Udacity Machine Learning NanoDegree

# Avito Demand Prediction Challenge

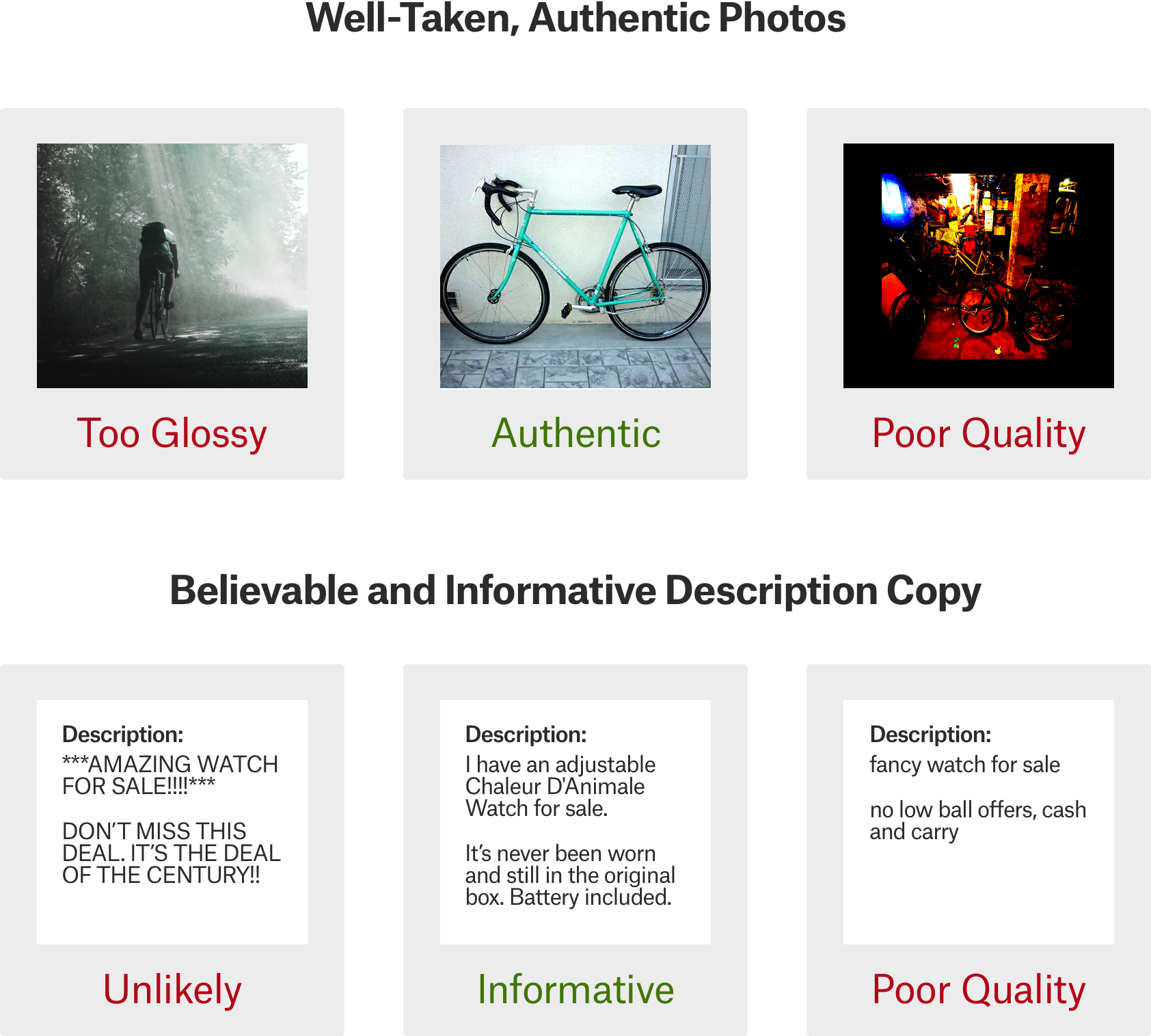
Capstone Proposal

*Sachin Lamba*

Feb 2019

* ***Domain Background***

When selling used goods online, a combination of tiny, nuanced details in a product description can make a big difference in drumming up interest. Details like:



And, even with an optimized product listing, demand for a product may simply not exist–frustrating sellers who may have over-invested in marketing.

[Avito](https://www.avito.ru/), Russia’s largest classified advertisements website, is deeply familiar with this problem. Sellers on their platform sometimes feel frustrated with both too little demand (indicating something is wrong with the product or the product listing) or too much demand (indicating a hot item with a good description was underpriced).

In their fourth Kaggle competition, Avito is challenging you to predict demand for an online advertisement based on its full description (title, description, images, etc.), its context (geographically where it was posted, similar ads already posted) and historical demand for similar ads in similar contexts. With this information, Avito can inform sellers on how to best optimize their listing and provide some indication of how much interest they should realistically expect to receive.

* ***Problem Statement***
* ***Datasets & Inputs***
* ***Abstract***
* ***Attributes Information***
* train.csv - Train data.
  + item\_id - Ad id.
  + user\_id - User id.
  + region - Ad region.
  + city - Ad city.
  + parent\_category\_name - Top level ad category as classified by Avito's ad model.
  + category\_name - Fine grain ad category as classified by Avito's ad model.
  + param\_1 - Optional parameter from Avito's ad model.
  + param\_2 - Optional parameter from Avito's ad model.
  + param\_3 - Optional parameter from Avito's ad model.
  + title - Ad title.
  + description - Ad description.
  + price - Ad price.
  + item\_seq\_number - Ad sequential number for user.
  + activation\_date- Date ad was placed.
  + user\_type - User type.
  + image - Id code of image. Ties to a jpg file in train\_jpg. Not every ad has an image.
  + image\_top\_1 - Avito's classification code for the image.
  + deal\_probability - The target variable. This is the likelihood that an ad actually sold something. It's not possible to verify every transaction with certainty, so this column's value can be any float from zero to one.
* test.csv - Test data. Same schema as the train data, minus deal\_probability.
* train\_active.csv - Supplemental data from ads that were displayed during the same period as train.csv. Same schema as the train data minus deal\_probability, image, and image\_top\_1.
* test\_active.csv - Supplemental data from ads that were displayed during the same period as test.csv. Same schema as the train data minus deal\_probability, image, and image\_top\_1.
* periods\_train.csv - Supplemental data showing the dates when the ads from train\_active.csv were activated and when they where displayed.
  + item\_id - Ad id. Maps to an id in train\_active.csv. IDs may show up multiple times in this file if the ad was renewed.
  + activation\_date - Date the ad was placed.
  + date\_from - First day the ad was displayed.
  + date\_to - Last day the ad was displayed.
* periods\_test.csv - Supplemental data showing the dates when the ads from test\_active.csv were activated and when they where displayed. Same schema as periods\_train.csv, except that the item ids map to an ad in test\_active.csv.
* train\_jpg.zip - Images from the ads in train.csv.
* test\_jpg.zip - Images from the ads in test.csv.
* sample\_submission.csv - A sample submission in the correct format.
  + ***Input Variable***
  + ***Output Variable***
* ***Solution Statement***
* ***Evaluation Metrics***

Root Mean Squared Error (RMSE)

RMSE =

is the predicted value and is the original value.

* ***Project Design***
* ***References***