

How to Kaggle

The background of the image is a dark purple color. Overlaid on it are several abstract geometric shapes in various colors: a large dark blue circle in the center containing the text; a pink circle with horizontal pink stripes to the left; a blue circle with horizontal blue stripes to the right; a purple circle with a dotted pattern on the bottom left; a yellow triangle pointing up on the left; a yellow triangle pointing down on the bottom left; a yellow dashed triangle pointing down in the center; a pink dashed circle on the bottom left; a blue dashed circle on the top left; a pink dashed triangle on the top right; a yellow dashed triangle on the bottom right; and a pink pentagon on the bottom right. There are also some thin, dashed lines connecting some of the shapes.



Carlos Sevilla
Barceló



@carlosevi94



Rodrigo Gómez
Rodríguez



@rodgomrod

Summary

- What is Kaggle?
- Our experience
- Competition details

What is Kaggle?

kaggleTM

What is Kaggle?



Some Stats

Users

110,624 registered users

10,000 active users per day

Competitions

332 total competitions

72% with money prizes

10,469,727\$ given prizes

What is Kaggle?

Kaggle Tiers

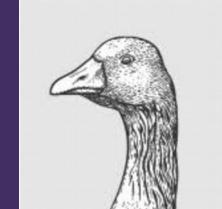
 141
Grandmasters

 1,185
Masters

 4,240
Experts

 47,772
Contributors

 57,286
Novices



What is Kaggle?



José A. Guerrero

Sevilla, Spain

Joined 8 years ago · last seen a month ago

Followers 126



Competitions
Grandmaster



Javier Tejedor Aguilera

Non Technical Losses Detection at Endesa (Enel Group)

Seville, Spain

Joined 4 years ago · last seen in the past day



Followers 35

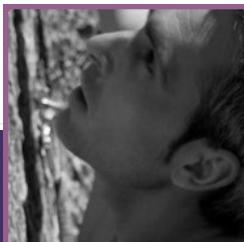
Following 18



Competitions
Master



Juan Galán



David Solis

Seville, Spain

Joined 4 years ago · last seen 4 months ago

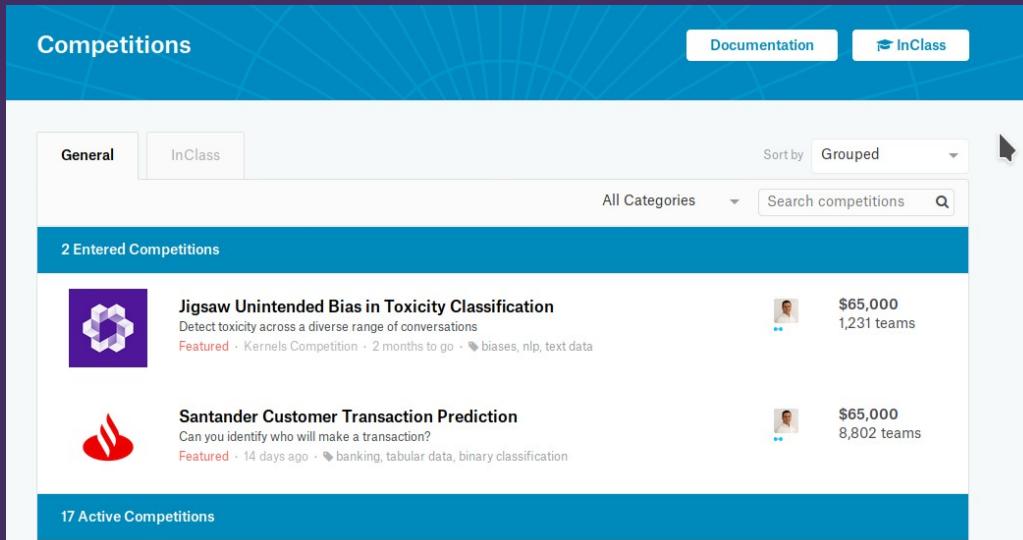


Competitions
Expert

Followers 3

What is Kaggle?

Competitions



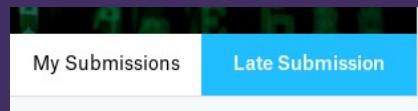
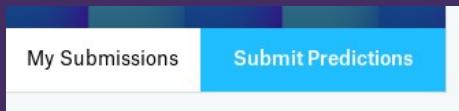
The image is a screenshot of the Kaggle Competitions page. At the top, there are buttons for 'Documentation' and 'InClass'. Below that, there are tabs for 'General' and 'InClass', and a dropdown for 'Sort by' set to 'Grouped'. There are also buttons for 'All Categories' and a search bar. The main content area shows '2 Entered Competitions':

- Jigsaw Unintended Bias in Toxicity Classification**
Detect toxicity across a diverse range of conversations
Featured · Kernels Competition · 2 months to go · biases, nlp, text data
\$65,000
1,231 teams
- Santander Customer Transaction Prediction**
Can you identify who will make a transaction?
Featured · 14 days ago · banking, tabular data, binary classification
\$65,000
8,802 teams

At the bottom, there is a section for '17 Active Competitions'.

What is Kaggle?

How to compete



What is Kaggle?

How to compete

Make a submission for [MapsTeam](#)

Step 1
Upload submission file

Upload Files

File Format
Your submission should be in CSV format. You can upload this in a zip/gz /rar/7z archive, if you prefer.

Number of Predictions
We expect the solution file to have 7853253 prediction rows. This file should have a header row. Please see sample submission file on the [data page](#).

Step 2
Describe submission

Briefly describe your submission

Make Submission

What is Kaggle?

How to compete

448	libinglin		0.69747	5	2mo
449	DL		0.69747	25	1mo
450	MapsaTeam		0.69746	37	1mo
Your Best Entry ↑					
Your submission scored 0.67762, which is not an improvement of your best score. Keep trying!					
451	wenliangth		0.69746	61	1mo
452	Experto		0.69746	5	1mo

What is Kaggle?

Leaderboard

[Public Leaderboard](#)

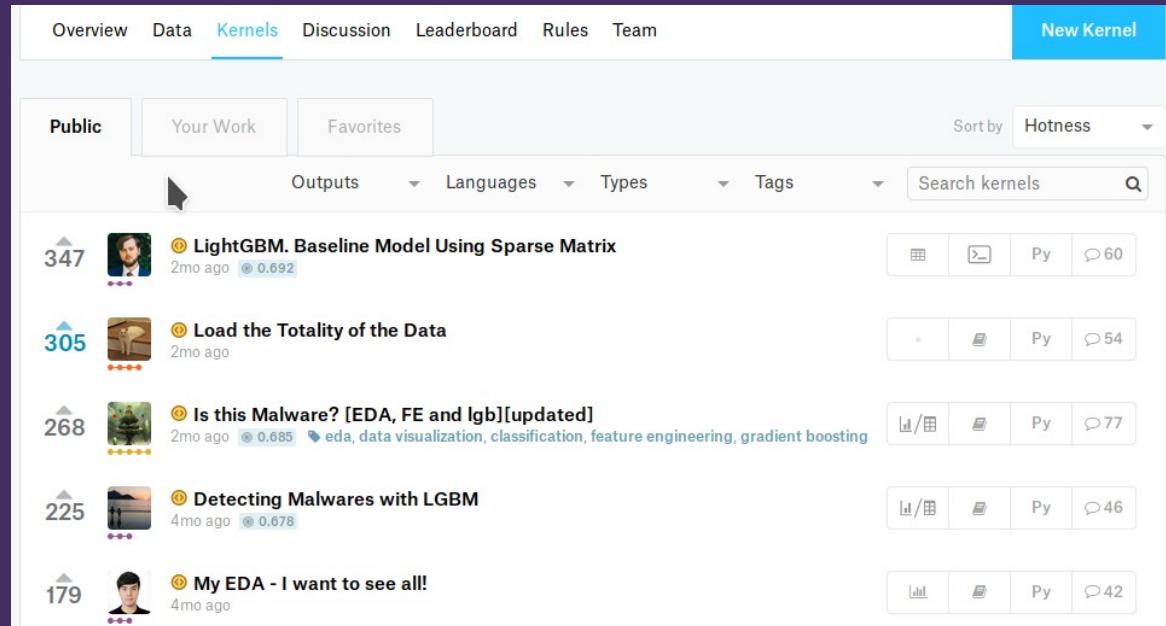
[Private Leaderboard](#)

This leaderboard is calculated with approximately 63% of the test data.

The final results will be based on the other 37%, so the final standings may be different.

What is Kaggle?

Kernels

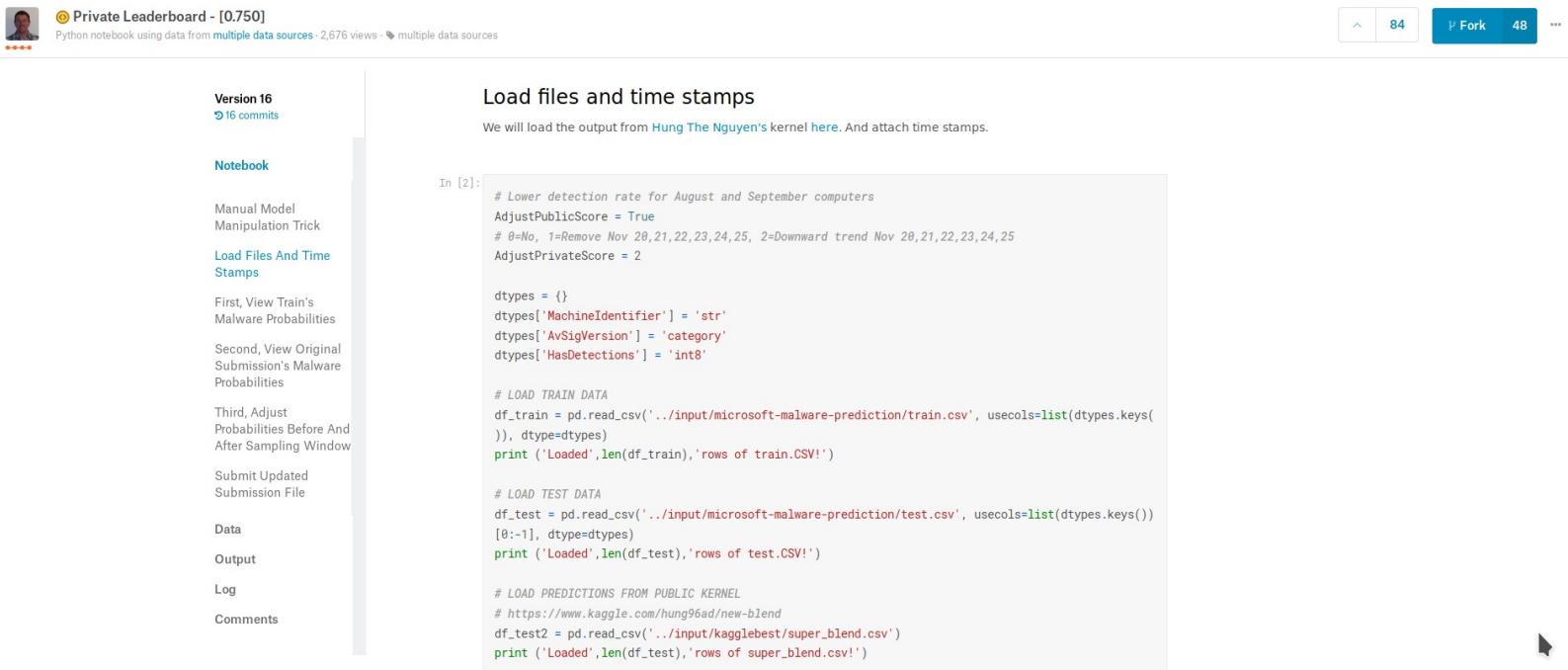


The screenshot shows the Kaggle Kernels interface. The top navigation bar includes links for Overview, Data, Kernels (which is underlined in blue), Discussion, Leaderboard, Rules, and Team, along with a 'New Kernel' button. Below the navigation is a filter bar with tabs for Public, Your Work, and Favorites, and dropdowns for Outputs, Languages, Types, and Tags, along with a search bar. The main content area displays a list of public kernels sorted by Hotness. Each kernel entry includes a rank, profile picture, title, last updated time, and a small icon indicating it's a notebook. To the right of each entry are buttons for viewing the kernel in a notebook or script format and a count of comments.

Rank	Profile Picture	Title	Last Updated	Comments	
347		LightGBM. Baseline Model Using Sparse Matrix	2mo ago @ 0.692	60	
305		Load the Totality of the Data	2mo ago	54	
268		Is this Malware? [EDA, FE and lgb][updated]	2mo ago @ 0.685	eda, data visualization, classification, feature engineering, gradient boosting	77
225		Detecting Malwares with LGBM	4mo ago @ 0.678	46	
179		My EDA - I want to see all!	4mo ago	42	

What is Kaggle?

Kernels



Private Leaderboard - [0.750]
Python notebook using data from [multiple data sources](#) · 2,676 views · [multiple data sources](#)

Version 16
16 commits

Notebook

- Manual Model Manipulation Trick
- Load Files And Time Stamps
- First, View Train's Malware Probabilities
- Second, View Original Submission's Malware Probabilities
- Third, Adjust Probabilities Before And After Sampling Window
- Submit Updated Submission File
- Data
- Output
- Log
- Comments

Load files and time stamps

We will load the output from [Hung The Nguyen's kernel here](#). And attach time stamps.

In [2]:

```
# Lower detection rate for August and September computers
AdjustPublicScore = True
# 0=No, 1=Remove Nov 20,21,22,23,24,25, 2=Downward trend Nov 20,21,22,23,24,25
AdjustPrivateScore = 2

dtypes = {}
dtypes['MachineIdentifier'] = 'str'
dtypes['AvSigVersion'] = 'category'
dtypes['HasDetections'] = 'int8'

# LOAD TRAIN DATA
df_train = pd.read_csv('../input/microsoft-malware-prediction/train.csv', usecols=list(dtypes.keys()), dtype=dtypes)
print ('Loaded',len(df_train),'rows of train.CSV!')

# LOAD TEST DATA
df_test = pd.read_csv('../input/microsoft-malware-prediction/test.csv', usecols=list(dtypes.keys())[0:-1], dtype=dtypes)
print ('Loaded',len(df_test),'rows of test.CSV!')

# LOAD PREDICTIONS FROM PUBLIC KERNEL
# https://www.kaggle.com/hung96ad/new-blend
df_test2 = pd.read_csv('../input/kagglebest/super_blend.csv')
print ('Loaded',len(df_test2),'rows of super_blend.csv!')
```

What is Kaggle?

Discussion

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions New Topic

304 topics Follow Sort by Recent Comments

All Mine Upvoted Search topics

11			Leaderboard is Finalized, Congratulations to our Winners Addison Howard a month ago	last comment by dierdier 1mo ago	 5
31			Competition Submissions, Academic License Request Addison Howard 4 months ago	last comment by airwindow 1mo ago	 13
23			Welcome! Rob McCann 4 months ago	last comment by Chris Deotte 2mo ago	 23
2			Load dataset ruchibahl 4 months ago	last comment by Mohit Singh 6d ago	 9
18			Summary of Microsoft Malware Shakeup Chris Deotte a month ago	last comment by Rob Rose 11d ago	 4
12			Worst Top Shakeup? NxGTR a month ago	last comment by Rob Rose 11d ago	 21

What is Kaggle?

Discussion

Overview Data Kernels **Discussion** Leaderboard Rules Team My Submissions New Topic

 **Carlos Sevilla**
9th place

[Edit](#) | [Options](#)

Is Neural Networks working for you?
posted in [Microsoft Malware Prediction](#) 2 months ago

Me team is using LightGBM. We wanna try use a Neural Network, but I'm doing some test, and my accuracy score is so low (0.56).
I'm trying with Dense layers, but I see that Embedding could works fine.
Anybody want to share experience? I know we are nearly to deadline, but I only want a +0.6 accuracy in NN

Comments (82) Sort by [Newest](#)

 Click here to enter a comment...

 **Chris Deotte** • (1435th in this Competition) • a month ago • [Options](#) • [Reply](#) 0

For those of you that are interested in seeing an embeddings NN kernel. I posted my NN [here](#). If you run it multiple times, it scores 0.696 (+ - 0.001) public LB and 0.770 (+ - 0.003) private LB. If you ensemble five of them, it scores 0.002 LB higher.

What is Kaggle?

Kaggle Community



What is Kaggle?

Competitive spirit



What is Kaggle?



What is Kaggle?

1	Wizardry		0.92727	91	14d
2	三人寄れば文殊の知恵（本当か？		0.92580	101	14d
879	Daisuke N		0.90116	14	14d
880	user1392		0.90116	26	14d
881	Leandro Mineti		0.90115	5	19d
882	riSun		0.90115	7	14d
4014	MuP62		0.90000	5	2mo
4015	Kevin Tang		0.90000	5	14d
4016	newbeeeee		0.89999	5	1mo
4017	San Check		0.89999	2	2mo

What is Kaggle?

Nobody loses at Kaggle

Every second that you're testing some stuff,
you're learning. And that improves you.

What is Kaggle?



Our experience

Mapsa Team



Our experience

Carlos

Rodrigo

Our experience



The screenshot shows the homepage of the Microsoft Malware Prediction competition on Kaggle. The title 'Microsoft Malware Prediction' is displayed, along with the tagline 'Can you predict if a machine will soon be hit with malware?'. It shows 2,426 teams participated a month ago. The competition offers \$25,000 in prize money. The 'Late Submission' tab is currently selected in the navigation bar.

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Late Submission



The 'Overview' page provides a summary of the competition. It includes sections for Description, Evaluation, Prizes, and Timeline. The Prizes section notes that the malware industry is well-organized and funded, and that Microsoft takes the problem seriously. The Timeline section highlights that Microsoft has over one billion enterprise and consumer customers and is deeply invested in improving security. The Description section explains that the competition challenges data scientists to predict malware attacks. The page also features a Microsoft logo composed of four colored squares (orange, green, blue, yellow) arranged in a 2x2 grid.

Overview

Description	The malware industry continues to be a well-organized, well-funded market dedicated to evading traditional security measures. Once a computer is infected by malware, criminals can hurt consumers and enterprises in many ways.
Evaluation	
Prizes	
Timeline	With more than <i>one billion</i> enterprise and consumer customers, Microsoft takes this problem very seriously and is deeply invested in improving security.

As one part of their overall strategy for doing so, Microsoft is challenging the data science community to develop techniques to predict if a machine will soon be hit with malware. As with their previous, [Malware Challenge \(2015\)](#), Microsoft is providing Kagglers with an unprecedented malware dataset to encourage open-source progress on effective techniques for predicting malware occurrences.

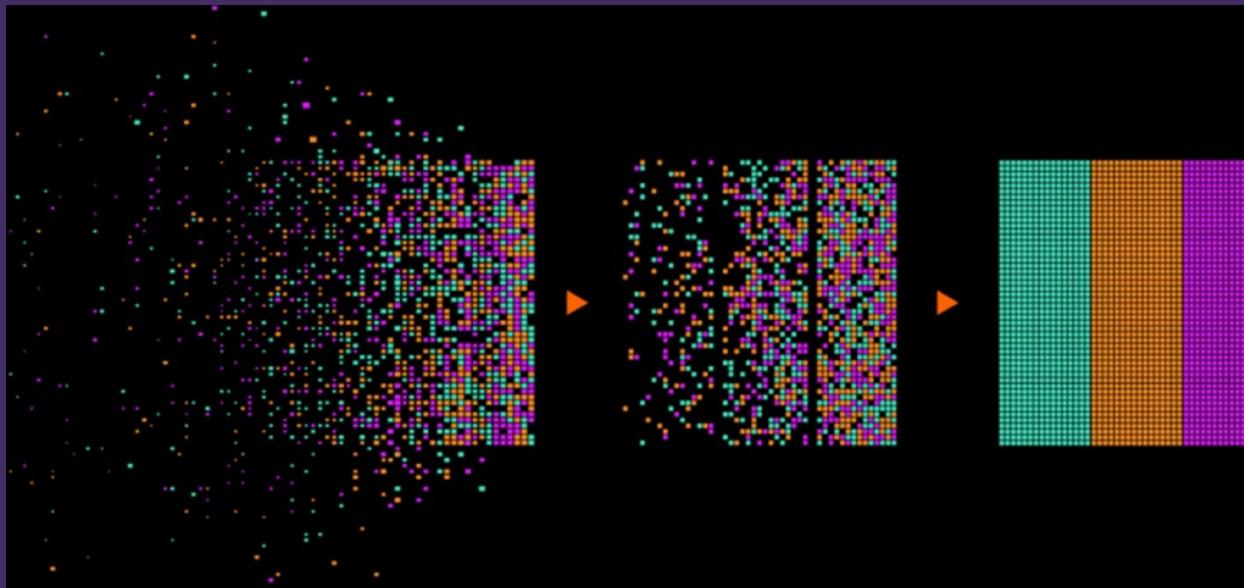
Can you help protect more than one billion machines from damage BEFORE it happens?

Our experience

Carlos

Rodrigo

Our experience



Our experience

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

 Microsoft · 1,800 teams · 22 days to go (15 days to go until merger deadline)

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Submit Predictions

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
<input type="radio"/> Next in line	just now	0 seconds	0 seconds	

Deep learning is shallow next to the depths of your greatness...

We are scoring your submission...

C

Our experience

1801	Badner		0.468	1	6h
1802	MapsaTeam		0.340	1	9m
Your Best Entry ↑					
	Your submission scored 0.340	 Tweet this!			
1803	dict		0.336	1	7d

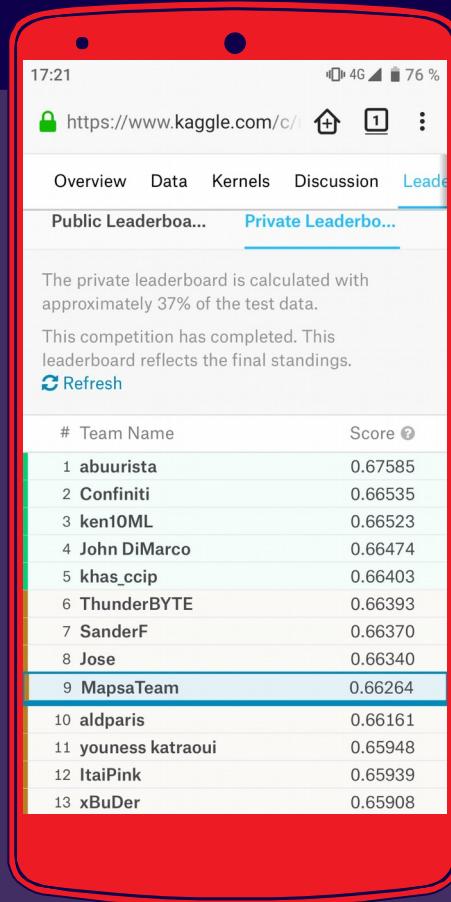
Our experience

Maps
Team

Our experience



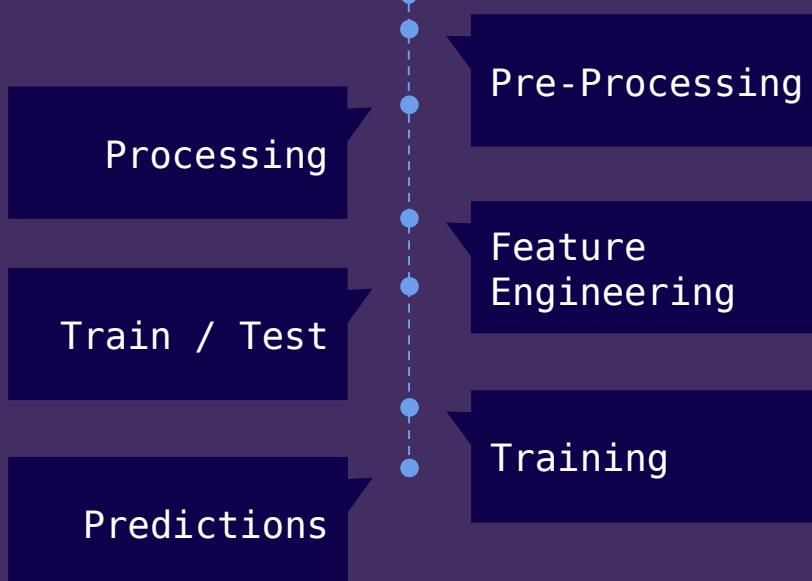
Our experience



Competition details

Competition details

Modeling



Competition details

Pre-Processing

- Train and test union
- Numerical / categorical split
- Data cleaning

```
def transformaciones_OSEdition(x):  
    if x == '#':  
        return None  
    elif x == '00426-OEM-8992662-00006':  
        return 'Ultimate'  
    elif x == 'HomePremium' or x == 'HomeBasic':  
        return 'Home'  
    elif x == 'Window 10 Enterprise' or x == 'Enterprise 2015 LTSB' or x == 'EnterpriseN' \  
        or x == 'EnterpriseE':  
        return 'Enterprise'  
    elif x == 'ServerDatacenterACor' or x == 'ServerDatacenterEval':  
        return 'ServerDatacenter'  
    elif x == 'ProfessionalSingleLanguage' or x == 'PRO' or x == 'Pro' or x == 'professional' \  
        or x == 'ProfessionalCountrySpecific':  
        return 'Professional'  
    elif x == 'ProfessionalEducationN' or x == 'EducationN' or x == 'ProfessionalEducation':  
        return 'Education'  
    elif x == 'CloudN':  
        return 'Cloud'  
    elif x == 'ProfessionalWorkstationN':  
        return 'ProfessionalWorkstation'  
    else:  
        return x  
  
udf_OSEdition = udf(lambda z: transformaciones_OSEdition(z), StringType())
```

Competition details

Processing

- Label Encoding
- NaN imputation
- Version split & L.E.

```
data = data.withColumn('Census_InternalBatteryType_informed',
                      when(col('Census_InternalBatteryType').isNotNull(),1).otherwise(0))

print('\nPipeline de Indexers para las columnas {0}\n'.format(cols_le))
indexers = [StringIndexer(inputCol=c, outputCol=c+"_index", handleInvalid="keep").fit(data) for c in cols_le]
pipeline = Pipeline(stages=indexers)
data = pipeline.fit(data).transform(data)
data = data.drop(*cols_le)

print("Persist intermedio 0\n")
data.persist()
print(data.first())

# Transformamos las columnas de versiones "x.y.z.t"
# Conversion a LabelEncoding

# print('Transformacion columnas de versiones\n')
print('\tCensus_OSVersion\n')
data = data.withColumn('Census_OSVersion_0', concat(split(data['Census_OSVersion'], '\'.')[0],
                                                 split(data['Census_OSVersion'], '\'.')[1]))
.withColumn('Census_OSVersion_1', concat(split(data['Census_OSVersion'], '\'.')[0],
                                                 split(data['Census_OSVersion'], '\'.')[1],
                                                 split(data['Census_OSVersion'], '\'.')[2]))
```

Competition details

Feature Engineering

- Frequency encoding
- Dates
- Extra info.

```
N1 = ['Census_OSEdition', 'SmartScreen', 'Census_OSBranch', 'Census_OSSkuName']

N2 = ['Census_ChassisTypeName', 'Census_FlightRing', 'Census_MDC2FormFactor', 'ProductName']

N3 = ['Census_PrimaryDiskTypeName', 'Census_PowerPlatformRoleName', 'Census_ProcessorClass']

N4 = ['Processor', 'Census_OSIinstallTypeName', 'OsVer', 'Census_GenuineStateName', 'PuMode']

N5 = ['Census_OSVersion', 'EngineVersion', 'AppVersion', 'AvSigVersion']

write_path = 'data/df_groupby_cat_0'

spark = SparkSession.builder.appName("Microsoft_Kaggle").getOrCreate()

data = spark.read.csv('data/df_cat_prep0/*.csv', header=True, inferSchema=True).select(N1+N2+N3+N4+N5+['MachineIdentifier'])

print('Guardamos el DF en {}'.format(write_path))
data.join(data.groupBy(N1).agg(count('*').alias('count1')), N1, 'left')\
    .join(data.groupBy(N2).agg(count('*').alias('count2')), N2, 'left')\
    .join(data.groupBy(N3).agg(count('*').alias('count3')), N3, 'left')\
    .join(data.groupBy(N4).agg(count('*').alias('count4')), N4, 'left')\
    .join(data.groupBy(N5).agg(count('*').alias('count5')), N5, 'left')\
    .select('MachineIdentifier', 'count1', 'count2', 'count3', 'count4', 'count5')\
    .write.csv(write_path, sep=',', mode="overwrite", header=True)
```

Competition details

Train & Test

- `train.csv` generation
- `test.csv` generation

```
df_gby_cat = spark.read.csv('data/df_groupby_cat_0/*.csv', header=True).repartition(80, ['MachineIdentifier'])
df_gby_cat.persist()
df_gby_cat.count()

df_gby_num = spark.read.csv('data/df_groupby_num_0/*.csv', header=True).repartition(80, ['MachineIdentifier'])
df_gby_num.persist()
df_gby_num.count()

full_df = df_num.join(df_cat, ['MachineIdentifier'])\
    .join(df_dates, ['MachineIdentifier'])\
    .join(df_kmeans, ['MachineIdentifier'])\
    .join(df_avsigver, ['MachineIdentifier'])\
    .join(df_gby_cat, ['MachineIdentifier'])\
    .join(df_gby_num, ['MachineIdentifier']).repartition(80, ['MachineIdentifier'])

full_df.persist()
full_df.count()

train = full_df.filter(col('HasDetections').isNotNull()).fillna('-1')

test = full_df.filter(col('HasDetections').isNull()).fillna('-1')
```

Competition details

```
print('Comienza entrenamiento del modelo LightGBM')
lgb_model = lgb.LGBMClassifier(**params)

ft_importances = np.zeros(X_train.shape[1])

counter = 1
for train_index, test_index in kf.split(train_ids, y_train):
    print('Fold {}\n'.format(counter))

    X_fit, X_val = X_train.iloc[train_index, :], X_train.iloc[test_index, :]
    y_fit, y_val = y_train.iloc[train_index], y_train.iloc[test_index]

    lgb_model.fit(X_fit,
                  y_fit,
                  eval_set=[(X_val, y_val)],
                  verbose=100,
                  early_stopping_rounds=100)

    del X_fit
    del X_val
    del y_fit
    del y_val
    del train_index
    del test_index
    gc.collect()
```

Competition details

Predictions

- LightGBM
- Ensemble
- Submit to Kaggle

```
X_test = test.loc[:, sel_cols]
X_machines = test.loc[:, 'MachineIdentifier']
del test
del list_
gc.collect()

lgb_test_preds = np.zeros(X_test.shape[0])

for i in range(1, k+1):
    model = joblib.load('saved_models/{}_{}.pkl'.format(model_name, i))
    print('Realizando predicciones. FOLD = {}'.format(i))
    lgb_test_preds += model.predict_proba(X_test)[:, 1]

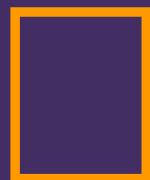
    del model
    gc.collect()

del X_test
gc.collect()

print('Haciendo la media y guardando CSV')
final_prds = lgb_test_preds/k

df_prds = pd.DataFrame({'MachineIdentifier': X_machines, 'HasDetections': final_prds})

df_prds.to_csv('submissions/{}.csv'.format(model_name), index=None)
```



Thanks!

Any questions?