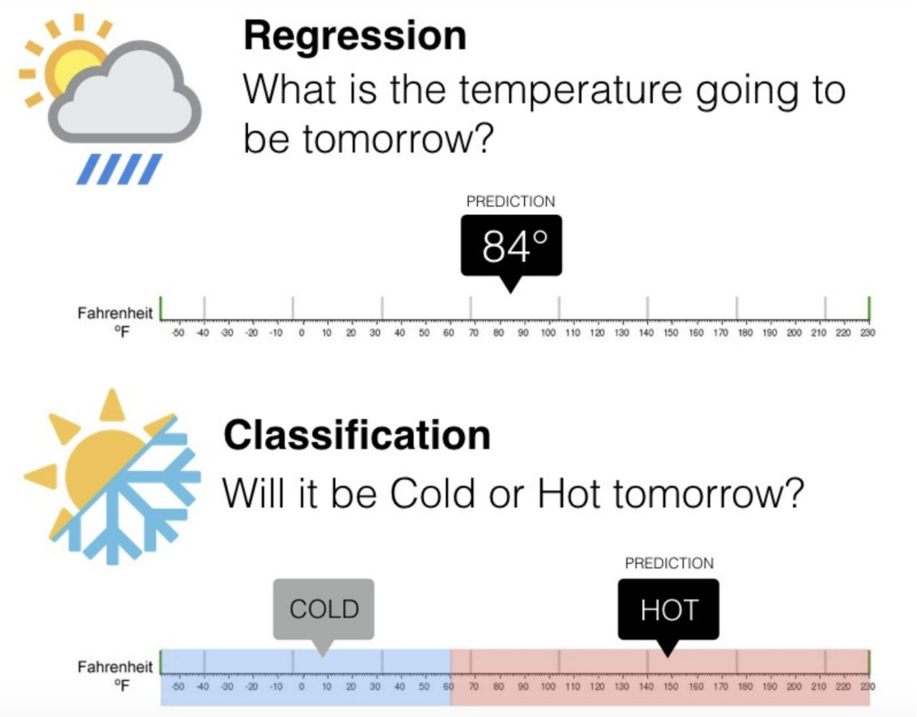
Next step: examine results and errors to understand why and how is the model failing or succeeding

Why: validation and iteration



Performance metrics can be deceptive, on highly imbalanced datasets a classifier can reach very high accuracy without any predictive power

Binary classification



Regression

What is the temperature going to be tomorrow?

Classification

Will it be Cold or Hot tomorrow?

https://towardsdatascience.com/regression-or-classification-linear-or-logistic-f093e8757b9c

Goal

Validate your model - inspect how it is performing

There are lot of ways to do this.. You want to contrast data (target and/or features) and predictions

- For **regression**: looking at residuals, for example doing EDA on residuals and inspecting the outliers
- For **classification**: one can start with a confusion matrix, breaking results in true class and predictions

Confusion Matrix

- Counts how often the model predicted correctly and how often it got confused
- False Positive: false alarm / type I error
- False Negative: missed detection / type II error

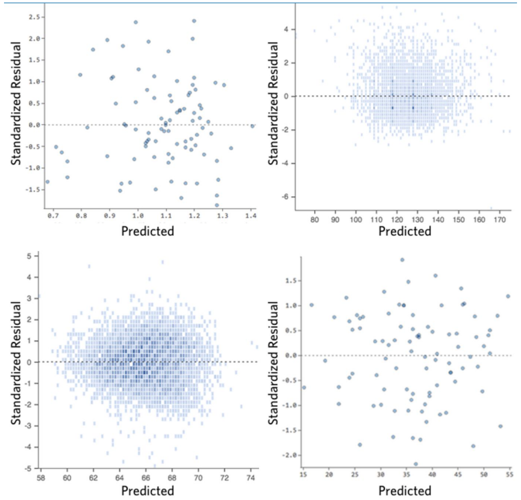
What do the misclassified examples have in common?

		Predicted	
		Negatives	Positives
Actual	Negatives	TN	FP
	Positives	FN	TP

Regression

Residual analysis

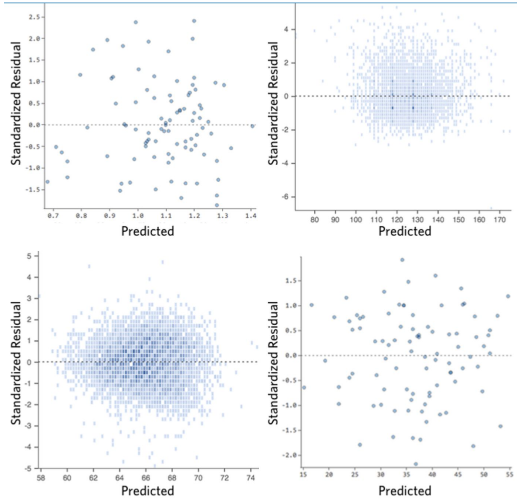
- This is like EDA again but on residuals (predicted - observed)
- Plot residuals /and standardized residuals vs predicted
- We want our residuals to have no patterns, to be symmetrically distributed, centered in the middle of the plot

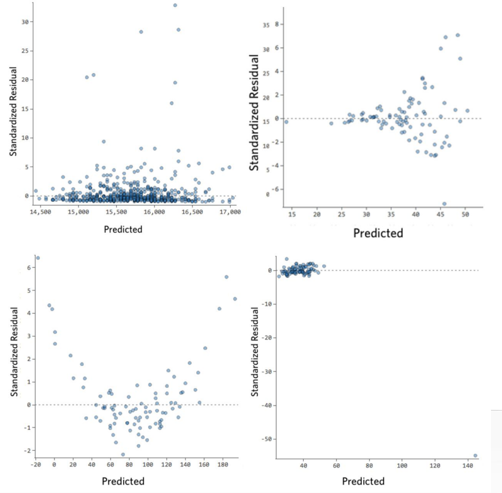


Regression

Residual analysis

- This is like EDA again but on residuals (predicted - observed)
- Plot residuals /and standardized residuals vs predicted
- We want our residuals to have no patterns, to be symmetrically distributed, centered in the middle of the plot
- IF not.. Then there is room for improvement in the model





What if my residuals look like this walkthrough: https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Resources

https://svpg.com/what-is-a-product/
https://medium.com/analytics-vidhya/root-mean-square-log-error-rmse-vs-rmlse-935c6cc1802a
Building Machine Learning Powered Applications - Emmanuel Ameisen
https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/
https://www.scikit-yb.org/en/latest/api/regressor/residuals.html

Example of EDA with error analysis
https://www.kaggle.com/elitcohen/forest-cover-type-eda-modeling-error-analysis#Error-Analysis
https://www.kaggle.com/pestipeti/error-analysis
https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

ML Project Topics

Kickstarter Project Success

Analyse and model success factors of kickstarter campaigns. Give new projects an idea what is needed for a successful funding and potentially even predict campaign success upfront.

- 221811 rows of data on campaigns
- (medium/hard)

Kickstarter Project

Tanzania Tourism Prediction

Can you use tourism survey data and ML to predict how much money a tourist will spend when visiting Tanzania?

- Survey Data from 6476 participants
- (easy/medium)

Zindi-Tansania-Tourism

Fraud Detection Challenge in Electricity and Gas Consumption

- Based on client’s billing history detect clients involved in fraudulent activities
- (medium)

Fraud Detection Challenge

Urban Air Pollution Challenge

Predict air quality levels and empower communities to plan and protect their health

- weather data and daily observations collected from Sentinel 5P satellite tracking various pollutants in the atmosphere
- (medium/advanced -> domain knowledge helpful)

Air Pollution Challenge

Flight Delay Prediction Challenge

Predict airline delays for Tunisian aviation company, Tunisair

- Data on flight delays. Can be combined with airport locations
- (medium)

Flight Delay Prediction Challenge

Financial Inclusion in Africa

Can you predict who in Africa is most likely to have a bank account?

- Survey data on financial inclusion of ~33,600 participants
- (easy/medium)

Financial Inclusion in Africa

Turtle Rescue Forecast Challenge

Anticipate the number of turtles to rescue

- Lots of data cleaning
- (easy/medium)

Turtle Rescue Forecast Challenge