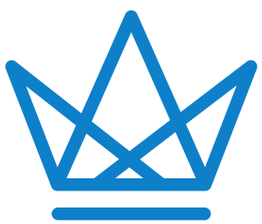
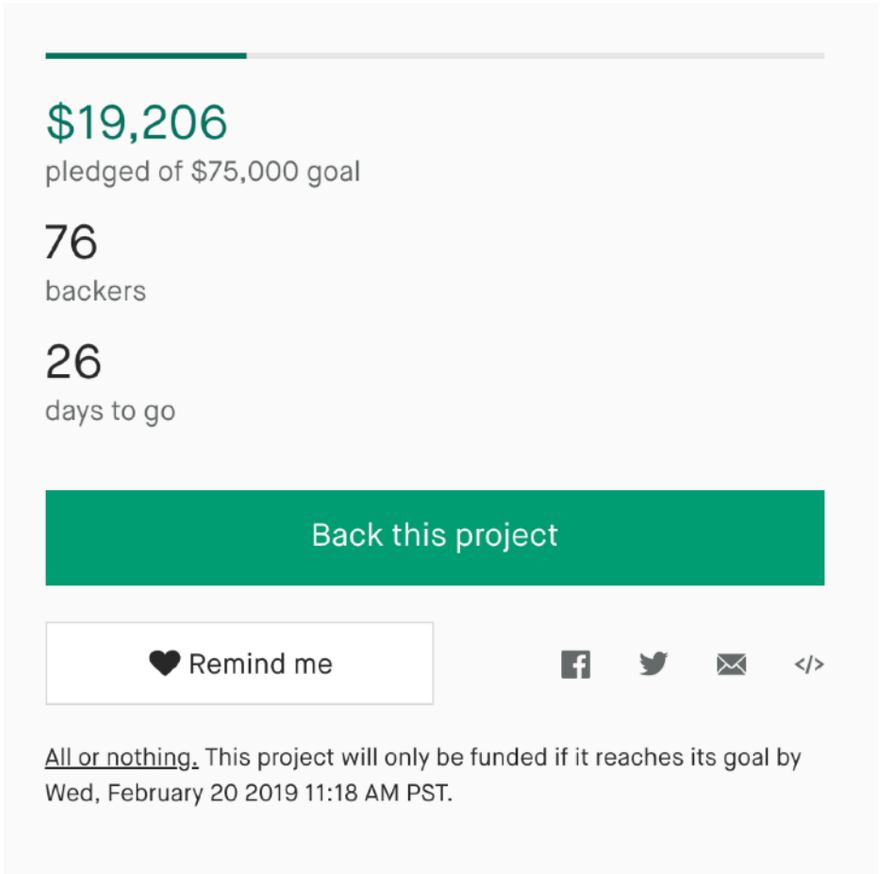


KICKSTARTER PROJECT FUNDING

MARC KELECHAVA

KICKSTARTER FUNDING PREDICTION



Kickstarter follows an **all-or-nothing** funding model: if the goal is not met, then no one is charged and the project gets nothing.

Prediction Goal - Identify if the content of a Kickstarter project page can predict project funding (pledged amount)

Inference Goal - What is the relative impact of page content on pledged amount?

Product Enhancement - Imagine a system where Kickstarter could automatically suggest content improvements to project creators

TYPICAL KICKSTARTER PAGE

Timmy and Tommy will be made from super soft minky fabric, will be about 7" in size, will be stuffed with poly stuffing, and will cost about $4000 to start production on (including KS costs, etc).

Pledge $20 or more

~TIMMY AND TOMMY THE TWO HEADED TURTLE~
Pledge this amount and you will receive a Creepy Kawaii digital wallpaper, and "Timmy and Tommy the Two Headed Turtle" plush.

ESTIMATED DELIVERY
May 2015

SHIPS TO
Anywhere in the world

80 backers

Pledge $20 or more

~ANY UNLOCKED PLUSH !~
Pledge this amount and you will receive a Creepy Kawaii digital wallpaper, and any unlocked plush doll.

ESTIMATED DELIVERY
May 2015

SHIPS TO
Anywhere in the world

31 backers

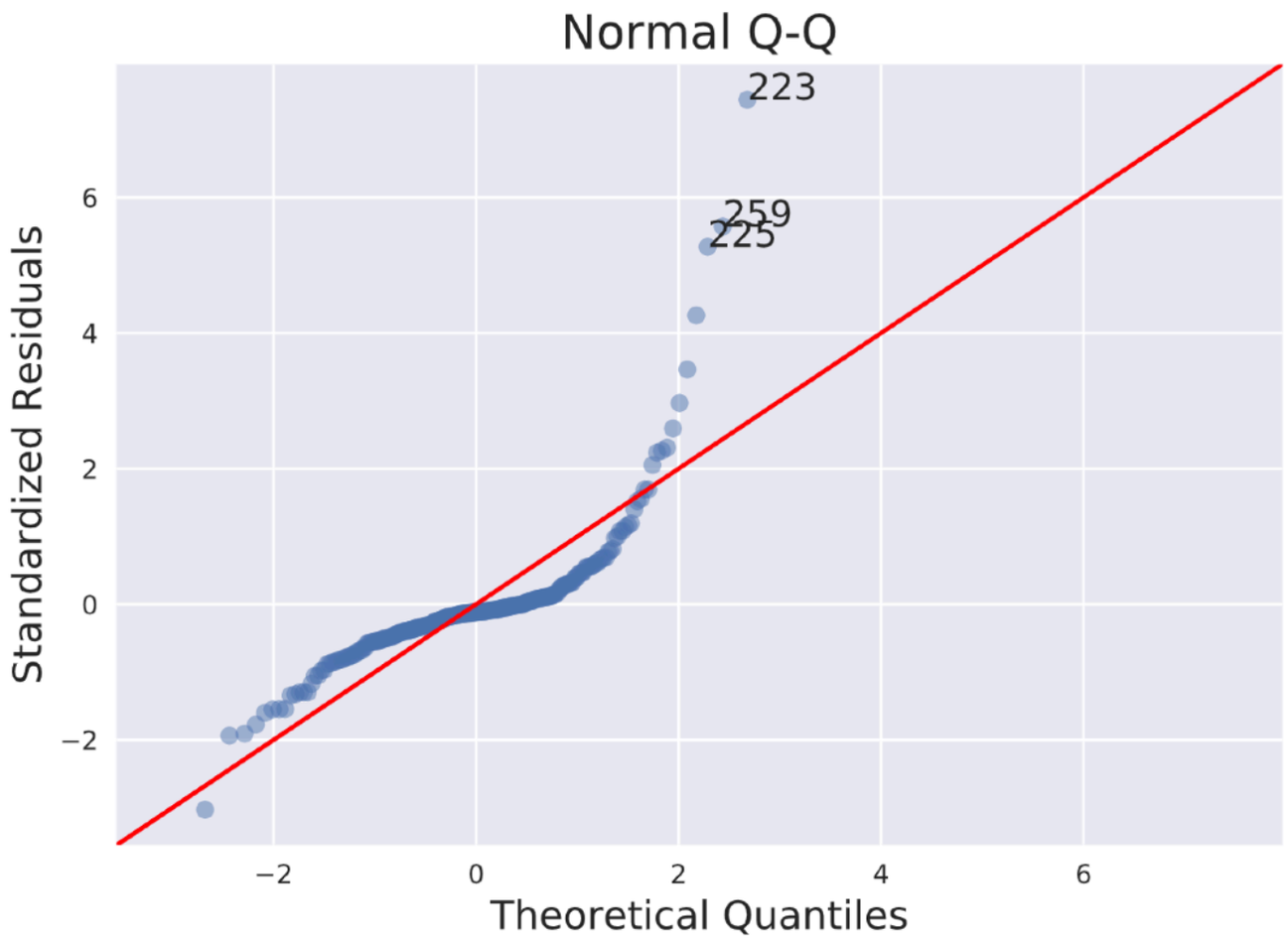
SCRAPEABLE ELEMENTS

- Description Text Sentiment*
- Pledged $, Goal $
- Project Length
- Number of gift options
- Number of photos
- Video Header
- Length of text description
- Average gift option $

*Scored chunks of text with NLTK Vader Sentiment Analysis Library

NAIVE REGRESSION APPROACH

A FAR CRY FROM NORMALITY



LINEAR REGRESSION

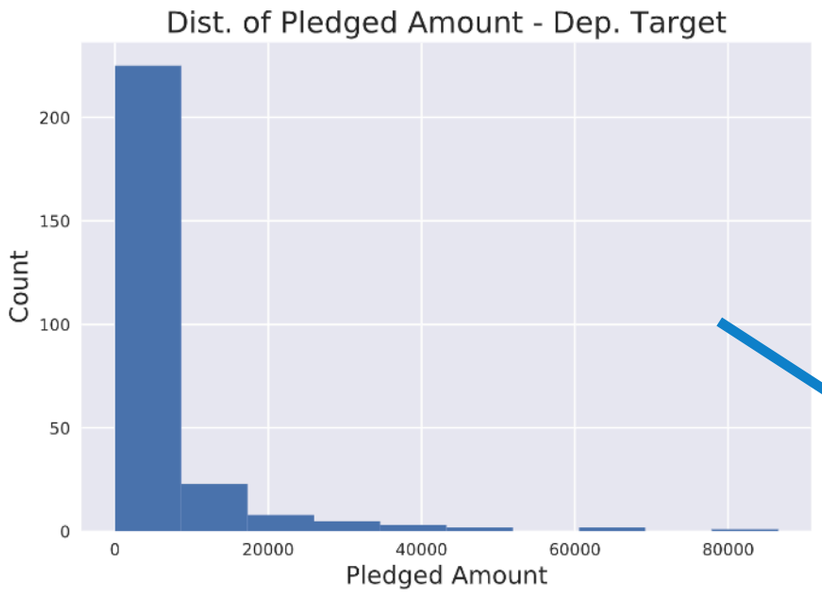
ORDINARY LEAST SQUARES

To meet the assumptions of OLS Linear Regression, the **prediction error** between true pledged $ and predicted pledged $ must be approximately normally distributed.

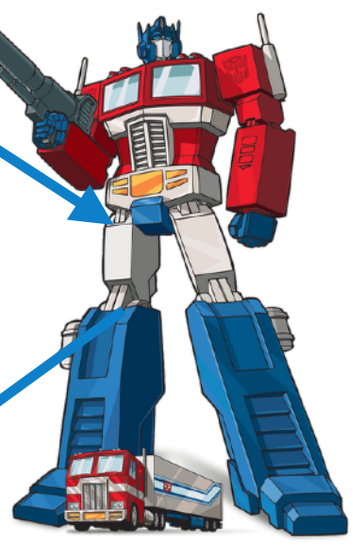
Clearly this is not the case.

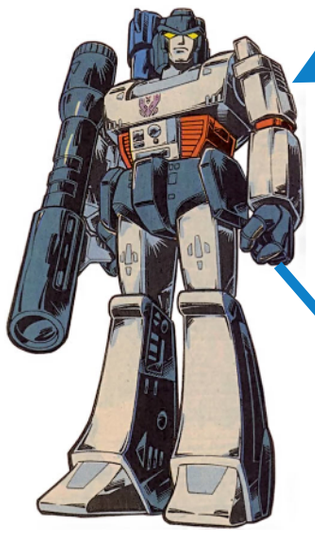
TRANSFORMERS

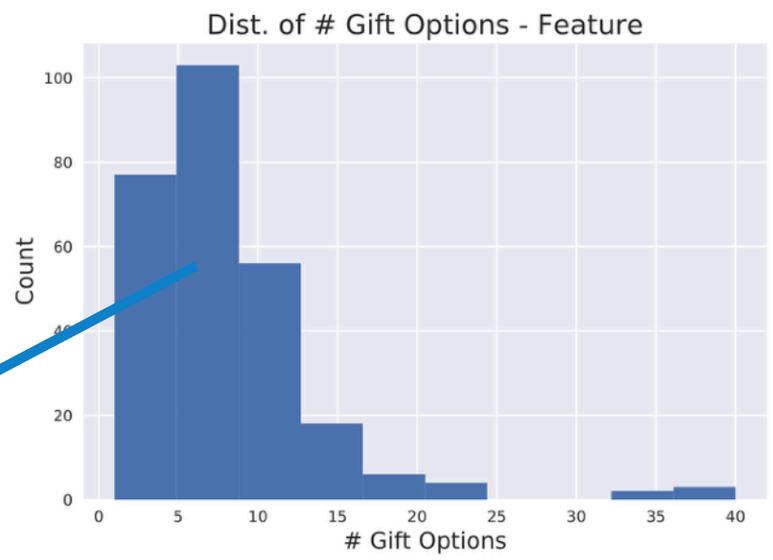
ON TARGET AND SOME FEATURES

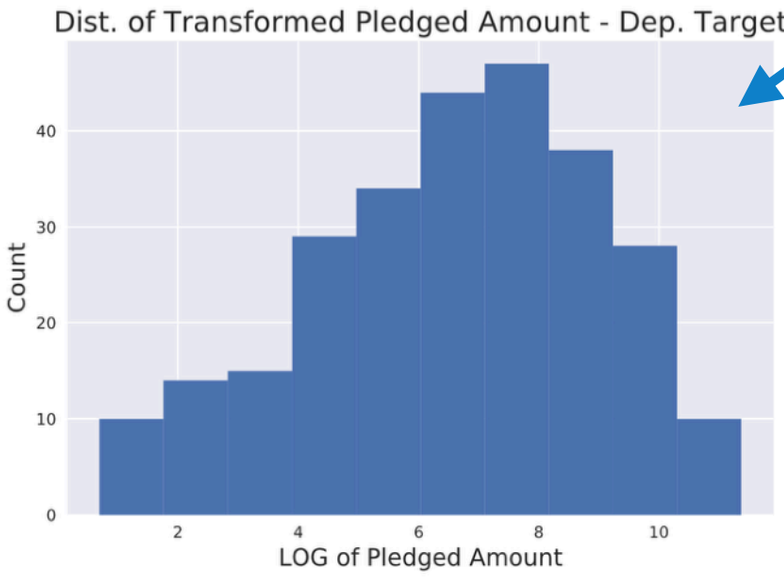


LOG TRANSFORMATION

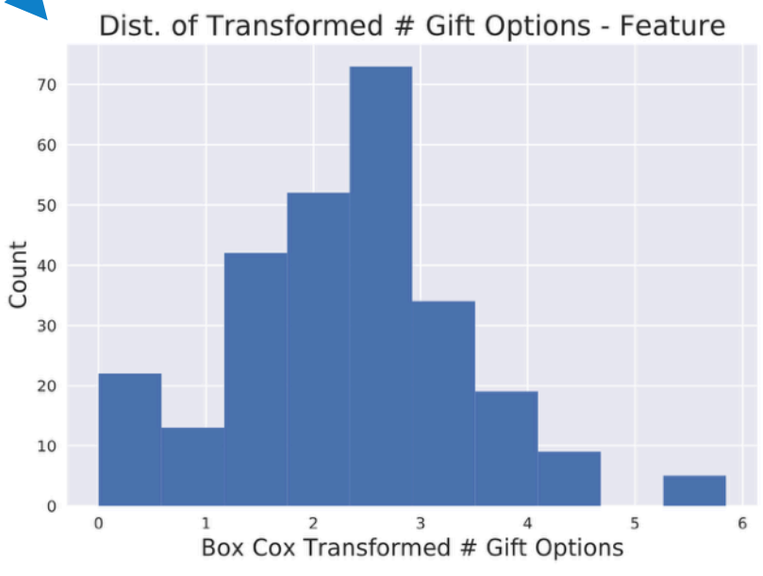








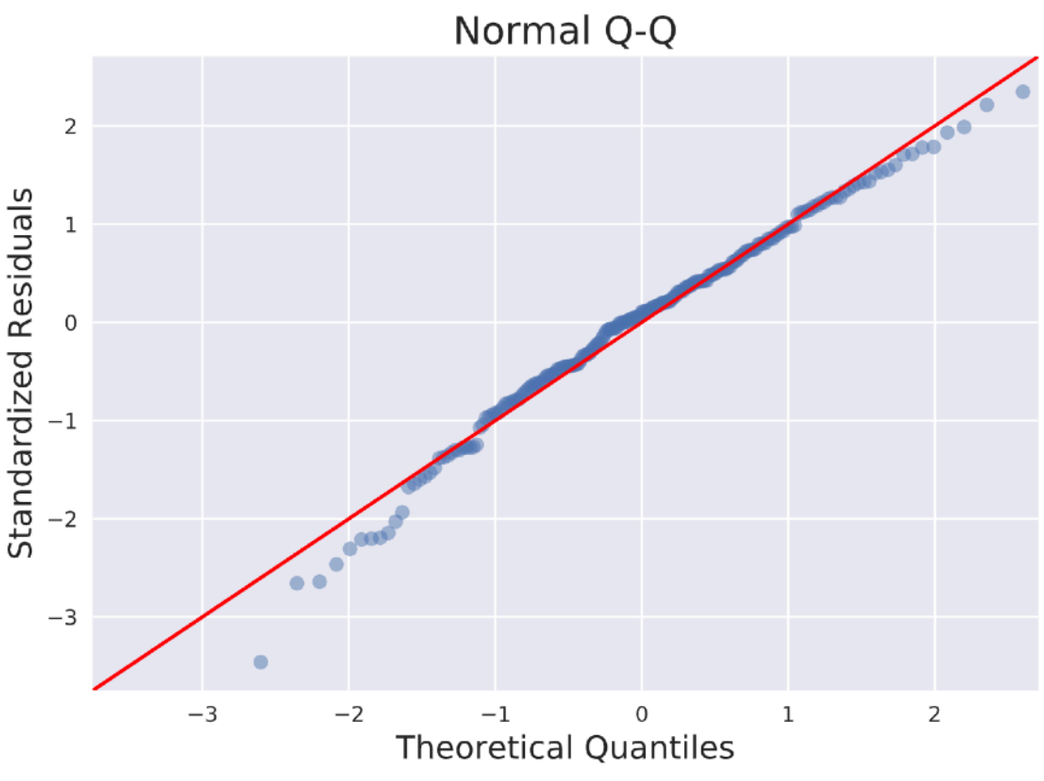
BOX-COX TRANSFORMATION



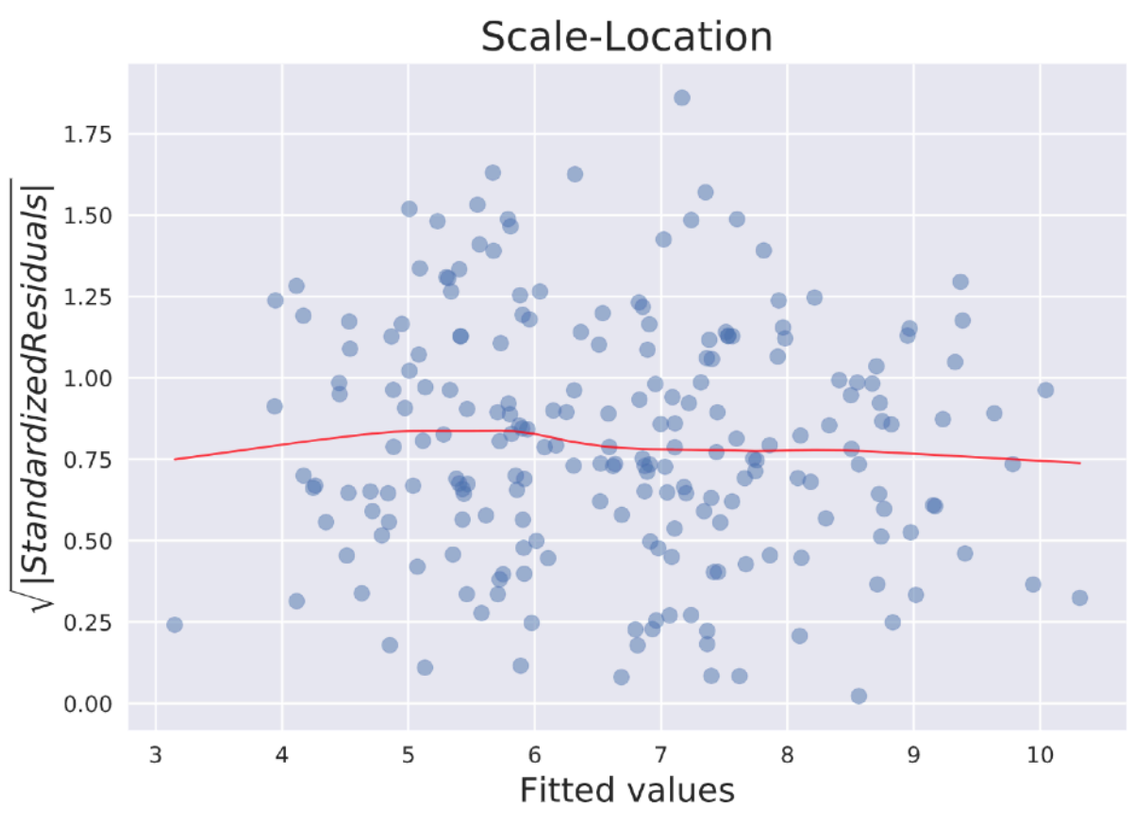
OLS ASSUMPTION IMPROVEMENT

AFTER VARIABLE TRANSFORMATION

ALIGNMENT WITH RED LIGHT INDICATES **NORMALLY DISTRIBUTED ERRORS**



HORIZONTAL RED LINE INDICATES **CONSTANT VARIANCE IN ERRORS**



*Variance plot code from Robert's blog

THE PIPELINE STRUCTURE

TRAIN

VALIDATE

TEST

MODEL TRAINING

SIMPLE REGRESSION

RIDGE REGRESSION

HUBER REGRESSION

} 3-FOLD CV*

MODEL SCORING

DATA SPARSITY

RANDOM RESAMPLING

Problem: < 250 observations led to model instability

Solution: **1000 loops** of different random states

*A separate function searched for best regularizers for Ridge and Huber

MODEL ADJUSTED R-SQUARED

ON 1000 RUNS OF 3-FOLD CV - VALIDATION

0.304
SIMPLE REGRESSION
ADJ. R-SQUARED

0.309
RIDGE REGRESSION
ADJ. R-SQUARED

0.311
HUBER REGRESSION
ADJ. R-SQUARED

TEST

0.342
HUBER REGRESSION
(TEST ADJ. R-SQUARED)

Model performance on unseen data generalizes well!

Features:
Photos, Text Sentiment (-), Text Length, Project Length, # Gift Options

COEFFICIENT INTERPRETATION

CHANGE IN Y FOR ONE UNIT IN X

PROJECT LENGTH

Every additional unit increase (+1 day) in Length leads to **-11% decrease** in pledged amount

TEXT SENTIMENT (-)

Every additional unit increase (+1%) in Negative Sentiment Score leads to **-0.22% decrease** in pledged amount

LOG TRANSFORMED 'Y'

LOG PLEDGED AMOUNT

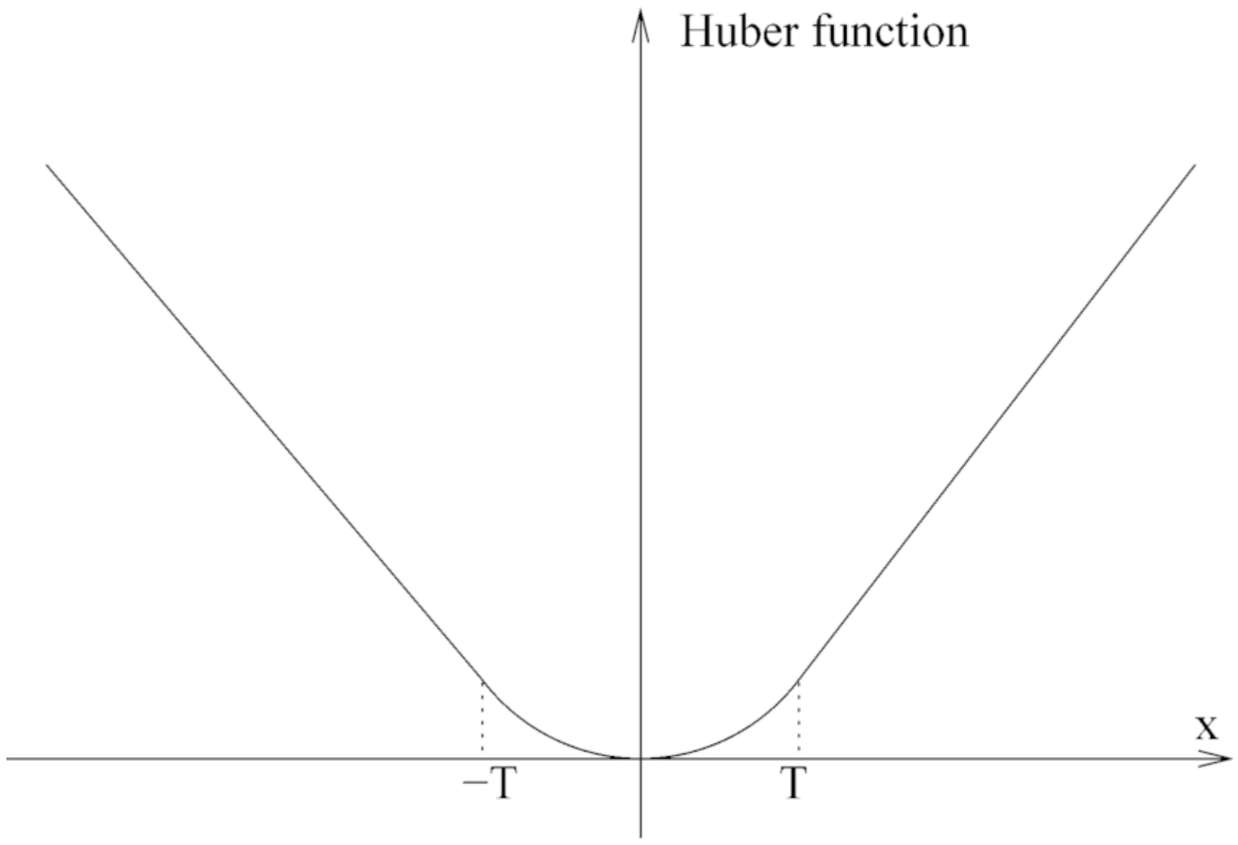
For the non-transformed X-values, we can say a unit increase in x is associated with an increase of

$$\approx 100 * (e^{\beta})$$

percent change in pledged amount

WHY HUBER?

LESS PUNISHING ON OUTLIERS



DATA FILLED WITH OUTLIERS

FROM HIGHLY-FUNDED PROJECTS

Goal was to make sure the loss function wasn't too heavily influenced by outliers, while not completely ignoring their effect.

Huber Regressor solves for this problem explicitly.

Similar to Ridge & Lasso (i.e. regularization parameter).
Squared loss in 'T' Range, Absolute Value loss otherwise.

IMPROVEMENTS

01

SCHEDULED DATA PULLS

Can only scrape new data every few days to get projects near the end of their funding window.

Need to set up a script to run every day and keep pulling the most recent data available.

02

IMPROVED SENTIMENT ANALYSIS

NLTK Vader simply sums up the sentiment for each word in the text using a pre-built dictionary.

Could apply more advanced NLP models on the ‘persuasive-ness’ of the campaign

03

COEFFICIENT INTERPRETATION

Since 3 of the Xs had to be transformed to meet OLS assumptions, the interpretability at the end is loss.

Ideally would find a way to solve for this.

THANK YOU

MARC KELECHAVA

github.com/marcmuon