Kötücül Yazılım Tahmini İçin Pratik Veri Bilimi Tarifleri

Uğur Ünal, Aykut Çayır

Danışmanlar: Prof. Dr. Hasan Dağ, Öğr. Gör. Işıl Yenidoğan Dağ

CCIP
CENTER FOR CYBERSECURITY &
CRITICAL INFRASTRUCTURE PROTECTION

data istanbul

Agenda

- KHAS CCIP
- Introduction to Kaggle and Microsoft Malware Prediction Competition
- Recipe 0: Describe the Dataset with Pandas
- Recipe 1: Feature Engineering with FeatEngine
- Recipe 2: Feature Selection with LOFO
- Recipe 3: Using Gradient Boosting Models for Tabular Datasets (Xgboost, Catboost, LightGBM)
- Recipe 4: Don’t be Overfit!
- Recipe 5: Bayesian Hyperparameter Tuning with Hyperopt
- Recipe 6: Define a Pipeline
- Conclusion

CCIP
CENTER FOR CYBERSECURITY & CRITICAL INFRASTRUCTURE PROTECTION

data istanbul

KHAS CCIP

Kadir Has University's Center for Cybersecurity & Critical Infrastructure Protection (KHAS_CCIP) is one of the first established centres focusing on CIP research, such as security of industrial control systems, energy and other critical infrastructures.

The centre aims to become a "hub" for the researches on Industrial Control Systems both in Turkey and its close region. Centre's research capacity depends on the axis of three disciplines: Engineering and Information Systems, International Relations within the context of Social Science part, and Law.

Based on its multidisciplinary team, KHAS_CCIP approaches Cybersecurity-related research and application issues (e.g. attacks, vulnerabilities, defence, etc.) with an innovative and unconventional methodologies in principal. In that respect, CCIP's mission is to develop innovative methods, tools, and strategies to detect, identify, and mitigate all types of cybersecurity related threats for institutions, companies, and nationally important industrial control systems. More importantly, since "human remains as the weakest link" in cyber security chain, training and education located at the heart of Centre's mission.

Reaching scientific excellency is one of the headline goals of KHAS_CCIP. Thus, R&D activities are divided into three main activity areas: Solutions towards industry and scientific community, trainings for researchers and professionals, and capacity building and networking.

CCIP
CENTER FOR CYBERSECURITY &
CRITICAL INFRASTRUCTURE PROTECTION

data istanbul

Introduction to Kaggle and Microsoft Malware Prediction Competition-I

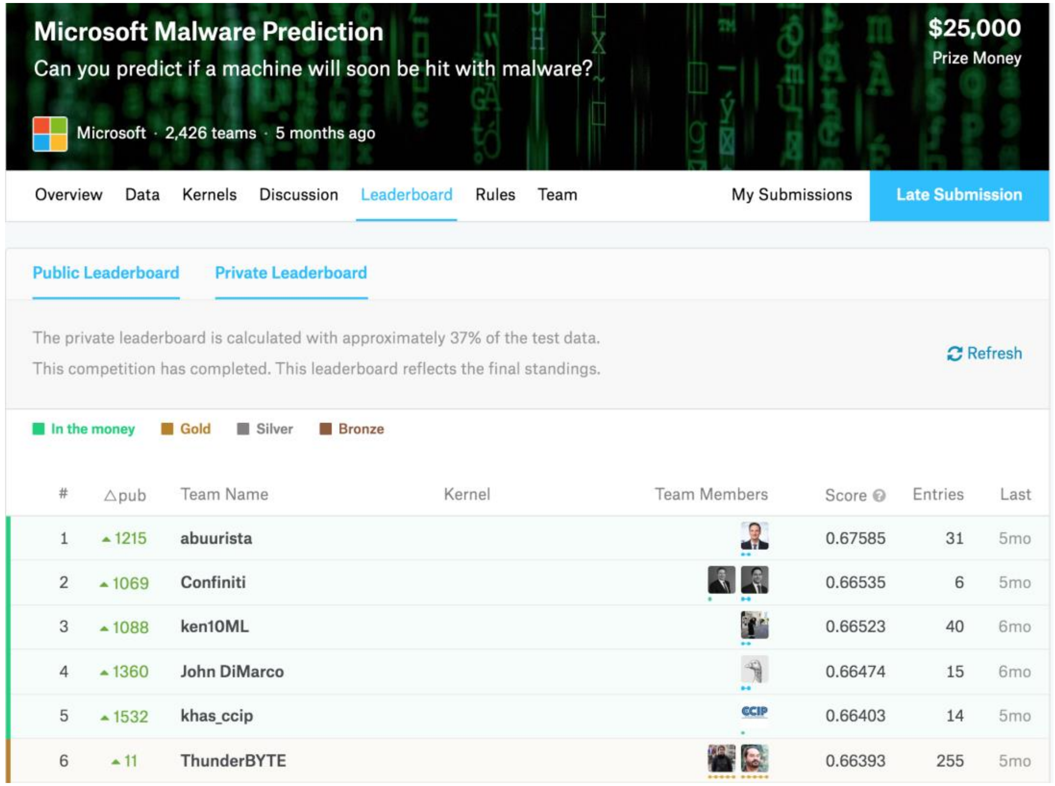
Kaggle

- Founded in April 2010 by Anthony Goldbloom and Ben Hamner
- Acquired by Google in 2017
- Competition Categories (7)
  - Featured
  - Research
  - Recruitment
  - Getting started
  - Masters
  - Playground
  - Analytics

Introduction to Kaggle and Microsoft Malware Prediction Competition-II

Microsoft Malware Prediction

- 2,426 Teams (Worldwide)
- 2,874 Competitors
- 43,696 Entries
- 4 Months
- Total Prize Money: $25,000
  - 1st Place - $12,000
  - 2nd Place - $7,000
  - 3rd Place - $3,000
  - 4th Place - $2,000
  - 5th Place - $1,000
- Evaluation Metric: Area Under ROC



Microsoft Malware Prediction

Can you predict if a machine will soon be hit with malware?

Microsoft · 2,426 teams · 5 months ago

$25,000
Prize Money

The private leaderboard is calculated with approximately 37% of the test data.

This competition has completed. This leaderboard reflects the final standings.

🟩 In the money ■ Gold ■ Silver ■ Bronze

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲1215	abuurista			0.67585	31	5mo
2	▲1069	Confiniti			0.66535	6	5mo
3	▲1088	ken10ML			0.66523	40	6mo
4	▲1360	John DiMarco			0.66474	15	6mo
5	▲1532	khas_ccip		CCIP	0.66403	14	5mo
6	▲11	ThunderBYTE			0.66393	255	5mo

CCIP
CENTER FOR CYBERSECURITY & CRITICAL INFRASTRUCTURE PROTECTION

data istanbul

Recipe 0: Describe the Dataset with Pandas

Original Training Set

- 8.9M Samples
- “HasDetections” is the binary target
- 83 Columns including target
- Balanced

Original Test Set

- 7.8M Samples
- 82 Columns
- No targets

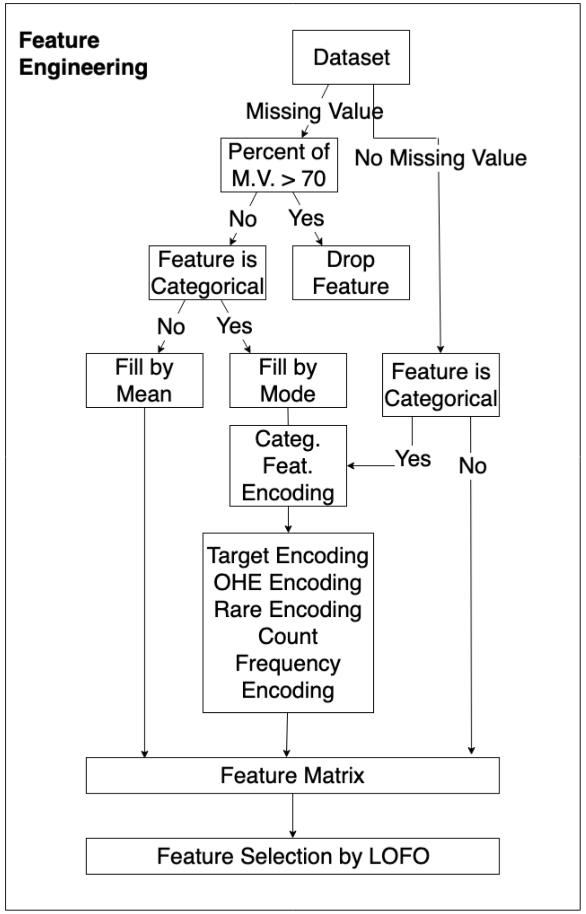
Recipe 1: Feature Engineering with FeatEngine

*Categorical Feature Encoding in for the Dataset

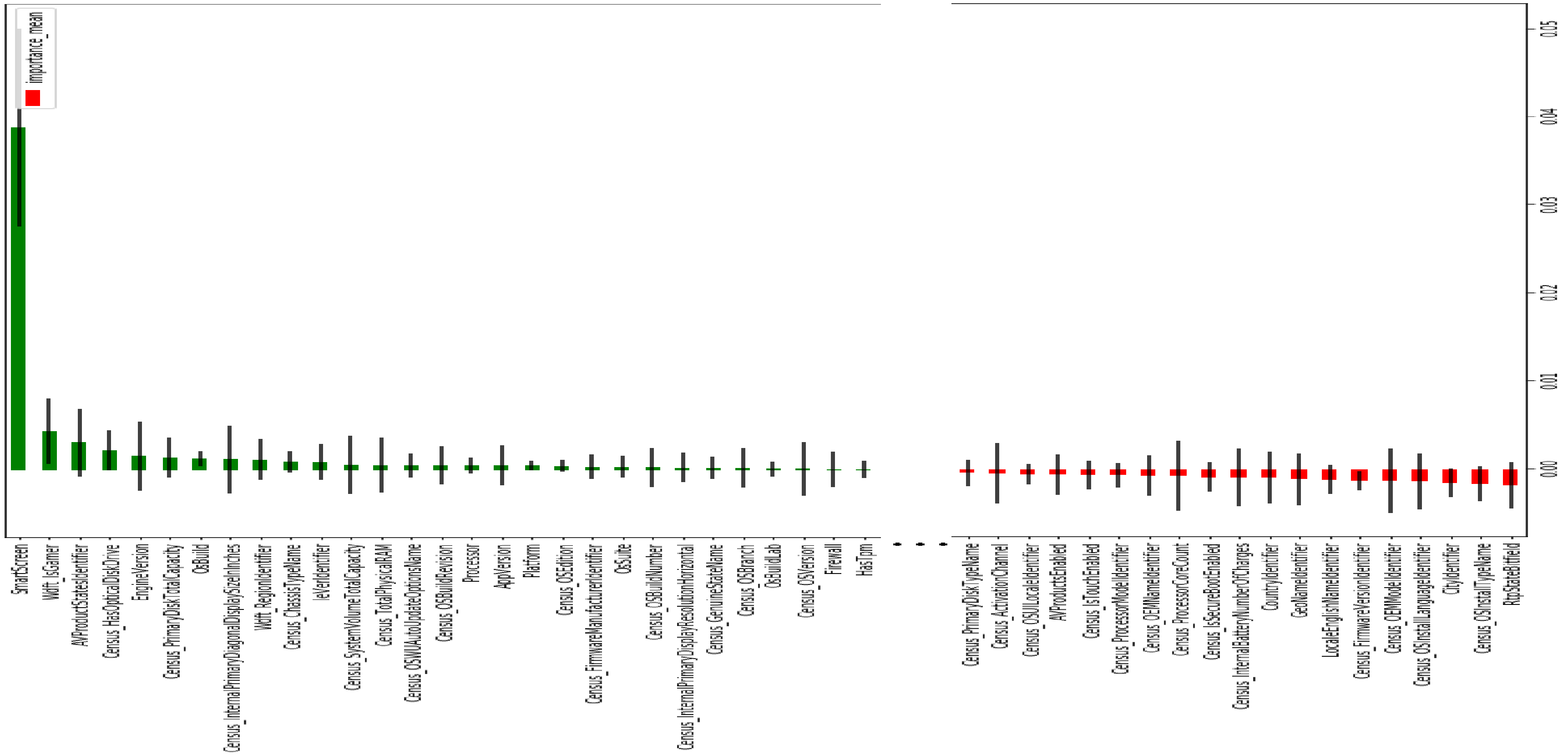
Because of high cardinality of categorical features, One-hot encoding has not been used

Using target encoding has increased overfitting

Using Rare Encoding + Count Frequency Encoding has reduced overfitting (Threshold for Rare Encoding is 0.005)



Recipe 2: Feature Selection with LOFO

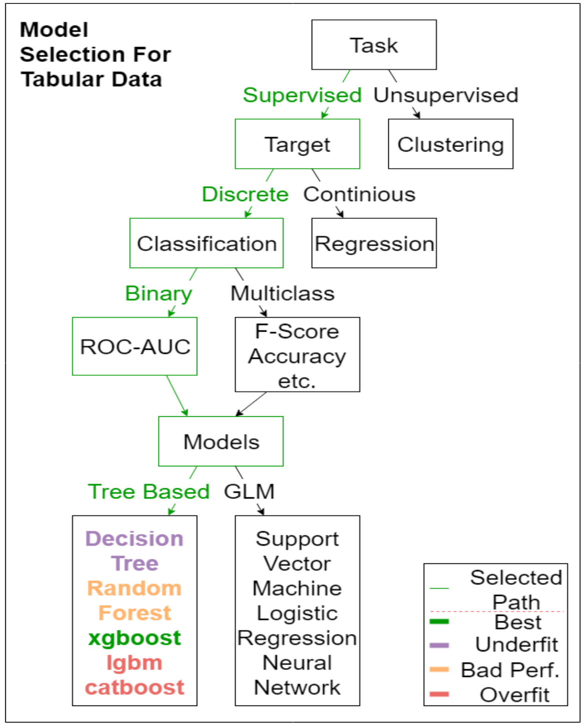


CCIP
CENTER FOR CYBERSECURITY &
CRITICAL INFRASTRUCTURE PROTECTION

data istanbul

Recipe 3: Using Gradient Boosting Models for Tabular Datasets (Xgboost, Catboost, LightGBM)

- There is “no free lunch” theorem in ML & DS, but if your dataset is tabular then you must try tree based models and their ensemble versions.
- XGBoost model has been selected, because our feature engineering methods give the best results with XGBoost



CCIP
CENTER FOR CYBERSECURITY & CRITICAL INFRASTRUCTURE PROTECTION

data istanbul

Recipe 4: Don’t Be Overfit!

- Create a holdout validation set
  - 90% of the original training set is used for training phase
  - 10% of the original training set is used for holdout set for validation step
- Compare AUC (or what the metric of the competition is) score of training to validation score.
- They must be close to each other!
- Kaggle Competition Results

  Public Leaderboard

    63% of Original Test Set

  Private Leaderboard

    37% of Original Test Set

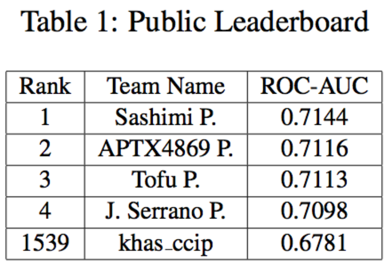


Table 1: Public Leaderboard

Rank	Team Name	ROC-AUC
1	Sashimi P.	0.7144
2	APTX4869 P.	0.7116
3	Tofu P.	0.7113
4	J. Serrano P.	0.7098
1539	khas_ccip	0.6781

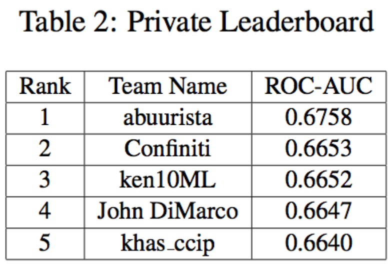


Table 2: Private Leaderboard

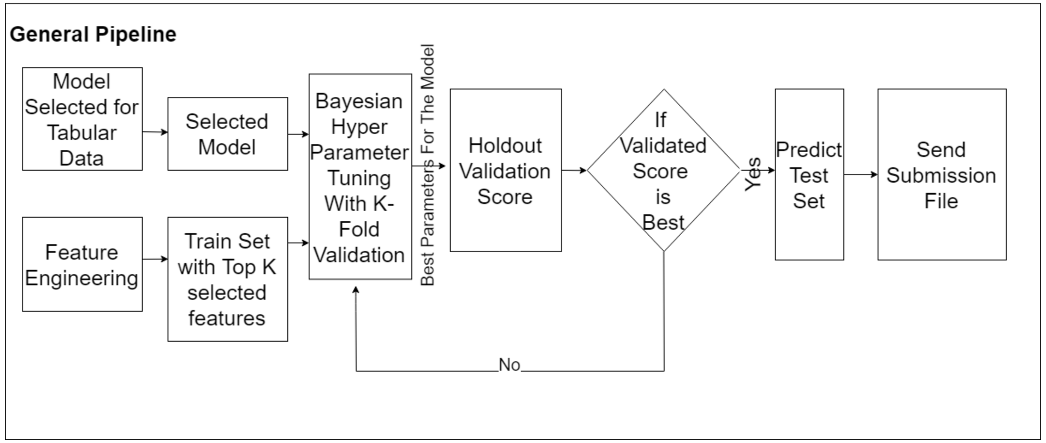
Rank	Team Name	ROC-AUC
1	abuurista	0.6758
2	Confiniti	0.6653
3	ken10ML	0.6652
4	John DiMarco	0.6647
5	khas_ccip	0.6640

Recipe 5: Bayesian Hyperparameter Tuning with Hyperopt

- Use Bayesian optimization
  - Best framework is Hyperopt
- Maximize AUC score of K-Folds
  - Use reasonable K values because if your dataset has many samples it takes too much time. In our case, optimization phase took 12 hours
- Don’t forget to validate your best model on your holdout validation set

Recipe 6: Define a Pipeline

- Design a pipeline
  - Our pipeline is generic and it can be applied for any data science competition that uses tabular dataset
  - A good pipeline is useful for a real world implementation of the competition (Microsoft Cyber Security Research Team thinks that our solution is very applicable and reproducible)
  - Designing a pipeline reduces your code complexity



General Pipeline

Model Selected for Tabular Data

Selected Model

Feature Engineering

Train Set with Top K selected features

Bayesian Hyper Parameter Tuning With K-Fold Validation

Best Parameters For The Model

Holdout Validation Score

If Validated Score is Best

Yes

No

Predict Test Set

Send Submission File

Conclusion

- Each recipe is applicable for all data science problems and competitions independent from domain.
- If the dataset is tabular then tree based ensemble models are very useful for your problem. If your dataset is unstructured (images, voice, signals, videos, etc.) then you must think deep learning architectures.
- Be careful to avoid being overfit.
- Designing a pipeline makes your solution simple and reproducible.

CCIP
CENTER FOR CYBERSECURITY &
CRITICAL INFRASTRUCTURE PROTECTION

data
istanbul