**Youtube**

The aim is to identify what drives higher impact and provide insights into the prediction logic. The model was built to find importance of features involved in higher numbers of views from a Youtube prenumerant based on numeric or categoric data.

**Prep and clean**

Data read in chunks. **Outliers** removed. The model must not learn extremes instead of majority of data. **Missing values** handled by mean imputation. To save rows and get better generalization.

The variable date (join\_date) made to (weekday) to be useful for weekday patterns to be seen. Seven categories, 0-6 were every number is a weekday.

**Model**

**TensorFlow + Keras,** dense feed forwardneural network

Scaled numeric data and SVD-textfeatures

Layers

Dense (256) –Dropout (0,3) –BatchNorm

Dense (128) – Dropout (0,3)

Dense (64) Output (1)

Regularisation L2 and Dropout to avoid overfitting.

Optimices with Adam to get lossfunction (MAE)

**Training**

**K-fold Cross validation** (n=3) to avoid overfitting and get robust error measure. Tests all data in folds and no data is wasted. Every fold has test and training but not at the same time.

**Early Stopp (patience 5)** stops training when an improvement ceases.

**Result**

**MAE ≈ 0.94 log-view for every prenumerant**

**≈ 2 view/sub**

**SHARP**

**What did the model learn**

**Summary\_plot** visualizefeatures that are of importance

**Text\_svd\_56** and **text\_svd\_141** tells us that the views are greatly influenced by the content of the channel's description, name, and keywords.

Then the model explores what Text\_svd\_56 contains. When the variable was examined, it emerged that words as lifestyle, family and food channels, with a focus on sharing, home content and recipes was very interesting and draw in high numbers of views.

Text\_svd\_141 did specifically point out that the words **subscribe**, **new video** or **watch now** in text form on the channel affect views highly.