SHELTER ANIMAL OUTCOMES

kaggle

Bertino Enrico
Donizetti Alberto
Necchi Pierpaolo
Negrini Alessandro
Toschi Marta

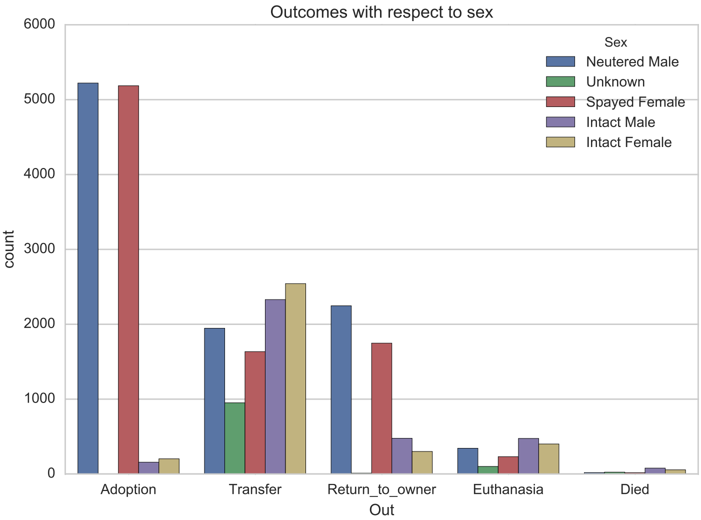
Plan

Introduction

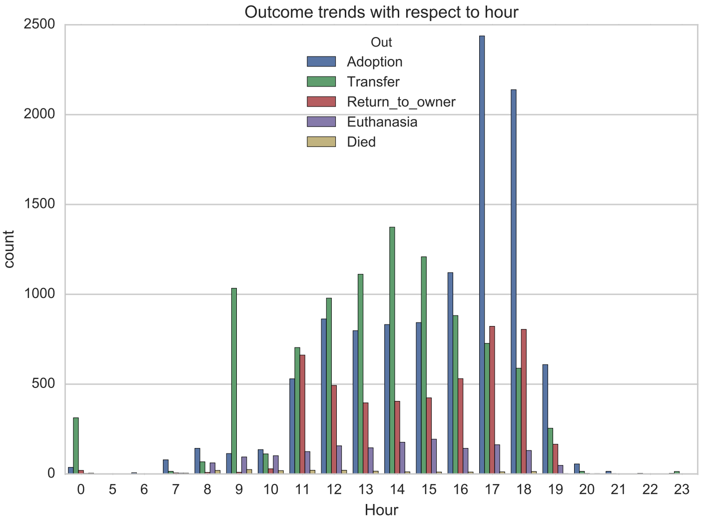
TODO

- Put kaggle logo
- Describe animal shelter competition (animal photo?)

Animal Status



Hourly Patterns



Leak

Two sources of leak

- Data is gathered at outcome time
  - Animal status is a strong outcome predictor
- Training set and test set overlap in time
  - Outcome time provides very rich information

Features

Original Variable	Type	Variables obtained	Type	Le
Name	String	Length of name	Numerical	No
Date and time	Datetime	Year	Numerical	Ye
		Season	Numerical	No
		Holidays	Categorical	No
		Month	Numerical	No
		Day of week	Numerical	No
		Day	Numerical	Ye
		Day of year	Numerical	Ye
		Hour	Numerical	Ye
		Minute	Numerical	Ye
		Minute of day	Numerical	Ye
		Outcomes clusters	Numerical	Ye
Animal type	Categorical	Animal Type	Categorical	No

Outcomes Temporal Clustering

TODO: diagramma?

Classifiers and Software

Random Forests and Xgboost

- High flexibility and ability to handle “mixed” data-types.
- Typically work well out-of-the-box
- Xgboost has proven extremely successful in past Kaggle competitions.
- Quite easy to fine-tune.

May the python be with you

- Pandas
- Scikit-Learn
- Xgboost

Model Validation and Parameter Tuning

- Extracted a stratified holdout set from the training set
- Used early stopping to avoid overfitting in xgboost classifier
- Evaluated several performance metrics on the holdout set
- Tuned xgboost parameters using CV-based grid search
- Bagged several xgboost classifiers to reduce variance

Project Milestones

Description	Score	Leaderboard
Bagged xgboost classifier with no leak	0.91586	667
Added animal status	0.81768	454
Added day, hour and minute information	0.69699	21
Added outcome clusters	0.64574	4
Tuned xgboost parameters by grid search	0.62799	4
Hierarchical xgboost & random forest classifier	0.62713	4

Conclusions and Further Developments

Conclusions

- “Leak” variables have a huge predictive power

Conclusions and Further Developments

Conclusions

- "Leak" variables have a huge predictive power
- Exploiting temporal clusters of outcomes, we reached the 3rd position worldwide.

Conclusions and Further Developments

Conclusions

- "Leak" variables have a huge predictive power
- Exploiting temporal clusters of outcomes, we reached the 3rd position worldwide.
- Fine-tuning of the xgboost parameters yields large improvements in classification accuracy

Conclusions and Further Developments

Conclusions

- "Leak" variables have a huge predictive power
- Exploiting temporal clusters of outcomes, we reached the 3rd position worldwide.
- Fine-tuning of the xgboost parameters yields large improvements in classification accuracy

Conclusions and Further Developments

Conclusions

- "Leak" variables have a huge predictive power
- Exploiting temporal clusters of outcomes, we reached the 3rd position worldwide.
- Fine-tuning of the xgboost parameters yields large improvements in classification accuracy

Further Developments

- Design features to separate adoptions and return to owners

Conclusions and Further Developments

Conclusions

- "Leak" variables have a huge predictive power
- Exploiting temporal clusters of outcomes, we reached the 3rd position worldwide.
- Fine-tuning of the xgboost parameters yields large improvements in classification accuracy

Further Developments

- Design features to separate adoptions and return to owners
- Combine different classifiers able to learn different aspects

References

Tianqi Chen and Tong He.
xgboost: extreme gradient boosting.
R package version 0.4-2, 2015.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani.
The Elements Of Statistical Learning, Data Mining Inference And Prediction.
Springer, 2 edition, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
Scikit-learn: Machine learning in Python.
Journal of Machine Learning Research, 12:2825–2830, 2011.