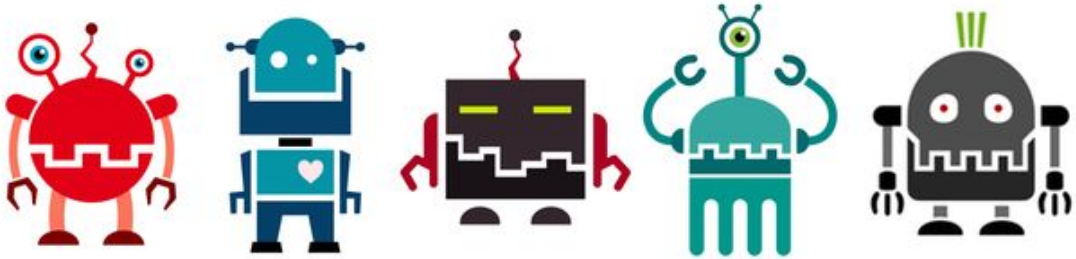
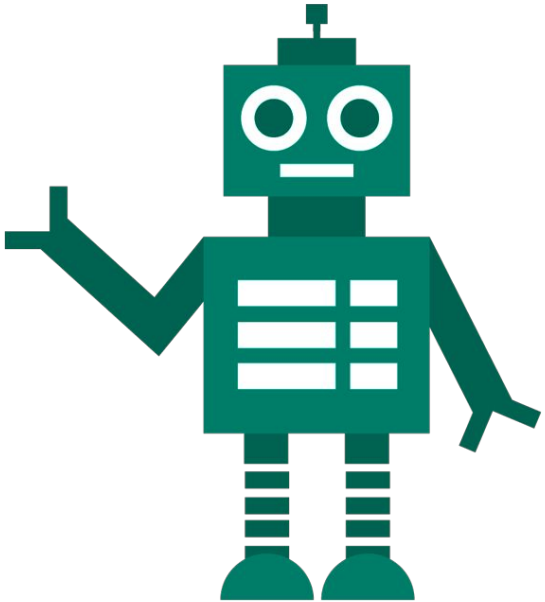
# Bots vs. Humans



# Action plan

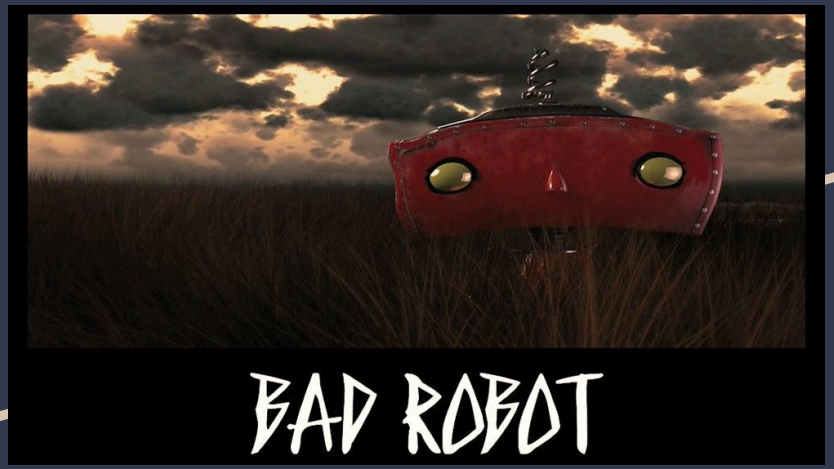
- Project scope
- Data summary
- Models
- Findings
- Next steps



# Predict if an online bid is made by a machine or human


- 2015 Kaggle Competition
- 7.6 million bids on different auctions
- Use classification methods to distinguish between two classes

# Why bots are bad?




- Create a bad experience for real users
- Users less likely to come back to the auction site
- Over 95% of unique users are human

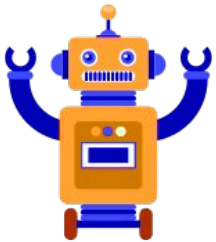
# Initial hiccups

**Class imbalance**

- 7+ million bids
- 1910 real users
- 103 bots

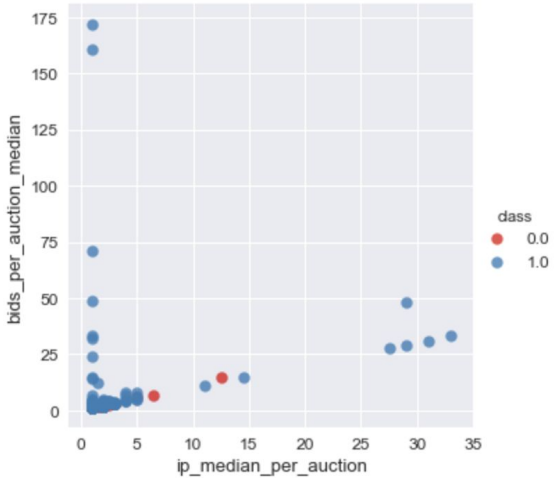
**Time obfuscation**

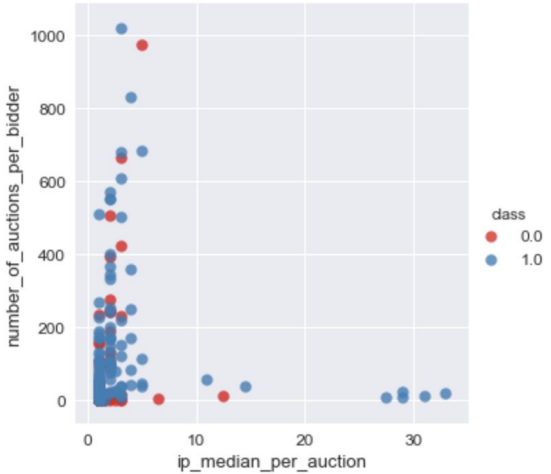
- Time stamps have been hidden
- Ex. 9709222052631578
- Time is essential in spotting bots



# Modeled data

- SVM basic
- SVM RBF kernel
- KNN, GridSearchCV

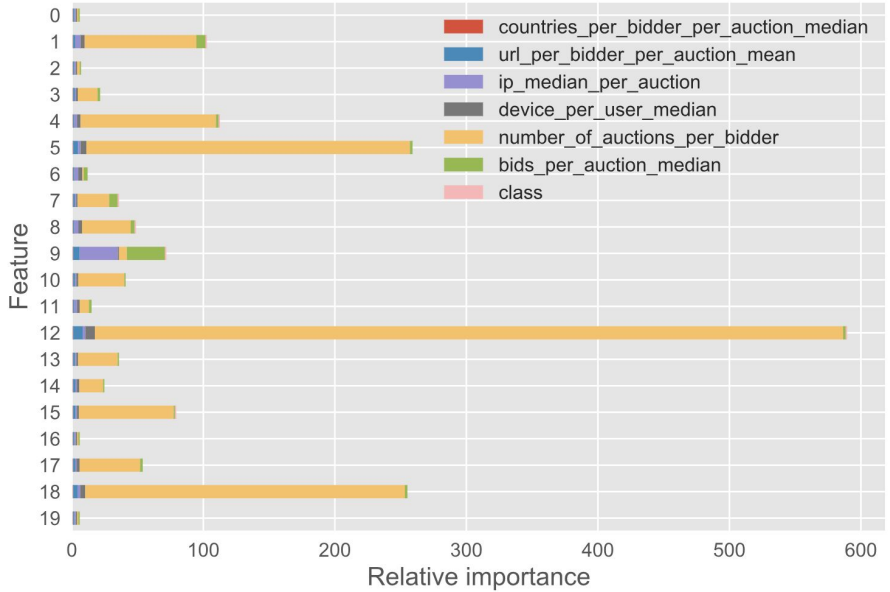


# Model results

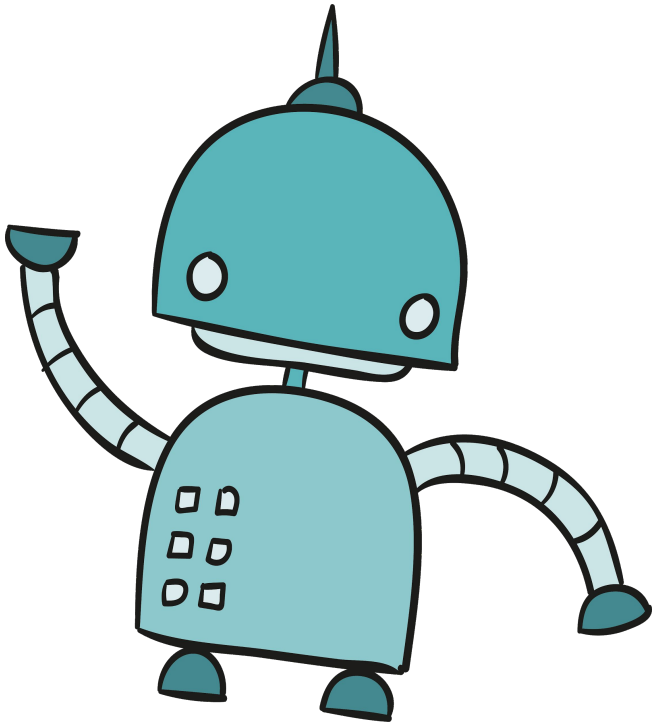
- RBF kernel gives best accuracy scores (0.878)
- Linear split accuracy is lower (.80)
- Best KNN result on K=5 (~.83)

# Importance summary



# Further Steps

- Oversampling the Bots class
- Time stamp
- Other models
  - Random Forest
  - Various Kernels



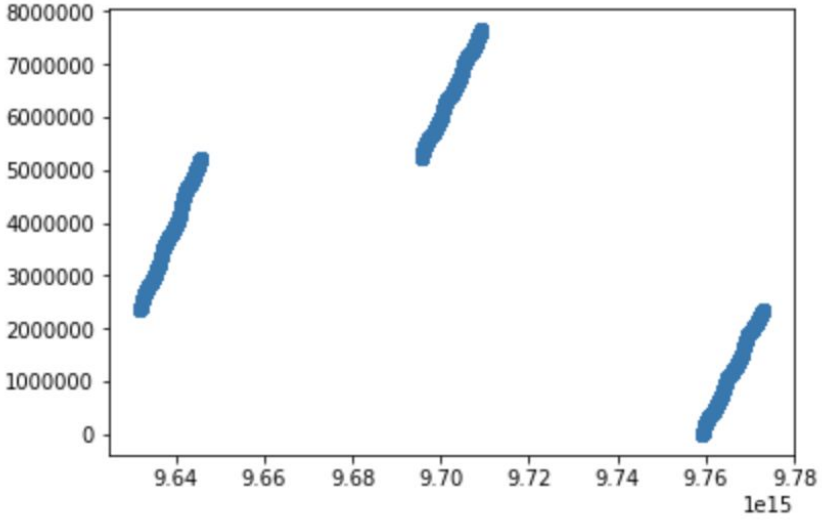
# Thank you!

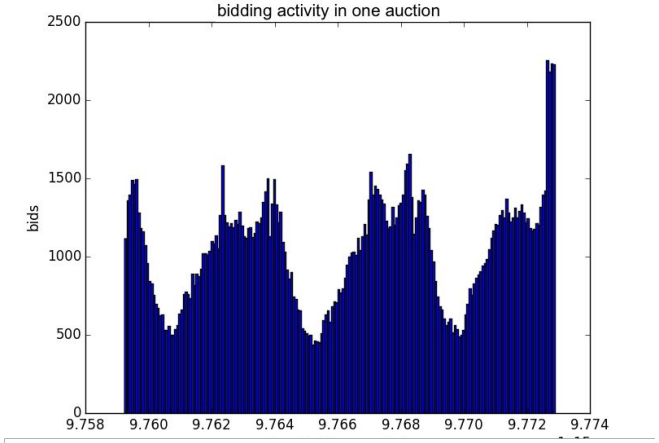
Iris Borkovsky
i.borkovsky@gmail.com

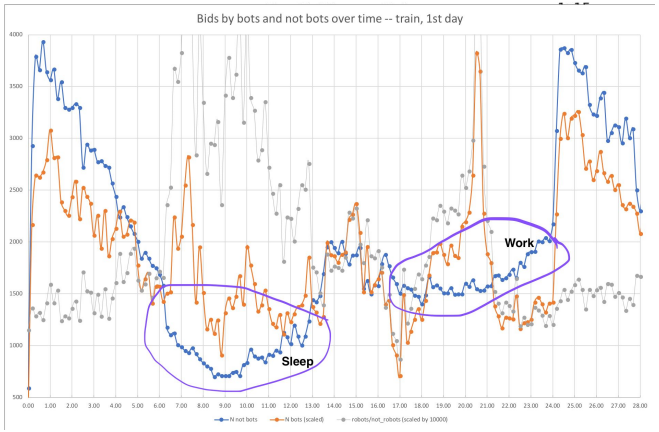

# Appendix

# Dealing with time obfuscation

- Time obfuscation maintains monotonicity
- Use smallest possible time delta from given numbers
- Quadrillion ($10^{15}$)




# Time column EDA

# Class imbalance

- 1910 Classified Humans
- 103 Classified Bots
- Randomly undersampling real bidders before modeling