*Lambda School Data Science — Tree Ensembles*

# Random Forests

## Pre-read

- Scikit-Learn User Guide, Ensemble Methods (https://scikit-learn.org/stable/modules/ensemble.html)
- Coloring with Random Forests (http://structuringtheunstructured.blogspot.com/2017/11/coloring-with-random-forests.html)
- Beware Default Random Forest Importances (https://explained.ai/rf-importance/index.html)

## More

- Machine Learning Explainability: Permutation Importance (https://www.kaggle.com/dansbecker/permutation-importance)
- eli5: Permutation Importance (https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html)
- eli5: Explaining XGBoost predictions on the Titanic dataset (https://eli5.readthedocs.io/en/latest/_notebooks/xgboost-titanic.html)
- The Mechanics of Machine Learning: Categorically Speaking (https://mlbook.explained.ai/catvars.html)

Selecting good features – Part III: random forests (https://blog.datadive.net/selecting-good-features-part-iii-random-forests/)

> There are a few things to keep in mind when using the impurity based ranking. Firstly, feature selection based on impurity reduction is biased towards preferring variables with more categories.
>
> Secondly, when the dataset has two (or more) correlated features, then from the point of view of the model, any of these correlated features can be used as the predictor, with no concrete preference of one over the others. But once one of them is used, the importance of others is significantly reduced since effectively the impurity they can remove is already removed by the first feature. As a consequence, they will have a lower reported importance. This is not an issue when we want to use feature selection to reduce overfitting, since it makes sense to remove features that are mostly duplicated by other features. But when interpreting the data, it can lead to the incorrect conclusion that one of the variables is a strong predictor while the others in the same group are unimportant, while actually they are very close in terms of their relationship with the response variable.

An Introduction to Statistical Learning (http://www-bcf.usc.edu/~gareth/ISL/), Chapter 8.2.1, Out-of-Bag Error Estimation

> It turns out that **there is a very straightforward way to estimate the test error of a bagged model, without the need to perform cross-validation or the validation set approach.**
>
> Recall that the key to bagging is that trees are repeatedly fit to bootstrapped subsets of the observations. One can show that on average, each bagged tree makes use of around two-thirds of the observations. The remaining one-third of the **observations not used to fit a given bagged tree are referred to as the out-of bag (OOB) observations.**
>
> We can predict the response for the ith observation using each of the trees in which that observation was OOB. This will yield around B/3 predictions for the ith observation. In order to obtain a single prediction for the ith observation, we can average these predicted responses (if regression is the goal) or can take a majority vote (if classification is the goal).
>
> This leads to a single OOB prediction for the ith observation. An OOB prediction can be obtained in this way for each of the n observations, from which the overall OOB MSE (for a regression problem) or classification error (for a classification problem) can be computed. The resulting **OOB error is a valid estimate of the test error for the bagged model, since the response for each observation is predicted using only the trees that were not fit using that observation.** ...
>
> It can be shown that with B sufficiently large, OOB error is virtually equivalent to leave-one-out cross-validation error. The OOB approach for estimating the test error is particularly **convenient when performing bagging on large data sets for which cross-validation would be computationally onerous.**

# Libraries

- eli5 (https://github.com/TeamHG-Memex/eli5): `conda install -c conda-forge eli5 / pip install eli5`
- category_encoders (https://github.com/scikit-learn-contrib/categorical-encoding): `conda install -c conda-forge category_encoders / pip install category_encoders`
- mlxtend (https://github.com/rasbt/mlxtend): `pip install mlxtend`
- ipywidgets (https://ipywidgets.readthedocs.io/en/stable/examples/Using%20Interact.html): included with Anaconda, doesn't work on Google Colab

# ipywidgets revisited: Decision Tree vs Random Forest

In [1]:

```
!pip install pip -U
!pip install numpy -U
!pip install pandas -U
!pip install scikit -U
!pip install sklearn -U

!pip install eli5
!pip install category_encoders
!pip install mlxtend
!pip install ipywidgets
```

```
Collecting pip
  Downloading https://files.pythonhosted.org/packages/62/ca/94d32a65
16ed197a491d17d46595ce58a83cbb2fca280414e57cd86b84dc/pip-19.2.1-py2.
py3-none-any.whl (1.4MB)
    100% |████████████████████████████████| 1.4MB 15.9MB/s ta 0:00:0
1
Installing collected packages: pip
  Found existing installation: pip 10.0.1
    Uninstalling pip-10.0.1:
      Successfully uninstalled pip-10.0.1
Successfully installed pip-19.2.1
Collecting numpy
  Downloading https://files.pythonhosted.org/packages/19/b9/bda9781f
0a74b90ebd2e046fde1196182900bd4a8e1ea503d3ffebc50e7c/numpy-1.17.0-cp
36-cp36m-manylinux1_x86_64.whl (20.4MB)
     |████████████████████████████████| 20.4MB 3.0MB/s eta 0:00:01
Installing collected packages: numpy
  Found existing installation: numpy 1.14.3
    Uninstalling numpy-1.14.3:
      Successfully uninstalled numpy-1.14.3
Successfully installed numpy-1.17.0
Collecting pandas
  Downloading https://files.pythonhosted.org/packages/1d/9a/7eb9952f
4b4d73fbd75ad1d5d6112f407e695957444cb695cbb3cdab918a/pandas-0.25.0-c
p36-cp36m-manylinux1_x86_64.whl (10.5MB)
     |████████████████████████████████| 10.5MB 3.0MB/s eta 0:00:01
Requirement already satisfied, skipping upgrade: python-dateutil>=2.
6.1 in /home/ec2-user/anaconda3/envs/python3/lib/python3.6/site-pack
ages (from pandas) (2.7.3)
Requirement already satisfied, skipping upgrade: numpy>=1.13.3 in /h
ome/ec2-user/anaconda3/envs/python3/lib/python3.6/site-packages (fro
m pandas) (1.17.0)
Requirement already satisfied, skipping upgrade: pytz>=2017.2 in /ho
```

```
me/ec2-user/anaconda3/envs/python3/lib/python3.6/site-packages (from
pandas) (2018.4)
Requirement already satisfied, skipping upgrade: six>=1.5 in /home/e
c2-user/anaconda3/envs/python3/lib/python3.6/site-packages (from pyt
hon-dateutil>=2.6.1->pandas) (1.11.0)
Installing collected packages: pandas
  Found existing installation: pandas 0.24.2
    Uninstalling pandas-0.24.2:
      Successfully uninstalled pandas-0.24.2
Successfully installed pandas-0.25.0
Collecting scikit
  ERROR: Could not find a version that satisfies the requirement sci
kit (from versions: none)
ERROR: No matching distribution found for scikit
Collecting sklearn
  Downloading https://files.pythonhosted.org/packages/1e/7a/dbb3be0c
e9bd5c8b7e3d87328e79063f8b263b2b1bfa4774cb1147bfcd3f/sklearn-0.0.tar
.gz
Requirement already satisfied, skipping upgrade: scikit-learn in /ho
me/ec2-user/anaconda3/envs/python3/lib/python3.6/site-packages (from
sklearn) (0.20.3)
Requirement already satisfied, skipping upgrade: scipy>=0.13.3 in /h
ome/ec2-user/anaconda3/envs/python3/lib/python3.6/site-packages (fro
m scikit-learn->sklearn) (1.1.0)
Requirement already satisfied, skipping upgrade: numpy>=1.8.2 in /ho
me/ec2-user/anaconda3/envs/python3/lib/python3.6/site-packages (from
scikit-learn->sklearn) (1.17.0)
Building wheels for collected packages: sklearn
  Building wheel for sklearn (setup.py) ... done
  Created wheel for sklearn: filename=sklearn-0.0-py2.py3-none-any.w
hl size=1321 sha256=a9031a8eb83a3a443461b8d100ac2208daf744c7a84f9db5
5e42fddb6b3441d0
  Stored in directory: /home/ec2-user/.cache/pip/wheels/76/03/bb/589
d421d27431bcd2c6da284d5f2286c8e3b2ea3cf1594c074
Successfully built sklearn
Installing collected packages: sklearn
Successfully installed sklearn-0.0
Collecting eli5
  Downloading https://files.pythonhosted.org/packages/73/cb/e773ec38
d6d8b48b18537020e7782ac038581b52047d41b6500135e4bdc7/eli5-0.9.0-py2.
py3-none-any.whl (94kB)
     |████████████████████████████████| 102kB 4.3MB/s ta 0:00:011
Collecting tabulate>=0.7.7 (from eli5)
  Downloading https://files.pythonhosted.org/packages/c2/fd/202954b3
f0eb896c53b7b6f07390851b1fd2ca84aa95880d7ae4f434c4ac/tabulate-0.8.3.
tar.gz (46kB)
     |████████████████████████████████| 51kB 32.8MB/s eta 0:00:01
Requirement already satisfied: six in /home/ec2-user/anaconda3/envs/
python3/lib/python3.6/site-packages (from eli5) (1.11.0)
Collecting graphviz (from eli5)
```

```
  Downloading https://files.pythonhosted.org/packages/5c/b1/016e6575
86843f40b4daa66127ce1ee9e3285ff15baf5d80946644a98aeb/graphviz-0.11.1
-py2.py3-none-any.whl
Requirement already satisfied: typing in /home/ec2-user/anaconda3/en
vs/python3/lib/python3.6/site-packages (from eli5) (3.6.4)
Requirement already satisfied: scikit-learn>=0.18 in /home/ec2-user/
anaconda3/envs/python3/lib/python3.6/site-packages (from eli5) (0.20
.3)
Requirement already satisfied: jinja2 in /home/ec2-user/anaconda3/en
vs/python3/lib/python3.6/site-packages (from eli5) (2.10)
Requirement already satisfied: scipy in /home/ec2-user/anaconda3/env
s/python3/lib/python3.6/site-packages (from eli5) (1.1.0)
Requirement already satisfied: numpy>=1.9.0 in /home/ec2-user/anacon
da3/envs/python3/lib/python3.6/site-packages (from eli5) (1.17.0)
Requirement already satisfied: attrs>16.0.0 in /home/ec2-user/anacon
da3/envs/python3/lib/python3.6/site-packages (from eli5) (18.1.0)
Requirement already satisfied: MarkupSafe>=0.23 in /home/ec2-user/an
aconda3/envs/python3/lib/python3.6/site-packages (from jinja2->eli5)
(1.0)
Building wheels for collected packages: tabulate
  Building wheel for tabulate (setup.py) ... done
  Created wheel for tabulate: filename=tabulate-0.8.3-cp36-none-any.
whl size=22563 sha256=1e2ee7a8d4bcb8a7799467ea09d9c946758d85ae5c216e
82c943a60a1cd48bf4
  Stored in directory: /home/ec2-user/.cache/pip/wheels/2b/67/89/414
471314a2d15de625d184d8be6d38a03ae1e983dbda91e84
Successfully built tabulate
Installing collected packages: tabulate, graphviz, eli5
Successfully installed eli5-0.9.0 graphviz-0.11.1 tabulate-0.8.3
Collecting category_encoders
  Downloading https://files.pythonhosted.org/packages/6e/a1/f7a22f14
4f33be78afeb06bfa78478e8284a64263a3c09b1ef54e673841e/category_encode
rs-2.0.0-py2.py3-none-any.whl (87kB)
     |████████████████████████████████| 92kB 3.9MB/s eta 0:00:011
Requirement already satisfied: scipy>=0.19.0 in /home/ec2-user/anaco
nda3/envs/python3/lib/python3.6/site-packages (from category_encoder
s) (1.1.0)
Requirement already satisfied: pandas>=0.21.1 in /home/ec2-user/anac
onda3/envs/python3/lib/python3.6/site-packages (from category_encode
rs) (0.25.0)
Requirement already satisfied: patsy>=0.4.1 in /home/ec2-user/anacon
da3/envs/python3/lib/python3.6/site-packages (from category_encoders
) (0.5.0)
Requirement already satisfied: scikit-learn>=0.20.0 in /home/ec2-use
r/anaconda3/envs/python3/lib/python3.6/site-packages (from category_
encoders) (0.20.3)
Requirement already satisfied: statsmodels>=0.6.1 in /home/ec2-user/
anaconda3/envs/python3/lib/python3.6/site-packages (from category_en
coders) (0.9.0)
Requirement already satisfied: numpy>=1.11.3 in /home/ec2-user/anaco
```

```
nda3/envs/python3/lib/python3.6/site-packages (from category_encoder
s) (1.17.0)
Requirement already satisfied: pytz>=2017.2 in /home/ec2-user/anacon
da3/envs/python3/lib/python3.6/site-packages (from pandas>=0.21.1->c
ategory_encoders) (2018.4)
Requirement already satisfied: python-dateutil>=2.6.1 in /home/ec2-u
ser/anaconda3/envs/python3/lib/python3.6/site-packages (from pandas>
=0.21.1->category_encoders) (2.7.3)
Requirement already satisfied: six in /home/ec2-user/anaconda3/envs/
python3/lib/python3.6/site-packages (from patsy>=0.4.1->category_enc
oders) (1.11.0)
Installing collected packages: category-encoders
Successfully installed category-encoders-2.0.0
Collecting mlxtend
  Downloading https://files.pythonhosted.org/packages/52/04/c362f34f
666f0ddc7cf593805e64d64fa670ed96fd9302e68549dd48287d/mlxtend-0.17.0-
py2.py3-none-any.whl (1.3MB)
     |████████████████████████████████| 1.3MB 3.1MB/s eta 0:00:01
Requirement already satisfied: matplotlib>=3.0.0 in /home/ec2-user/a
naconda3/envs/python3/lib/python3.6/site-packages (from mlxtend) (3.
0.3)
Requirement already satisfied: scikit-learn>=0.20.3 in /home/ec2-use
r/anaconda3/envs/python3/lib/python3.6/site-packages (from mlxtend)
(0.20.3)
Collecting joblib>=0.13.2 (from mlxtend)
  Downloading https://files.pythonhosted.org/packages/cd/c1/50a758e8
247561e58cb87305b1e90b171b8c767b15b12a1734001f41d356/joblib-0.13.2-p
y2.py3-none-any.whl (278kB)
     |████████████████████████████████| 286kB 25.7MB/s eta 0:00:01
Requirement already satisfied: pandas>=0.24.2 in /home/ec2-user/anac
onda3/envs/python3/lib/python3.6/site-packages (from mlxtend) (0.25.
0)
Requirement already satisfied: setuptools in /home/ec2-user/anaconda
3/envs/python3/lib/python3.6/site-packages (from mlxtend) (39.1.0)
Collecting scipy>=1.2.1 (from mlxtend)
  Using cached https://files.pythonhosted.org/packages/29/50/a552a5a
ff252ae915f522e44642bb49a7b7b31677f9580cfd11bcc869976/scipy-1.3.1-cp
36-cp36m-manylinux1_x86_64.whl
Requirement already satisfied: numpy>=1.16.2 in /home/ec2-user/anaco
nda3/envs/python3/lib/python3.6/site-packages (from mlxtend) (1.17.0
)
Requirement already satisfied: kiwisolver>=1.0.1 in /home/ec2-user/a
naconda3/envs/python3/lib/python3.6/site-packages (from matplotlib>=
3.0.0->mlxtend) (1.0.1)
Requirement already satisfied: cycler>=0.10 in /home/ec2-user/anacon
da3/envs/python3/lib/python3.6/site-packages (from matplotlib>=3.0.0
->mlxtend) (0.10.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.
0.1 in /home/ec2-user/anaconda3/envs/python3/lib/python3.6/site-pack
ages (from matplotlib>=3.0.0->mlxtend) (2.2.0)
```

```
Requirement already satisfied: python-dateutil>=2.1 in /home/ec2-use
r/anaconda3/envs/python3/lib/python3.6/site-packages (from matplotli
b>=3.0.0->mlxtend) (2.7.3)
Requirement already satisfied: pytz>=2017.2 in /home/ec2-user/anacon
da3/envs/python3/lib/python3.6/site-packages (from pandas>=0.24.2->m
lxtend) (2018.4)
Requirement already satisfied: six in /home/ec2-user/anaconda3/envs/
python3/lib/python3.6/site-packages (from cycler>=0.10->matplotlib>=
3.0.0->mlxtend) (1.11.0)
Installing collected packages: joblib, scipy, mlxtend
  Found existing installation: scipy 1.1.0
    Uninstalling scipy-1.1.0:
      Successfully uninstalled scipy-1.1.0
Successfully installed joblib-0.13.2 mlxtend-0.17.0 scipy-1.3.1
Requirement already satisfied: ipywidgets in /home/ec2-user/anaconda
3/envs/python3/lib/python3.6/site-packages (7.4.0)
Requirement already satisfied: traitlets>=4.3.1 in /home/ec2-user/an
aconda3/envs/python3/lib/python3.6/site-packages (from ipywidgets) (
4.3.2)
Requirement already satisfied: ipykernel>=4.5.1 in /home/ec2-user/an
aconda3/envs/python3/lib/python3.6/site-packages (from ipywidgets) (
4.8.2)
Requirement already satisfied: ipython>=4.0.0; python_version >= "3.
3" in /home/ec2-user/anaconda3/envs/python3/lib/python3.6/site-packa
ges (from ipywidgets) (6.4.0)
Requirement already satisfied: nbformat>=4.2.0 in /home/ec2-user/ana
conda3/envs/python3/lib/python3.6/site-packages (from ipywidgets) (4
.4.0)
Requirement already satisfied: widgetsnbextension~=3.4.0 in /home/ec
2-user/anaconda3/envs/python3/lib/python3.6/site-packages (from ipyw
idgets) (3.4.2)
Requirement already satisfied: ipython_genutils in /home/ec2-user/an
aconda3/envs/python3/lib/python3.6/site-packages (from traitlets>=4.
3.1->ipywidgets) (0.2.0)
Requirement already satisfied: six in /home/ec2-user/anaconda3/envs/
python3/lib/python3.6/site-packages (from traitlets>=4.3.1->ipywidge
ts) (1.11.0)
Requirement already satisfied: decorator in /home/ec2-user/anaconda3
/envs/python3/lib/python3.6/site-packages (from traitlets>=4.3.1->ip
ywidgets) (4.3.0)
Requirement already satisfied: jupyter_client in /home/ec2-user/anac
onda3/envs/python3/lib/python3.6/site-packages (from ipykernel>=4.5.
1->ipywidgets) (5.2.3)
Requirement already satisfied: tornado>=4.0 in /home/ec2-user/anacon
da3/envs/python3/lib/python3.6/site-packages (from ipykernel>=4.5.1-
>ipywidgets) (5.0.2)
Requirement already satisfied: simplegeneric>0.8 in /home/ec2-user/a
naconda3/envs/python3/lib/python3.6/site-packages (from ipython>=4.0
.0; python_version >= "3.3"->ipywidgets) (0.8.1)
Requirement already satisfied: pexpect; sys_platform != "win32" in /
```

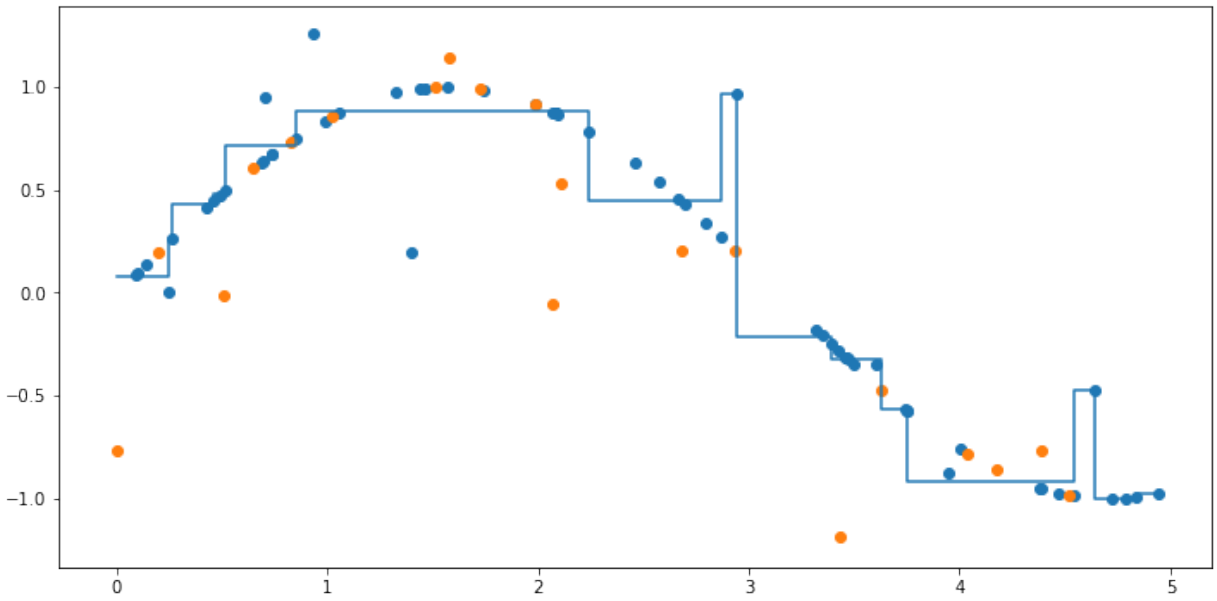
```
home/ec2-user/anaconda3/envs/python3/lib/python3.6/site-packages (fr
om ipython>=4.0.0; python_version >= "3.3"->ipywidgets) (4.5.0)
Requirement already satisfied: backcall in /home/ec2-user/anaconda3/
envs/python3/lib/python3.6/site-packages (from ipython>=4.0.0; pytho
n_version >= "3.3"->ipywidgets) (0.1.0)
Requirement already satisfied: prompt-toolkit<2.0.0,>=1.0.15 in /hom
e/ec2-user/anaconda3/envs/python3/lib/python3.6/site-packages (from
ipython>=4.0.0; python_version >= "3.3"->ipywidgets) (1.0.15)
Requirement already satisfied: pickleshare in /home/ec2-user/anacond
a3/envs/python3/lib/python3.6/site-packages (from ipython>=4.0.0; py
thon_version >= "3.3"->ipywidgets) (0.7.4)
Requirement already satisfied: jedi>=0.10 in /home/ec2-user/anaconda
3/envs/python3/lib/python3.6/site-packages (from ipython>=4.0.0; pyt
hon_version >= "3.3"->ipywidgets) (0.12.0)
Requirement already satisfied: setuptools>=18.5 in /home/ec2-user/an
aconda3/envs/python3/lib/python3.6/site-packages (from ipython>=4.0.
0; python_version >= "3.3"->ipywidgets) (39.1.0)
Requirement already satisfied: pygments in /home/ec2-user/anaconda3/
envs/python3/lib/python3.6/site-packages (from ipython>=4.0.0; pytho
n_version >= "3.3"->ipywidgets) (2.2.0)
Requirement already satisfied: jsonschema!=2.5.0,>=2.4 in /home/ec2-
user/anaconda3/envs/python3/lib/python3.6/site-packages (from nbform
at>=4.2.0->ipywidgets) (2.6.0)
Requirement already satisfied: jupyter_core in /home/ec2-user/anacon
da3/envs/python3/lib/python3.6/site-packages (from nbformat>=4.2.0->
ipywidgets) (4.4.0)
Requirement already satisfied: notebook>=4.4.1 in /home/ec2-user/ana
conda3/envs/python3/lib/python3.6/site-packages (from widgetsnbexten
sion~=3.4.0->ipywidgets) (5.5.0)
Requirement already satisfied: pyzmq>=13 in /home/ec2-user/anaconda3
/envs/python3/lib/python3.6/site-packages (from jupyter_client->ipyk
ernel>=4.5.1->ipywidgets) (17.0.0)
Requirement already satisfied: python-dateutil>=2.1 in /home/ec2-use
r/anaconda3/envs/python3/lib/python3.6/site-packages (from jupyter_c
lient->ipykernel>=4.5.1->ipywidgets) (2.7.3)
Requirement already satisfied: ptyprocess>=0.5 in /home/ec2-user/ana
conda3/envs/python3/lib/python3.6/site-packages (from pexpect; sys_p
latform != "win32"->ipython>=4.0.0; python_version >= "3.3"->ipywidg
ets) (0.5.2)
Requirement already satisfied: wcwidth in /home/ec2-user/anaconda3/e
nvs/python3/lib/python3.6/site-packages (from prompt-toolkit<2.0.0,>
=1.0.15->ipython>=4.0.0; python_version >= "3.3"->ipywidgets) (0.1.7
)
Requirement already satisfied: parso>=0.2.0 in /home/ec2-user/anacon
da3/envs/python3/lib/python3.6/site-packages (from jedi>=0.10->ipyth
on>=4.0.0; python_version >= "3.3"->ipywidgets) (0.2.0)
Requirement already satisfied: jinja2 in /home/ec2-user/anaconda3/en
vs/python3/lib/python3.6/site-packages (from notebook>=4.4.1->widget
snbextension~=3.4.0->ipywidgets) (2.10)
Requirement already satisfied: terminado>=0.8.1 in /home/ec2-user/an
```

```
aconda3/envs/python3/lib/python3.6/site-packages (from notebook>=4.4
.1->widgetsnbextension~=3.4.0->ipywidgets) (0.8.1)
Requirement already satisfied: nbconvert in /home/ec2-user/anaconda3
/envs/python3/lib/python3.6/site-packages (from notebook>=4.4.1->wid
getsnbextension~=3.4.0->ipywidgets) (5.4.1)
Requirement already satisfied: Send2Trash in /home/ec2-user/anaconda
3/envs/python3/lib/python3.6/site-packages (from notebook>=4.4.1->wi
dgetsnbextension~=3.4.0->ipywidgets) (1.5.0)
Requirement already satisfied: MarkupSafe>=0.23 in /home/ec2-user/an
aconda3/envs/python3/lib/python3.6/site-packages (from jinja2->noteb
ook>=4.4.1->widgetsnbextension~=3.4.0->ipywidgets) (1.0)
Requirement already satisfied: mistune>=0.8.1 in /home/ec2-user/anac
onda3/envs/python3/lib/python3.6/site-packages (from nbconvert->note
book>=4.4.1->widgetsnbextension~=3.4.0->ipywidgets) (0.8.3)
Requirement already satisfied: entrypoints>=0.2.2 in /home/ec2-user/
anaconda3/envs/python3/lib/python3.6/site-packages (from nbconvert->
notebook>=4.4.1->widgetsnbextension~=3.4.0->ipywidgets) (0.2.3)
Requirement already satisfied: bleach in /home/ec2-user/anaconda3/en
vs/python3/lib/python3.6/site-packages (from nbconvert->notebook>=4.
4.1->widgetsnbextension~=3.4.0->ipywidgets) (2.1.3)
Requirement already satisfied: pandocfilters>=1.4.1 in /home/ec2-use
r/anaconda3/envs/python3/lib/python3.6/site-packages (from nbconvert
->notebook>=4.4.1->widgetsnbextension~=3.4.0->ipywidgets) (1.4.2)
Requirement already satisfied: testpath in /home/ec2-user/anaconda3/
envs/python3/lib/python3.6/site-packages (from nbconvert->notebook>=
4.4.1->widgetsnbextension~=3.4.0->ipywidgets) (0.3.1)
Requirement already satisfied: defusedxml in /home/ec2-user/anaconda
3/envs/python3/lib/python3.6/site-packages (from nbconvert->notebook
>=4.4.1->widgetsnbextension~=3.4.0->ipywidgets) (0.6.0)
Requirement already satisfied: html5lib!=1.0b1,!=1.0b2,!=1.0b3,!=1.0
b4,!=1.0b5,!=1.0b6,!=1.0b7,!=1.0b8,>=0.99999999pre in /home/ec2-user
/anaconda3/envs/python3/lib/python3.6/site-packages (from bleach->nb
convert->notebook>=4.4.1->widgetsnbextension~=3.4.0->ipywidgets) (1.
0.1)
Requirement already satisfied: webencodings in /home/ec2-user/anacon
da3/envs/python3/lib/python3.6/site-packages (from html5lib!=1.0b1,!
=1.0b2,!=1.0b3,!=1.0b4,!=1.0b5,!=1.0b6,!=1.0b7,!=1.0b8,>=0.99999999p
re->bleach->nbconvert->notebook>=4.4.1->widgetsnbextension~=3.4.0->i
pywidgets) (0.5.1)
```

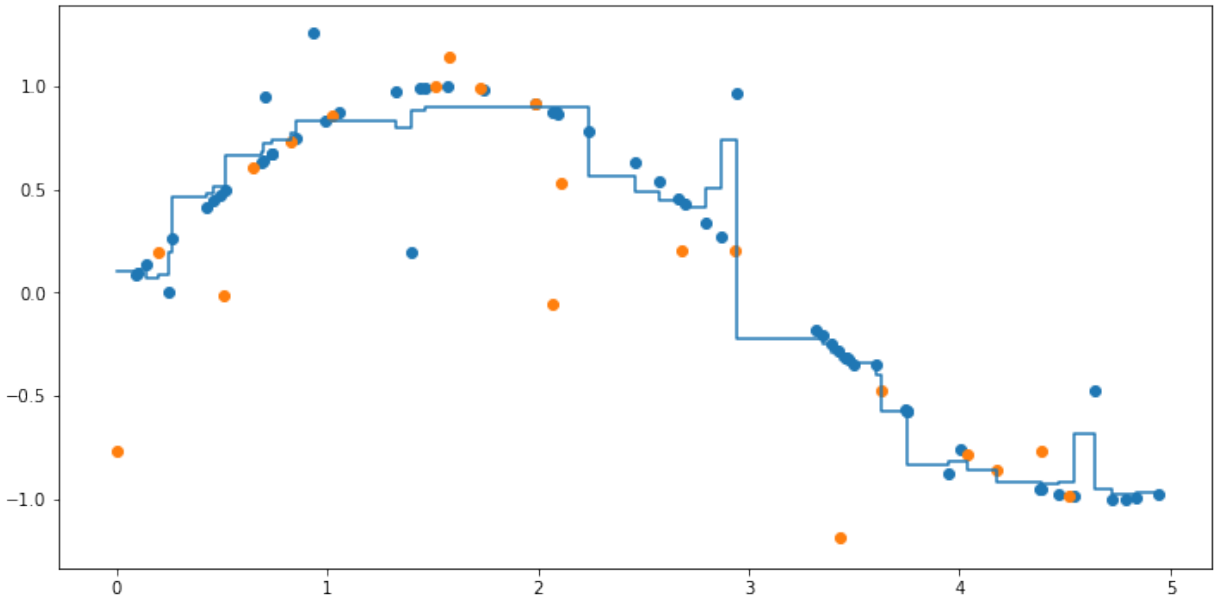
In [2]:

```
%matplotlib inline
from ipywidgets import interact
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor

# Example from http://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html
def make_data():
    import numpy as np
    rng = np.random.RandomState(1)
    X = np.sort(5 * rng.rand(80, 1), axis=0)
    y = np.sin(X).ravel()
    y[::5] += 2 * (0.5 - rng.rand(16))
    return X, y

X, y = make_data()

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.25, random_state=42)

def regress_wave(max_depth):
    dt = DecisionTreeRegressor(max_depth=max_depth)
    dt.fit(X_train, y_train)
    print('Decision Tree train R^2:', dt.score(X_train, y_train))
    print('Decision Tree test R^2:', dt.score(X_test, y_test))
    plt.gcf().set_size_inches(12, 6)
    plt.scatter(X_train, y_train)
    plt.scatter(X_test, y_test)
    plt.step(X, dt.predict(X))
    plt.show()

    rf = RandomForestRegressor(max_depth=max_depth, n_estimators=100, n_jobs=-1)
    rf.fit(X_train, y_train)
    print('Random Forest train R^2:', rf.score(X_train, y_train))
    print('Random Forest test R^2:', rf.score(X_test, y_test))
    plt.gcf().set_size_inches(12, 6)
    plt.scatter(X_train, y_train)
    plt.scatter(X_test, y_test)
    plt.step(X, rf.predict(X))
    plt.show()

interact(regress_wave, max_depth=(1,8,1));
```

```
Decision Tree train R^2: 0.9681759735712112
Decision Tree test R^2: 0.683265008917209
```




```
Random Forest train R^2: 0.9682443149423664
Random Forest test R^2: 0.702160140479906
```

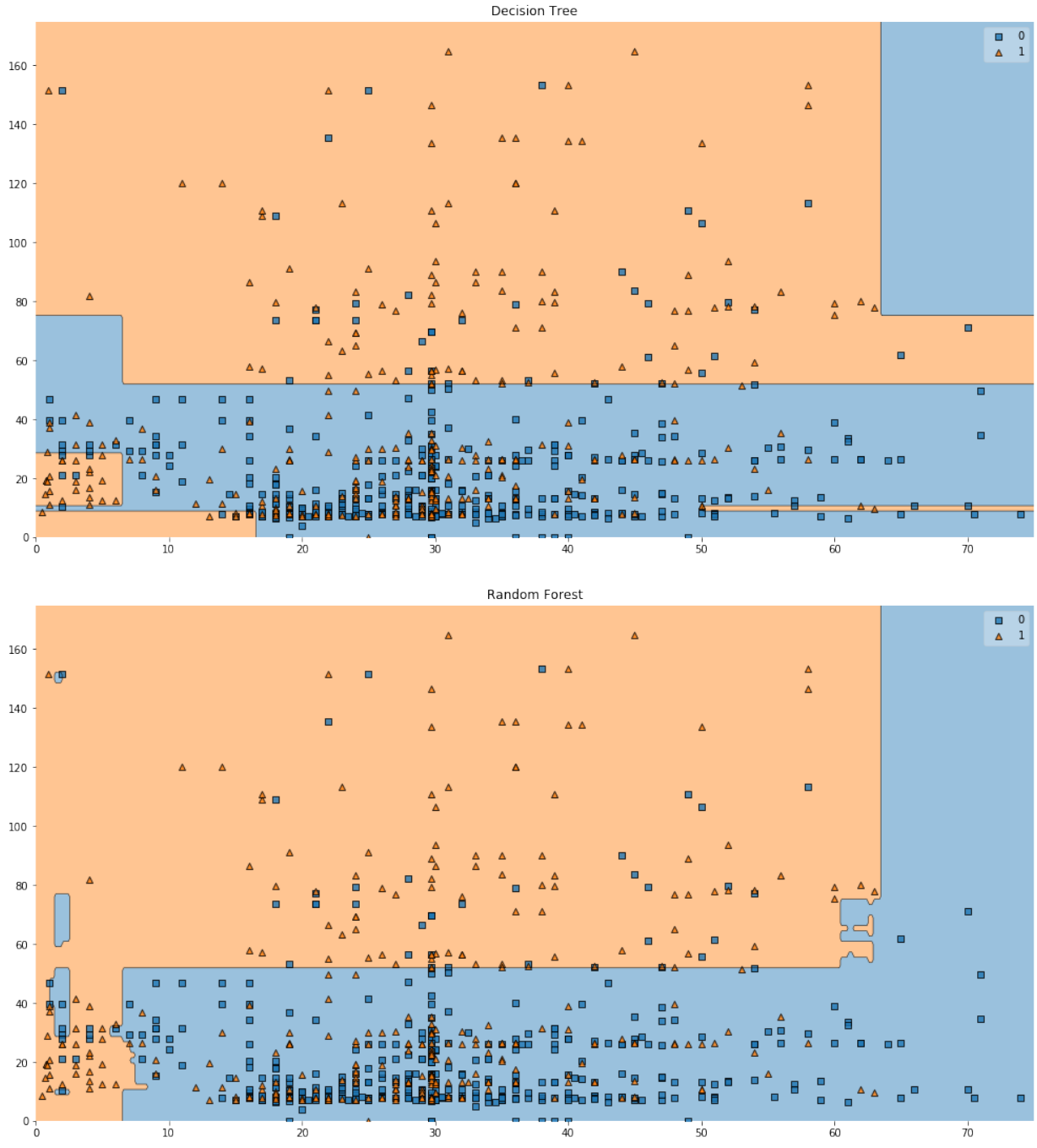



## Regressing a wave

## Titanic survival, by Age & Fare

In [3]:

```
from mlxtend.plotting import plot_decision_regions
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier
from sklearn.impute import SimpleImputer
from sklearn.tree import DecisionTreeClassifier

titanic = sns.load_dataset('titanic')
X = SimpleImputer().fit_transform(titanic[['age', 'fare']])
y = titanic['survived'].values

def classify_titanic(max_depth):
    dt = DecisionTreeClassifier(max_depth=max_depth)
    dt.fit(X, y)
    plot_decision_regions(X, y, dt)
    plt.gcf().set_size_inches(17, 9)
    plt.title('Decision Tree')
    plt.axis((0,75,0,175))
    plt.show()

    rf = RandomForestClassifier(max_depth=max_depth, n_estimators=100, n_jobs=-1
)
    rf.fit(X, y)
    plot_decision_regions(X, y, rf)
    plt.gcf().set_size_inches(17, 9)
    plt.title('Random Forest')
    plt.axis((0,75,0,175))
    plt.show()

interact(classify_titanic, max_depth=(1,8,1));
```







# Lending Club

Read csv files downloaded from [Kaggle (https://www.kaggle.com/c/ds2-tree-ensembles/data)]

In [4]:

```
import zipfile
import pandas as pd
ff = ["test_features.csv", "train_features.csv", "train_labels.csv", "sample_sub
mission.csv"]
with zipfile.ZipFile("data/ds2-tree-ensembles.zip") as z:
   with z.open(ff[0]) as f:
      test = pd.read_csv(f, header=0, delimiter=",")
      print(ff[0])
      print(test.sample())    # print the first 5 rows
   with z.open(ff[1]) as f:
      train = pd.read_csv(f, header=0, delimiter=",")
      print(ff[1])
      print(train.sample())    # print the first 5 rows
   with z.open(ff[2]) as f:
      train_labels = pd.read_csv(f, header=0, delimiter=",")
      print(ff[2])
      print(train_labels.sample())    # print the first 5 rows
   with z.open(ff[3]) as f:
      sample_submission = pd.read_csv(f, header=0, delimiter=",")
      print(ff[3])
      print(sample_submission.sample())    # print the first 5 rows
```

```
test_features.csv
            id  member_id  loan_amnt  funded_amnt        term int_ra
te  \
18428  1644268        NaN     6000.0       6000.0   36 months   15.6
1%

       installment grade sub_grade emp_title  ... sec_app_inq_last_6
mths  \
18428       209.79     D        D1        EA  ...
NaN

      sec_app_mort_acc  sec_app_open_acc  sec_app_revol_util  \
18428              NaN               NaN                 NaN

      sec_app_open_act_il sec_app_num_rev_accts  \
18428                 NaN                   NaN

      sec_app_chargeoff_within_12_mths sec_app_collections_12_mths_e
x_med  \
18428                              NaN
NaN

      sec_app_mths_since_last_major_derog  disbursement_method
18428                                 NaN                 Cash

[1 rows x 103 columns]
```

```
train_features.csv
             id  member_id  loan_amnt  funded_amnt        term int_r
ate  \
207155  1105790        NaN     6000.0       6000.0   36 months   23.
43%

        installment grade sub_grade emp_title  ... sec_app_inq_last_
6mths  \
207155       233.61     F        F1  Engineer  ...
NaN

       sec_app_mort_acc  sec_app_open_acc  sec_app_revol_util  \
207155              NaN               NaN                 NaN

       sec_app_open_act_il sec_app_num_rev_accts  \
207155                 NaN                   NaN

       sec_app_chargeoff_within_12_mths sec_app_collections_12_mths_
ex_med  \
207155                              NaN
NaN

       sec_app_mths_since_last_major_derog  disbursement_method
207155                                 NaN                 Cash

[1 rows x 103 columns]
train_labels.csv
             id  charged_off
666215  1259822            0
sample_submission.csv
          id  charged_off
2592  1854080          0.5
```

In [6]:

```
%%time
import pandas as pd
pd.options.display.max_columns = 200
pd.options.display.max_rows = 200

#X_train = pd.read_csv('/Users/wel51x/Downloads/ds2-tree-ensembles/train_features.csv')
#X_test = pd.read_csv('/Users/wel51x/Downloads/ds2-tree-ensembles/test_features.csv')
#y_train = pd.read_csv('/Users/wel51x/Downloads/ds2-tree-ensembles/train_labels.csv')['charged_off']
X_train = train
X_test = test
y_train = train_labels['charged_off']
#sample_submission = pd.read_csv('/Users/wel51x/Downloads/ds2-tree-ensembles/sample_submission.csv')

print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

```
(1309457, 103) (26724, 103) (1309457,) (20,)
CPU times: user 183 µs, sys: 107 µs, total: 290 µs
Wall time: 263 µs
```

Wrangle X_train and X_test in the same way

In [7]:

```
def wrangle(X):
    X = X.copy()

    # Drop some columns
    X = X.drop(columns='id')  # id is random
    X = X.drop(columns=['member_id', 'url', 'desc'])  # All null
    X = X.drop(columns='title')  # Duplicative of purpose
    X = X.drop(columns='grade')  # Duplicative of sub_grade

    # Transform sub_grade from "A1" - "G5" to 1.1 - 7.5
    def wrangle_sub_grade(x):
        first_digit = ord(x[0]) - 64
        second_digit = int(x[1])
        return first_digit + second_digit/10

    X['sub_grade'] = X['sub_grade'].apply(wrangle_sub_grade)

    # Convert percentages from strings to floats
    X['int_rate'] = X['int_rate'].str.strip('%').astype(float)
    X['revol_util'] = X['revol_util'].str.strip('%').astype(float)
```

```
    # Transform earliest_cr_line to an integer: how many days it's been open
    X['earliest_cr_line'] = pd.to_datetime(X['earliest_cr_line'], infer_datetime
_format=True)
    X['earliest_cr_line'] = pd.Timestamp.today() - X['earliest_cr_line']
    X['earliest_cr_line'] = X['earliest_cr_line'].dt.days

    # Create features for three employee titles: teacher, manager, owner
    X['emp_title'] = X['emp_title'].str.lower()
    X['emp_title_teacher'] = X['emp_title'].str.contains('teacher', na=False)
    X['emp_title_manager'] = X['emp_title'].str.contains('manager', na=False)
    X['emp_title_owner']   = X['emp_title'].str.contains('owner', na=False)

    # Drop categoricals with high cardinality
    X = X.drop(columns=['emp_title', 'zip_code'])

    # Transform features with many nulls to binary flags
    many_nulls = ['sec_app_mths_since_last_major_derog',
                  'sec_app_revol_util',
                  'sec_app_earliest_cr_line',
                  'sec_app_mort_acc',
                  'dti_joint',
                  'sec_app_collections_12_mths_ex_med',
                  'sec_app_chargeoff_within_12_mths',
                  'sec_app_num_rev_accts',
                  'sec_app_open_act_il',
                  'sec_app_open_acc',
                  'revol_bal_joint',
                  'annual_inc_joint',
                  'sec_app_inq_last_6mths',
                  'mths_since_last_record',
                  'mths_since_recent_bc_dlq',
                  'mths_since_last_major_derog',
                  'mths_since_recent_revol_delinq',
                  'mths_since_last_delinq',
                  'il_util',
                  'emp_length',
                  'mths_since_recent_inq',
                  'mo_sin_old_il_acct',
                  'mths_since_rcnt_il',
                  'num_tl_120dpd_2m',
                  'bc_util',
                  'percent_bc_gt_75',
                  'bc_open_to_buy',
                  'mths_since_recent_bc']

    for col in many_nulls:
        X[col] = X[col].isnull()

    # For features with few nulls, do mean imputation
```

```
    for col in X:
        if X[col].isnull().sum() > 0:
            X[col] = X[col].fillna(X[col].mean())

    # Return the wrangled dataframe
    return X


X_train = wrangle(X_train)
X_test  = wrangle(X_test)
X_train.shape, X_test.shape
```

Out[7]:

```
((1309457, 98), (26724, 98))
```

Now X_train (and X_test) have no nulls

In [8]:

```
null_counts = X_train.isnull().sum()
all(null_counts == 0)
```

Out[8]:

```
True
```

And no high cardinality categoricals

In [10]:

```
cardinality = X_train.select_dtypes(exclude='number').nunique()
#all(cardinality <= 50) False
all(cardinality <= 51)
```

Out[10]:

```
True
```

In [11]:

```
cardinality
```

Out[11]:

```
term                                  2
emp_length                            2
home_ownership                        6
purpose                              14
addr_state                           51
mths_since_last_delinq                2
mths_since_last_record                2
initial_list_status                   2
mths_since_last_major_derog           2
application_type                      2
annual_inc_joint                      2
dti_joint                             2
mths_since_rcnt_il                    2
il_util                               2
bc_open_to_buy                        2
bc_util                               2
mo_sin_old_il_acct                    2
mths_since_recent_bc                  2
mths_since_recent_bc_dlq              2
mths_since_recent_inq                 2
mths_since_recent_revol_delinq        2
num_tl_120dpd_2m                      2
percent_bc_gt_75                      2
revol_bal_joint                       2
sec_app_earliest_cr_line              2
sec_app_inq_last_6mths                2
sec_app_mort_acc                      2
sec_app_open_acc                      2
sec_app_revol_util                    2
sec_app_open_act_il                   2
sec_app_num_rev_accts                 2
sec_app_chargeoff_within_12_mths      2
sec_app_collections_12_mths_ex_med    2
sec_app_mths_since_last_major_derog   2
disbursement_method                   2
emp_title_teacher                     2
emp_title_manager                     2
emp_title_owner                       2
dtype: int64
```

# Decision Tree

In [12]:

```
%%time
import category_encoders as ce
from sklearn.model_selection import cross_val_score
from sklearn.pipeline import make_pipeline
from sklearn.tree import DecisionTreeClassifier

pipe = make_pipeline(
    ce.OrdinalEncoder(),
    DecisionTreeClassifier(max_depth=5, class_weight='balanced')
)

cross_val_score(pipe, X_train, y_train, cv=5, scoring='roc_auc')
```

```
CPU times: user 3min 4s, sys: 17.6 s, total: 3min 22s
Wall time: 3min 22s
```

In [13]:

```
%%time
from sklearn.ensemble import RandomForestClassifier

pipe = make_pipeline(
    ce.OrdinalEncoder(),
    RandomForestClassifier(
        n_estimators=100,
        class_weight='balanced',
        min_samples_leaf=0.005,
        oob_score=True,
        n_jobs=-1)
)

cross_val_score(pipe, X_train, y_train, cv=5, scoring='roc_auc', verbose=10)
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurr
ent workers.

[CV]  ................................................................
..
[CV] ....................... , score=0.7142322450064384, total= 1.1m
in
[CV]  ................................................................
..

[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:  1.1min remaini
ng:    0.0s

[CV] ....................... , score=0.7122351604705779, total= 1.0m
in
[CV]  ................................................................
..

[Parallel(n_jobs=1)]: Done   2 out of   2 | elapsed:  2.2min remaini
ng:    0.0s

[CV] ....................... , score=0.7132051103964225, total=  55.
3s
[CV]  ................................................................
..

[Parallel(n_jobs=1)]: Done   3 out of   3 | elapsed:  3.1min remaini
ng:    0.0s

[CV] ....................... , score=0.7158387078062994, total=  55.
2s
[CV]  ................................................................
..

[Parallel(n_jobs=1)]: Done   4 out of   4 | elapsed:  4.0min remaini
ng:    0.0s

[CV] ....................... , score=0.7146950484694643, total=  55.
0s
CPU times: user 3min 6s, sys: 1min 1s, total: 4min 7s
Wall time: 4min 55s

[Parallel(n_jobs=1)]: Done   5 out of   5 | elapsed:  4.9min remaini
ng:    0.0s
[Parallel(n_jobs=1)]: Done   5 out of   5 | elapsed:  4.9min finishe
d
```

# Out-of-Bag estimated score

Out-of-bag is a faster way to get an estimated score with Random Forest, using the parameter `oob_score=True`

# Random Forest

Improves ROC AUC compared to Decision Tree

In [14]:

```
from sklearn.metrics import roc_auc_score
```

In [15]:

```
%%time
pipe.fit(X_train, y_train)
y_pred_proba = pipe.named_steps['randomforestclassifier'].oob_decision_function_
[:, 1]
print('ROC AUC, Out-of-Bag estimate:', roc_auc_score(y_train, y_pred_proba))
```

```
ROC AUC, Out-of-Bag estimate: 0.7132605104073831
CPU times: user 15min 30s, sys: 18.9 s, total: 15min 49s
Wall time: 1min 18s
```

In [16]:

```
pipe.named_steps
```

Out[16]:

```
{'ordinalencoder': OrdinalEncoder(cols=['term', 'home_ownership', 'p
urpose', 'addr_state', 'initial_list_status', 'application_type', 'd
isbursement_method'],
         drop_invariant=False, handle_missing='value',
         handle_unknown='value',
         mapping=[{'col': 'term', 'mapping':  36 months    1
  60 months    2
 NaN          -2
 dtype: int64, 'data_type': dtype('O')}, {'col': 'home_ownership', '
mapping': MORTGAGE    1
 RENT        2
 OWN         3
 ANY         4
 OTHER       5
 NONE        6
 NaN        -2
 dtype: int64, 'data_type': dtype('O')}, {...od', 'mapping': Cash
1
 DirectPay    2
 NaN         -2
 dtype: int64, 'data_type': dtype('O')}],
         return_df=True, verbose=0),
 'randomforestclassifier': RandomForestClassifier(bootstrap=True, cl
ass_weight='balanced',
             criterion='gini', max_depth=None, max_features='auto',
             max_leaf_nodes=None, min_impurity_decrease=0.0,
             min_impurity_split=None, min_samples_leaf=0.005,
             min_samples_split=2, min_weight_fraction_leaf=0.0,
             n_estimators=100, n_jobs=-1, oob_score=True, random_sta
te=None,
             verbose=0, warm_start=False)}
```

## You can explore hyperparameter values

In [18]:

```
%%time

max_depths = list(range(2, 17, 2)) + [None]

for max_depth in max_depths:

    pipe = make_pipeline(
        ce.OrdinalEncoder(),
        RandomForestClassifier(
            n_estimators=100,
            class_weight='balanced',
            max_depth=max_depth,
            oob_score=True,
            n_jobs=-1
        )
    )

    pipe.fit(X_train, y_train)
    y_pred_proba = pipe.named_steps['randomforestclassifier'].oob_decision_funct
ion_[:, 1]
    print('Max Depth:', max_depth)
    print('ROC AUC, OOB:', roc_auc_score(y_train, y_pred_proba))

# Max Depth: 18
# ROC AUC, OOB: 0.7127616060911285
# Max Depth: 20
# ROC AUC, OOB: 0.7089254193634139
```

```
Max Depth: 2
ROC AUC, OOB: 0.697688412099441
Max Depth: 4
ROC AUC, OOB: 0.7068835972735312
Max Depth: 6
ROC AUC, OOB: 0.7117499487304035
Max Depth: 8
ROC AUC, OOB: 0.7158622618026634
Max Depth: 10
ROC AUC, OOB: 0.7186157500427036
Max Depth: 12
ROC AUC, OOB: 0.7197521082391606
Max Depth: 14
ROC AUC, OOB: 0.719424707668361
Max Depth: 16
ROC AUC, OOB: 0.7165724852701683
Max Depth: None
ROC AUC, OOB: 0.6982683952260536
CPU times: user 3h 7min 53s, sys: 2min 1s, total: 3h 9min 55s
Wall time: 12min 43s
```

# Feature Importances

We can look at feature importances. [But remember: (https://blog.datadive.net/selecting-good-features-part-iii-random-forests/)](https://blog.datadive.net/selecting-good-features-part-iii-random-forests/)

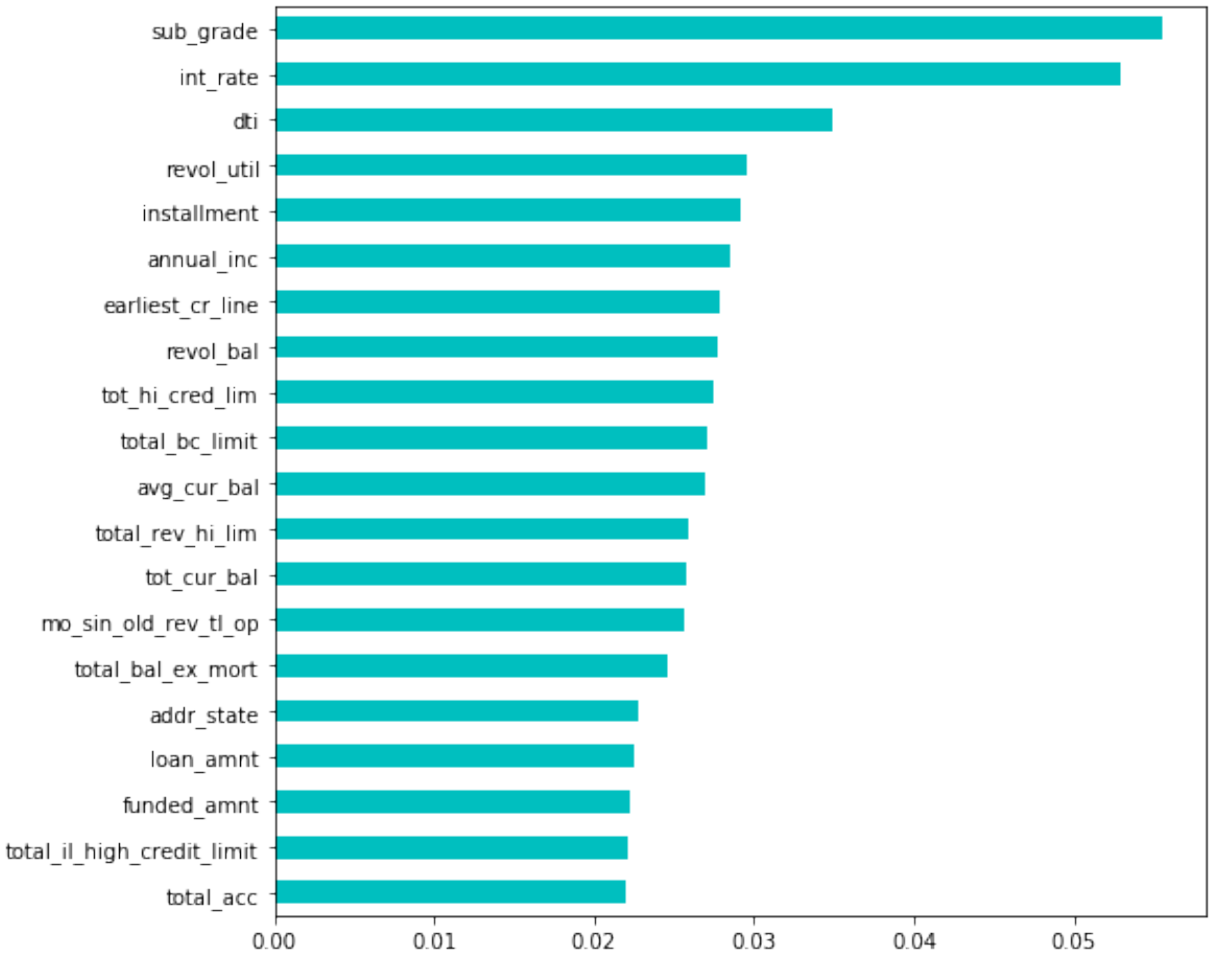
> Firstly, feature selection based on impurity reduction is biased towards preferring variables with more categories.
>
> Secondly, when the dataset has two (or more) correlated features, then from the point of view of the model, any of these correlated features can be used as the predictor, with no concrete preference of one over the others.

# Drop Column Importance / "Ablation Study"

`sub_grade` and `int_rate` are highly correlated. If we drop one of those features, the model uses the other more, so the score remains similar.

In [19]:

```
%%time
def show_feature_importances(
    pipe, X, y, estimator_name='randomforestclassifier',
    n=20, figsize=(8, 8), color='blue'):

    # pipe must not change dimensions of X dataframe
    pipe.fit(X, y)

    importances = pd.Series(
        pipe.named_steps[estimator_name].feature_importances_,
        X.columns)

    top_n = importances.sort_values(ascending=False)[:n]

    plt.figure(figsize=figsize)
    top_n.sort_values().plot.barh(color=color)


show_feature_importances(pipe, X_train, y_train, color='c')
```



In [20]:

```
%%time
cross_val_score(pipe, X_train.drop(columns='sub_grade'), y_train, cv=5, scoring='roc_auc')
```

```
CPU times: user 7min 11s, sys: 1min 24s, total: 8min 36s
Wall time: 8min 11s
```

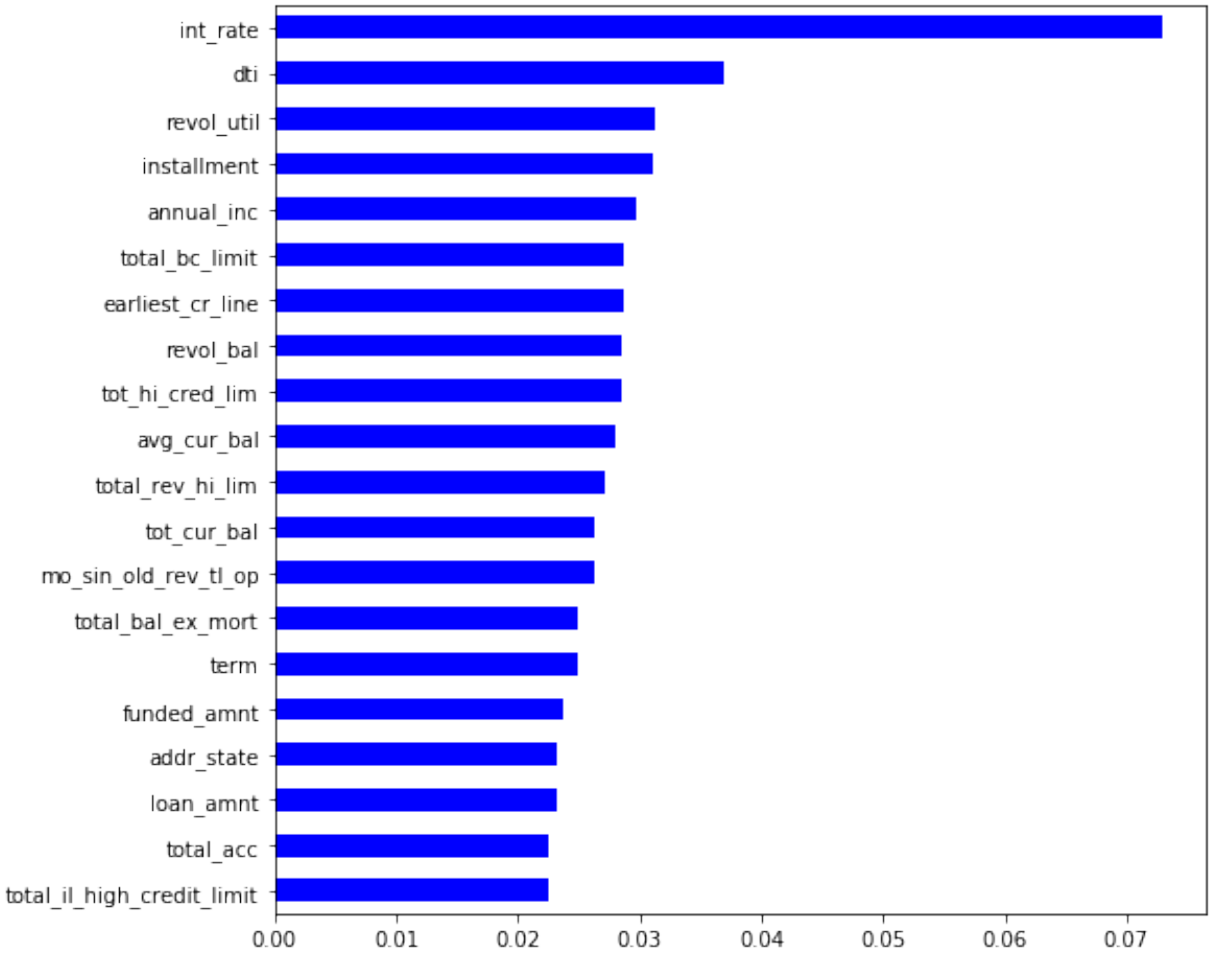
Out[20]:

```
array([0.71324234, 0.71119055, 0.71298131, 0.71590509, 0.71374193])
```

In [21]:

```
%%time
show_feature_importances(pipe, X_train.drop(columns='sub_grade'), y_train)
```

```
CPU times: user 35min 18s, sys: 12.8 s, total: 35min 30s
Wall time: 1min 52s
```



In [22]:

```
%%time
cross_val_score(pipe, X_train.drop(columns=['sub_grade']), y_train, cv=5, scorin
g='roc_auc')
```

```
CPU times: user 7min 6s, sys: 51.3 s, total: 7min 57s
Wall time: 7min 25s
```

Out[22]:

```
array([0.71563627, 0.71253498, 0.71285388, 0.71448075, 0.71513448])
```

In [27]:

```
sum([0.71324234, 0.71119055, 0.71298131, 0.71590509, 0.71374193])/5
```

Out[27]:

```
0.7134122439999999
```

In [28]:

```
sum([0.71563627, 0.71253498, 0.71285388, 0.71448075, 0.71513448])/5
```

Out[28]:

```
0.7141280720000001
```

But if we drop *both* features, then the score decreases:

In [29]:

```
%%time
cross_val_score(pipe, X_train.drop(columns=['sub_grade', 'int_rate']), y_train,
cv=5, scoring='roc_auc')
```

```
CPU times: user 7min 5s, sys: 1min 28s, total: 8min 34s
Wall time: 8min 4s
```

Out[29]:

```
array([0.70576527, 0.70152767, 0.70339744, 0.70515761, 0.70506773])
```

In [30]:

```
sum([0.70576527, 0.70152767, 0.70339744, 0.70515761, 0.70506773])/5
```

Out[30]:

```
0.7041831440000001
```

For more information, see Beware Default Random Forest Importances (https://explained.ai/rf-importance/index.html).

# Permutation Importance

Permutation Importance is a compromise between Feature Importance based on impurity reduction (which is the fastest) and Drop Column Importance (which is the "best.")

The ELI5 library documentation explains, (https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html)

> Importance can be measured by looking at how much the score (accuracy, F1, R^2, etc. - any score we're interested in) decreases when a feature is not available.
>
> To do that one can remove feature from the dataset, re-train the estimator and check the score. But it requires re-training an estimator for each feature, which can be computationally intensive. ...
>
> To avoid re-training the estimator we can remove a feature only from the test part of the dataset, and compute score without using this feature. It doesn't work as-is, because estimators expect feature to be present. So instead of removing a feature we can replace it with random noise - feature column is still there, but it no longer contains useful information. This method works if noise is drawn from the same distribution as original feature values (as otherwise estimator may fail). The simplest way to get such noise is to shuffle values for a feature, i.e. use other examples' feature values - this is how permutation importance is computed.
>
> The method is most suitable for computing feature importances when a number of columns (features) is not huge; it can be resource-intensive otherwise.

For more documentation on using this library, see:

- eli5.sklearn.PermutationImportance (https://eli5.readthedocs.io/en/latest/autodocs/sklearn.html#eli5.sklearn.permutation_importance.Permuta
- eli5.show_weights (https://eli5.readthedocs.io/en/latest/autodocs/eli5.html#eli5.show_weights)

In [31]:

```
%%time
import eli5
from eli5.sklearn import PermutationImportance

encoder = ce.OrdinalEncoder()
X_train_transformed = encoder.fit_transform(X_train)

model = RandomForestClassifier(
    n_estimators=100,
    class_weight='balanced',
    min_samples_leaf=0.005,
    n_jobs=-1)

model.fit(X_train_transformed, y_train)
permuter = PermutationImportance(model, scoring='roc_auc', n_iter=1, cv='prefit'
)
permuter.fit(X_train_transformed, y_train)
```

```
CPU times: user 58min 2s, sys: 16 s, total: 58min 18s
Wall time: 3min 39s
```

In [32]:

```
%%time
eli5.show_weights(permuter, top=None, feature_names=X_train_transformed.columns.
tolist())
```

```
CPU times: user 26.3 ms, sys: 51 µs, total: 26.4 ms
Wall time: 25.5 ms
```

Out[32]:

| Weight | Feature |
|---|---|
| 0.0240 ± 0.0000 | sub_grade |
| 0.0198 ± 0.0000 | int_rate |
| 0.0099 ± 0.0000 | term |
| 0.0033 ± 0.0000 | dti |
| 0.0023 ± 0.0000 | acc_open_past_24mths |
| 0.0016 ± 0.0000 | tot_hi_cred_lim |
| 0.0011 ± 0.0000 | annual_inc |
| 0.0011 ± 0.0000 | avg_cur_bal |
| 0.0010 ± 0.0000 | loan_amnt |
| 0.0009 ± 0.0000 | mort_acc |
| 0.0009 ± 0.0000 | funded_amnt |
| 0.0008 ± 0.0000 | home_ownership |
| 0.0008 ± 0.0000 | installment |
| 0.0008 ± 0.0000 | all_util |
| 0.0007 ± 0.0000 | total_bc_limit |
| 0.0005 ± 0.0000 | num_actv_rev_tl |
| 0.0005 ± 0.0000 | num_rev_tl_bal_gt_0 |
| 0.0004 ± 0.0000 | tot_cur_bal |
| 0.0004 ± 0.0000 | max_bal_bc |

| | |
|---|---|
| 0.0004 ± 0.0000 | revol_util |
| 0.0004 ± 0.0000 | num_tl_op_past_12m |
| 0.0003 ± 0.0000 | total_rev_hi_lim |
| 0.0003 ± 0.0000 | open_rv_24m |
| 0.0003 ± 0.0000 | emp_length |
| 0.0003 ± 0.0000 | mo_sin_rcnt_tl |
| 0.0003 ± 0.0000 | inq_last_6mths |
| 0.0002 ± 0.0000 | mths_since_rcnt_il |
| 0.0002 ± 0.0000 | il_util |
| 0.0002 ± 0.0000 | total_bal_il |
| 0.0002 ± 0.0000 | total_cu_tl |
| 0.0002 ± 0.0000 | mo_sin_old_rev_tl_op |
| 0.0001 ± 0.0000 | num_actv_bc_tl |
| 0.0001 ± 0.0000 | mo_sin_rcnt_rev_tl_op |
| 0.0001 ± 0.0000 | earliest_cr_line |
| 0.0001 ± 0.0000 | open_rv_12m |
| 0.0001 ± 0.0000 | initial_list_status |
| 0.0001 ± 0.0000 | open_acc_6m |
| 0.0001 ± 0.0000 | open_il_12m |
| 0.0001 ± 0.0000 | open_act_il |
| 0.0001 ± 0.0000 | num_op_rev_tl |
| 0.0000 ± 0.0000 | pct_tl_nvr_dlq |
| 0.0000 ± 0.0000 | revol_bal |
| 0.0000 ± 0.0000 | open_il_24m |
| 0.0000 ± 0.0000 | total_acc |
| 0.0000 ± 0.0000 | inq_last_12m |
| 0.0000 ± 0.0000 | inq_fi |
| 0.0000 ± 0.0000 | num_il_tl |
| 0.0000 ± 0.0000 | total_il_high_credit_limit |
| 0.0000 ± 0.0000 | num_bc_sats |
| 0.0000 ± 0.0000 | mths_since_recent_inq |
| 0.0000 ± 0.0000 | num_sats |
| 0.0000 ± 0.0000 | mths_since_last_delinq |
| 0.0000 ± 0.0000 | total_bal_ex_mort |
| 0.0000 ± 0.0000 | num_rev_accts |
| 0.0000 ± 0.0000 | delinq_2yrs |
| 0.0000 ± 0.0000 | num_bc_tl |
| 0.0000 ± 0.0000 | num_accts_ever_120_pd |
| 0.0000 ± 0.0000 | num_tl_120dpd_2m |
| 0.0000 ± 0.0000 | mths_since_last_record |
| 0.0000 ± 0.0000 | pub_rec |
| 0.0000 ± 0.0000 | open_acc |
| 0.0000 ± 0.0000 | addr_state |
| 0.0000 ± 0.0000 | mths_since_recent_bc_dlq |
| 0.0000 ± 0.0000 | tot_coll_amt |
| 0.0000 ± 0.0000 | mths_since_last_major_derog |
| 0.0000 ± 0.0000 | purpose |
| 0.0000 ± 0.0000 | mths_since_recent_revol_delinq |
| 0.0000 ± 0.0000 | num_tl_90g_dpd_24m |
| 0.0000 ± 0.0000 | pub_rec_bankruptcies |
| 0.0000 ± 0.0000 | bc_open_to_buy |
| 0.0000 ± 0.0000 | mo_sin_old_il_acct |
| 0 ± 0.0000 | dti_joint |
| 0 ± 0.0000 | sec_app_earliest_cr_line |
| 0 ± 0.0000 | emp_title_teacher |
| 0 ± 0.0000 | disbursement_method |
| 0 ± 0.0000 | sec_app_mths_since_last_major_derog |
| 0 ± 0.0000 | sec_app_collections_12_mths_ex_med |
| 0 ± 0.0000 | sec_app_chargeoff_within_12_mths |
| 0 ± 0.0000 | sec_app_num_rev_accts |
| 0 ± 0.0000 | sec_app_open_act_il |

```
        0 ± 0.0000    sec_app_revol_util
        0 ± 0.0000    sec_app_open_acc
        0 ± 0.0000    sec_app_mort_acc
        0 ± 0.0000    sec_app_inq_last_6mths
        0 ± 0.0000    revol_bal_joint
        0 ± 0.0000    annual_inc_joint
        0 ± 0.0000    tax_liens
        0 ± 0.0000    mths_since_recent_bc
        0 ± 0.0000    delinq_amnt
        0 ± 0.0000    chargeoff_within_12_mths
        0 ± 0.0000    emp_title_manager
        0 ± 0.0000    collections_12_mths_ex_med
        0 ± 0.0000    acc_now_delinq
        0 ± 0.0000    application_type
        0 ± 0.0000    emp_title_owner
  -0.0000 ± 0.0000    percent_bc_gt_75
  -0.0000 ± 0.0000    num_tl_30dpd
  -0.0000 ± 0.0000    bc_util
```

We can use Permutation Importance weights for feature selection. For example, we can remove features with zero weight. The model trains faster and the score does not decrease.

In [33]:

```
%%time
subset = X_train.columns[permuter.feature_importances_ > 0]
cross_val_score(pipe, X_train[subset], y_train, cv=5, scoring='roc_auc', verbose
=10)
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurr
ent workers.

[CV]  ................................................................
..

/home/ec2-user/anaconda3/envs/python3/lib/python3.6/site-packages/sk
learn/model_selection/_validation.py:542: FutureWarning: From versio
n 0.22, errors during fit will result in a cross validation score of
NaN by default. Use error_score='raise' if you want an exception rai
sed or error_score=np.nan to adopt the behavior from version 0.22.
  FutureWarning)

---------------------------------------------------------------------
-------
KeyError                                  Traceback (most recent cal
l last)
~/anaconda3/envs/python3/lib/python3.6/site-packages/pandas/core/ind
exes/base.py in get_loc(self, key, method, tolerance)
   2889             try:
-> 2890                 return self._engine.get_loc(key)
   2891             except KeyError:

pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()
```

```
pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.Py
ObjectHashTable.get_item()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.Py
ObjectHashTable.get_item()

KeyError: 'application_type'

During handling of the above exception, another exception occurred:

KeyError                                  Traceback (most recent cal
l last)
<timed exec> in <module>()

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/model_s
election/_validation.py in cross_val_score(estimator, X, y, groups,
scoring, cv, n_jobs, verbose, fit_params, pre_dispatch, error_score)
    400                                     fit_params=fit_params,
    401                                     pre_dispatch=pre_dispatch,
--> 402                                     error_score=error_score)
    403     return cv_results['test_score']
    404

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/model_s
election/_validation.py in cross_validate(estimator, X, y, groups, s
coring, cv, n_jobs, verbose, fit_params, pre_dispatch, return_train_
score, return_estimator, error_score)
    238             return_times=True, return_estimator=return_estim
ator,
    239             error_score=error_score)
--> 240         for train, test in cv.split(X, y, groups))
    241
    242     zipped_scores = list(zip(*scores))

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/externa
ls/joblib/parallel.py in __call__(self, iterable)
    915             # remaining jobs.
    916             self._iterating = False
--> 917             if self.dispatch_one_batch(iterator):
    918                 self._iterating = self._original_iterator is
not None
    919

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/externa
ls/joblib/parallel.py in dispatch_one_batch(self, iterator)
    757                 return False
    758             else:
```

```
--> 759                 self._dispatch(tasks)
    760                 return True
    761

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/externa
ls/joblib/parallel.py in _dispatch(self, batch)
    714         with self._lock:
    715             job_idx = len(self._jobs)
--> 716             job = self._backend.apply_async(batch, callback=
cb)
    717             # A job can complete so quickly than its callbac
k is
    718             # called before we get here, causing self._jobs
to

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/externa
ls/joblib/_parallel_backends.py in apply_async(self, func, callback)
    180     def apply_async(self, func, callback=None):
    181         """Schedule a func to be run"""
--> 182         result = ImmediateResult(func)
    183         if callback:
    184             callback(result)

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/externa
ls/joblib/_parallel_backends.py in __init__(self, batch)
    547         # Don't delay the application, to avoid keeping the
input
    548         # arguments in memory
--> 549         self.results = batch()
    550
    551     def get(self):

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/externa
ls/joblib/parallel.py in __call__(self)
    223         with parallel_backend(self._backend, n_jobs=self._n_
jobs):
    224             return [func(*args, **kwargs)
--> 225                     for func, args, kwargs in self.items]
    226
    227     def __len__(self):

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/externa
ls/joblib/parallel.py in <listcomp>(.0)
    223         with parallel_backend(self._backend, n_jobs=self._n_
jobs):
    224             return [func(*args, **kwargs)
--> 225                     for func, args, kwargs in self.items]
    226
    227     def __len__(self):
```

```
~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/model_s
election/_validation.py in _fit_and_score(estimator, X, y, scorer, t
rain, test, verbose, parameters, fit_params, return_train_score, ret
urn_parameters, return_n_test_samples, return_times, return_estimato
r, error_score)
    526             estimator.fit(X_train, **fit_params)
    527         else:
--> 528             estimator.fit(X_train, y_train, **fit_params)
    529
    530     except Exception as e:

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/pipelin
e.py in fit(self, X, y, **fit_params)
    263             This estimator
    264         """
--> 265         Xt, fit_params = self._fit(X, y, **fit_params)
    266         if self._final_estimator is not None:
    267             self._final_estimator.fit(Xt, y, **fit_params)

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/pipelin
e.py in _fit(self, X, y, **fit_params)
    228                 Xt, fitted_transformer = fit_transform_one_c
ached(
    229                     cloned_transformer, Xt, y, None,
--> 230                     **fit_params_steps[name])
    231                 # Replace the transformer of the step with t
he fitted
    232                 # transformer. This is necessary when loadin
g the transformer

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/externa
ls/joblib/memory.py in __call__(self, *args, **kwargs)
    340
    341     def __call__(self, *args, **kwargs):
--> 342         return self.func(*args, **kwargs)
    343
    344     def call_and_shelve(self, *args, **kwargs):

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/pipelin
e.py in _fit_transform_one(transformer, X, y, weight, **fit_params)
    612 def _fit_transform_one(transformer, X, y, weight, **fit_para
ms):
    613     if hasattr(transformer, 'fit_transform'):
--> 614         res = transformer.fit_transform(X, y, **fit_params)
    615     else:
    616         res = transformer.fit(X, y, **fit_params).transform(
X)

~/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/base.py
in fit_transform(self, X, y, **fit_params)
```

```
    465         else:
    466             # fit method of arity 2 (supervised transformati
on)
--> 467             return self.fit(X, y, **fit_params).transform(X)
    468
    469

~/anaconda3/envs/python3/lib/python3.6/site-packages/category_encode
rs/ordinal.py in fit(self, X, y, **kwargs)
    139             cols=self.cols,
    140             handle_unknown=self.handle_unknown,
--> 141             handle_missing=self.handle_missing
    142         )
    143         self.mapping = categories

~/anaconda3/envs/python3/lib/python3.6/site-packages/category_encode
rs/ordinal.py in ordinal_encoding(X_in, mapping, cols, handle_unknow
n, handle_missing)
    288             for switch in mapping:
    289                 column = switch.get('col')
--> 290                 X[column] = X[column].map(switch['mapping'])
    291
    292                 try:

~/anaconda3/envs/python3/lib/python3.6/site-packages/pandas/core/fra
me.py in __getitem__(self, key)
   2973             if self.columns.nlevels > 1:
   2974                 return self._getitem_multilevel(key)
-> 2975             indexer = self.columns.get_loc(key)
   2976             if is_integer(indexer):
   2977                 indexer = [indexer]

~/anaconda3/envs/python3/lib/python3.6/site-packages/pandas/core/ind
exes/base.py in get_loc(self, key, method, tolerance)
   2890                 return self._engine.get_loc(key)
   2891             except KeyError:
-> 2892                 return self._engine.get_loc(self._maybe_cast
_indexer(key))
   2893         indexer = self.get_indexer([key], method=method,
tolerance=tolerance)
   2894         if indexer.ndim > 1 or indexer.size > 1:

pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.Py
ObjectHashTable.get_item()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.Py
```

```
ObjectHashTable.get_item()
```

```
KeyError: 'application_type'
```

In [ ]: