Final Project
APANPS5335: Machine Learning

Due: Dec 6, 2018

Instructions: Please submit both the RMarkdown file and a PDF version. Make sure that your name and university number are at the top of your Project Report. Also, keep in mind this is a kaggle competition so don't forget to submit your results to leaderboard.

Predict if a painting purchased at auction is a fake.

> One of the biggest challenges of an purchasing a painting at an auction is the risk of that the painting might have serious issues that prevent it from being sold to customers. The artist community calls these unfortunate purchases "fakes","lemons", "kicks" or "Bad Buys". Machine Learning experts who can figure out which paintings have a higher risk of being kick can provide real value to auction houses trying to provide the best inventory selection possible to their customers. The challenge of this competition is to predict if the painting purchased at the Auction is a bad buy.

Dataset Description:

- RefID:Unique (sequential) number assigned to paintings
- IsBadBuy: Identifies if the kicked painting was an avoidable purchase
- PurchDate: The Date the painting was Purchased at Auction
- Auction: Auction provider at which the painting was purchased
- PaintingYear: The year of the painting
- PaintingAge: The Years elapsed since the Painting year
- Artist: painting Manufacturer
- Painting name: painting name. Each painting names may have painting categories embedded like "T1,T2,etc".
- Trim: painting Trim Level
- SubType: painting SubType
- Canvas Color: Canvas Color for painting
- Market: Market for painting (Commercial, Non-commercial)
- FrameTypeID: The type id of the painting frame
- FrameType: The painting frame type description (Wood,Metal)
- Bids: The total bids on paintings.
- Nationality: The artist's country
- Size: The size category of the painting
- TopThreeNYCName: Identifies if the artist is one of the top three NYC artists.
- MMRAcquisitionAuctionAveragePrice: Acquisition price for this painting in average condition at time of purchase

- MMRAcquisitionAuctionCleanPrice: Acquisition price for this painting in the above Average condition at time of purchase
- MMRAcquisitionRetailAveragePrice: Acquisition price for this painting in the retail market in average condition at time of purchase
- MMRAcquisitonRetailCleanPrice: Acquisition price for this painting in the retail market in above average condition at time of purchase
- MMRCurrentAuctionAveragePrice: Acquisition price for this painting in average condition as of current day
- MMRCurrentAuctionCleanPrice: Acquisition price for this painting in the above condition as of current day
- MMRCurrentRetailAveragePrice: Acquisition price for this painting in the retail market in average condition as of current day
- MMRCurrentRetailCleanPrice: Acquisition price for this painting in the retail market in above average condition as of current day
- PRIMEUNIT: Identifies if the painting would have a higher demand than a standard purchase
- AcquisitionType: Identifies how the painting was aquired (Auction buy, trade in, etc)
- AUCGUART: The level guarantee provided by auction for the painting (Green light - Guaranteed/arbitratable, Yellow Light - caution/issue, red light - sold as is)
- KickDate: Date the painting was kicked back to the auction
- BYRNO: Unique number assigned to the buyer that purchased the painting.
- VNZIP: Zipcode where the painting was purchased.
- VNST: State where the the painting was purchased.
- PaintingBCost: Acquisition cost paid for the painting at time of purchase.

- IsOnlineSale: Identifies if the painting was originally purchased online.
- WarrantyCost: Warranty price (term=36month)

Kaggle Competition Rules:

- One account per participant: You cannot sign up to Kaggle from multiple accounts and therefore you cannot submit from multiple accounts.
- Submission Limits: You may submit a maximum of 2 entries per day.
- You may select up to 2 final submissions for judging.
- Competition Deadline: 12/6/2018 12:00 am EST
- Each participant must enter competition with username : firstname_lastname_uniID For example:monil_shah_ms5836
- Training dataset is provided as train_dataset.
- Test dataset is provided as test_dataset.csv
- Sample output csv should contain two columns RefID and IsBadBuy. Please see attached sample submission file(sample_submission.csv)

Grading:

- Kaggle Competition: 40 points
  - Level 1 clearance: 20 points + relative grading for rank.
  - Level 2 clearance: 30 points + relative grading for rank.
- Project Report : 50 points
- Future Work : 10 points

Kickstart your project:

Question 1. (6 points)

1a) (2 points) Load the train_dataset.csv and split the data into train, cross-validation and test dataframes.

1b) (1 points) Plot the scatterplot for current retail clean price vs current retail average price. Is relationship linear?

1c) (3 points) Plot distributions for the Painting age, Bids, warranty costs, acquisitions costs, etc. Which distributions are skewed?

Question 2. (12 points) Feature Engineering

2a) (1 points) Drop the ID variable

2b) (2 points) Extract the painting categories T1-T13 from Painting Name and subtype. Create new columns for above extracted categories with 1 or 0 as outcome values.

2c) (1 points) The dataset describes the number of quality checks. to extract this use Painting Name. Each check is described three ways I4,I-4,I 4 and I6,I-6,I 6.

2d) (2 points) Painting size is also encoded in painting name and subtype. For example 3.5L . Extract that using grep function.

2e) (2 points) Sub Type contains categories of the painting. Genre, History, Still Life, Real Life, Landscape, Portrait and Fine Art. Extract this information.

2f) (2 points) Calculate ratio for prices and bids as part of feature engineering.

2g) (2 points) Compute dummmies for factors (Artist, painting Year, CanvasColor,etc). For example convert categorical variables to dummy indicator variables.

Question 3. (32 points) Try doing more feature engineering by separating day, month, year from purchase date. Fit a model and try predicting if the painting is badbuy. Use advance techniques like k-Fold cross validation, bootstrapping, boosting algorithms, bagging, regularization to improve your model accuracy. Submit your submission.csv on kaggle.

Question 4. (10 points) Given more time, improved data, better computing power, etc . How would you improve your model in the future.