Applied Machine Learning Workshop

Come to our workshop to learn how to build Python machine learning systems for your data science projects!

Thursday Feb. 28

Pupin 420, 8-9pm

Columbia Data Science Society

Connect with Us. Sign up for our News Letter.

https://cdssatcu.com

Basic Introductory Tutorial

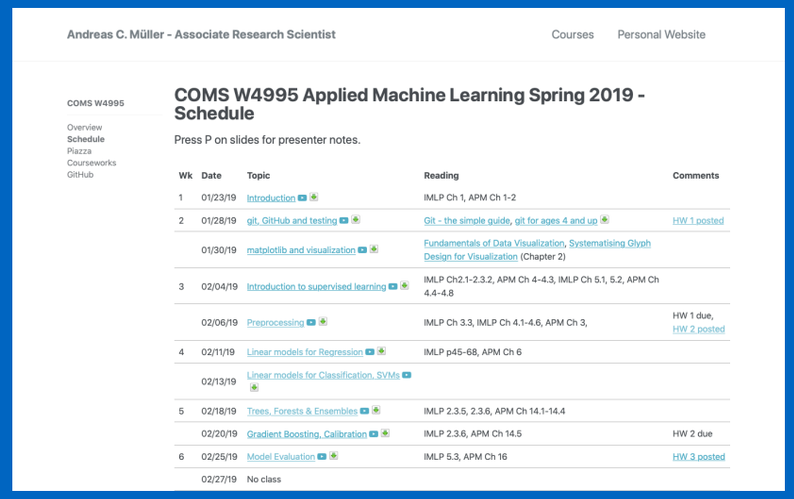
https://colab.research.google.com/notebooks/welcome.ipynb

Sci-Kit Learn: https://scikit-learn.org/stable/

XGBoost: https://xgboost.readthedocs.io/en/latest/

Acknowledements:

The content of this talk is motivated from the two classes, I am taking this semester at Columbia Data Science Institute.



Andreas C. Müller - Associate Research Scientist

Courses Personal Website

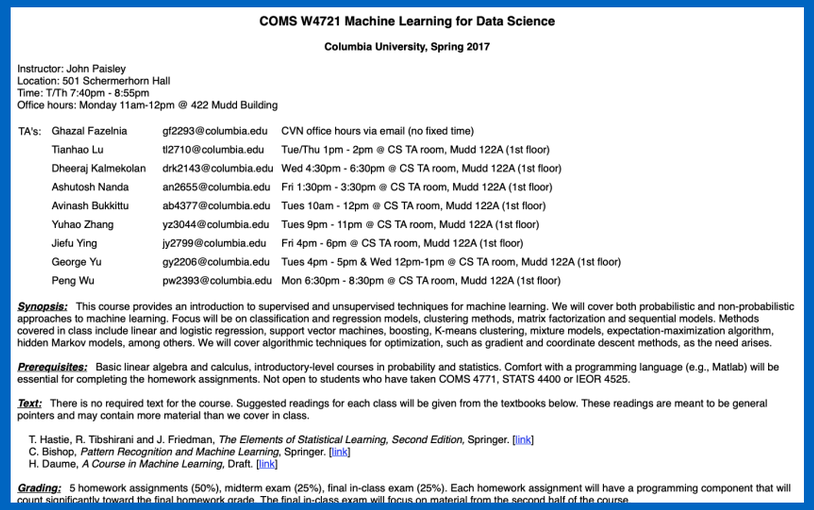
COMS W4995

Overview
Schedule
Piazza
Courseworks
GitHub

COMS W4995 Applied Machine Learning Spring 2019 - Schedule

Press P on slides for presenter notes.

Wk	Date	Topic	Reading	Comments
1	01/23/19	Introduction	IMLP Ch 1, APM Ch 1-2	
2	01/28/19	git, GitHub and testing	Git - the simple guide, git for ages 4 and up	HW 1 posted
	01/30/19	matplotlib and visualization	Fundamentals of Data Visualization, Systematising Glyph Design for Visualization (Chapter 2)	
3	02/04/19	Introduction to supervised learning	IMLP Ch2.1-2.3.2, APM Ch 4-4.3, IMLP Ch 5.1, 5.2, APM Ch 4.4-4.8	
	02/06/19	Preprocessing	IMLP Ch 3.3, IMLP Ch 4.1-4.6, APM Ch 3,	HW 1 due, HW 2 posted
4	02/11/19	Linear models for Regression	IMLP p45-68, APM Ch 6	
	02/13/19	Linear models for Classification, SVMs		
5	02/18/19	Trees, Forests & Ensembles	IMLP 2.3.5, 2.3.6, APM Ch 14.1-14.4	
	02/20/19	Gradient Boosting, Calibration	IMLP 2.3.6, APM Ch 14.5	HW 2 due
6	02/25/19	Model Evaluation	IMLP 5.3, APM Ch 16	HW 3 posted
	02/27/19	No class		



COMS W4721 Machine Learning for Data Science

Columbia University, Spring 2017

Instructor: John Paisley
Location: 501 Schermerhorn Hall
Time: T/Th 7:40pm - 8:55pm
Office hours: Monday 11am-12pm @ 422 Mudd Building

TA's:	Ghazal Fazelnia	gf2293@columbia.edu	CVN office hours via email (no fixed time)
	Tianhao Lu	tl2710@columbia.edu	Tue/Thu 1pm - 2pm @ CS TA room, Mudd 122A (1st floor)
	Dheeraj Kalmekolan	drk2143@columbia.edu	Wed 4:30pm - 6:30pm @ CS TA room, Mudd 122A (1st floor)
	Ashutosh Nanda	an2655@columbia.edu	Fri 1:30pm - 3:30pm @ CS TA room, Mudd 122A (1st floor)
	Avinash Bukkittu	ab4377@columbia.edu	Tues 10am - 12pm @ CS TA room, Mudd 122A (1st floor)
	Yuhao Zhang	yz3044@columbia.edu	Tues 9pm - 11pm @ CS TA room, Mudd 122A (1st floor)
	Jiefu Ying	jy2799@columbia.edu	Fri 4pm - 6pm @ CS TA room, Mudd 122A (1st floor)
	George Yu	gy2206@columbia.edu	Tues 4pm - 5pm & Wed 12pm-1pm @ CS TA room, Mudd 122A (1st floor)
	Peng Wu	pw2393@columbia.edu	Mon 6:30pm - 8:30pm @ CS TA room, Mudd 122A (1st floor)

Synopsis: This course provides an introduction to supervised and unsupervised techniques for machine learning. We will cover both probabilistic and non-probabilistic approaches to machine learning. Focus will be on classification and regression models, clustering methods, matrix factorization and sequential models. Methods covered in class include linear and logistic regression, support vector machines, boosting, K-means clustering, mixture models, expectation-maximization algorithm, hidden Markov models, among others. We will cover algorithmic techniques for optimization, such as gradient and coordinate descent methods, as the need arises.

Prerequisites: Basic linear algebra and calculus, introductory-level courses in probability and statistics. Comfort with a programming language (e.g., Matlab) will be essential for completing the homework assignments. Not open to students who have taken COMS 4771, STATS 4400 or IEOR 4525.

Text: There is no required text for the course. Suggested readings for each class will be given from the textbooks below. These readings are meant to be general pointers and may contain more material than we cover in class.

T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning, Second Edition,* Springer. [link]
C. Bishop, *Pattern Recognition and Machine Learning,* Springer. [link]
H. Daume, *A Course in Machine Learning,* Draft. [link]

Grading: 5 homework assignments (50%), midterm exam (25%), final in-class exam (25%). Each homework assignment will have a programming component that will count significantly toward the final homework grade. The final in-class exam will focus on material from the second half of the course.

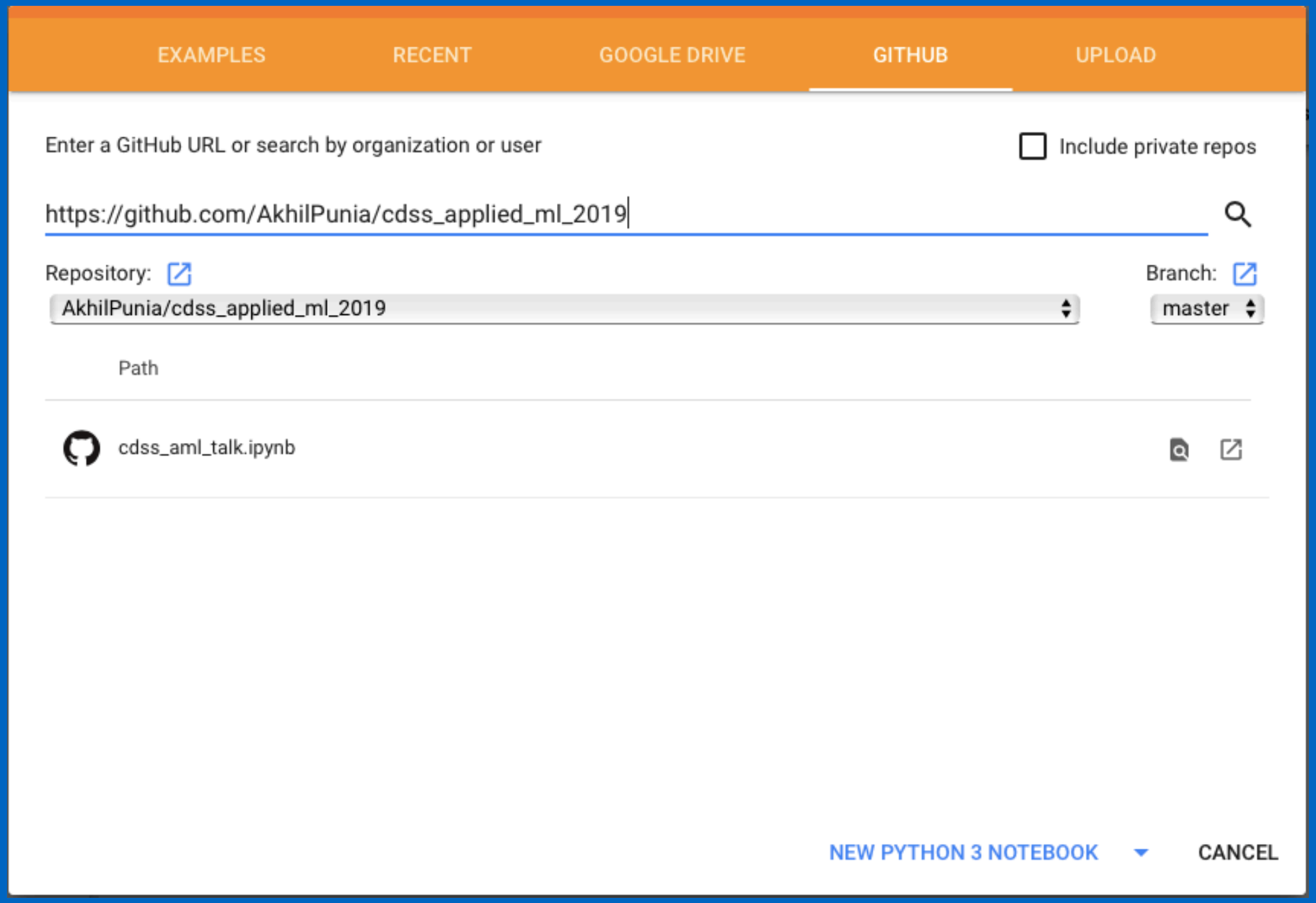
COMS W4995 Applied Machine Learning

taught by Andread C. Müller, core contributer of sci-kit learn and author of the O'Reilly book "Introduction to machine learning with Python", describing a practical approach to machine learning with python and scikit-learn.

COMS W4721 Machine Learning for Data Science

taught by Prof. John Paisley. He also teaches a course on Machine Learning through edx.The content of his on-campus course largely overlaps with his edx class.

Let’s get started.



Part 1/5

Let's talk about the Data

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

4,189 teams · Ongoing

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Team | My Submissions | Submit Predictions

Overview

- Description
- Evaluation
- Tutorials
- Frequently Asked Questions

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Featured Prediction Competition

Zillow Prize: Zillow's Home Value Prediction (Zestimate)

Can you improve the algorithm that changed the world of real estate?

$1,200,000
Prize Money

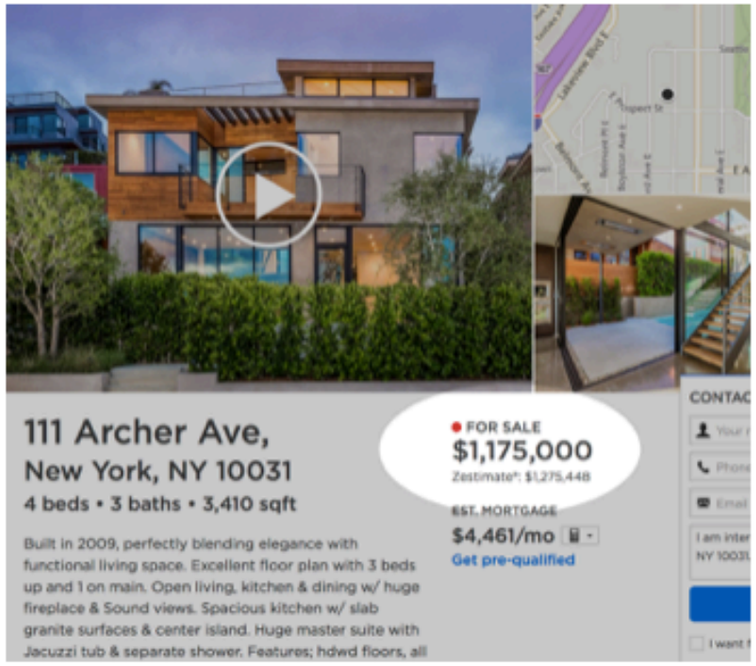
Zillow · 3,779 teams · a year ago

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Team | My Submissions | **Late Submission**

Overview

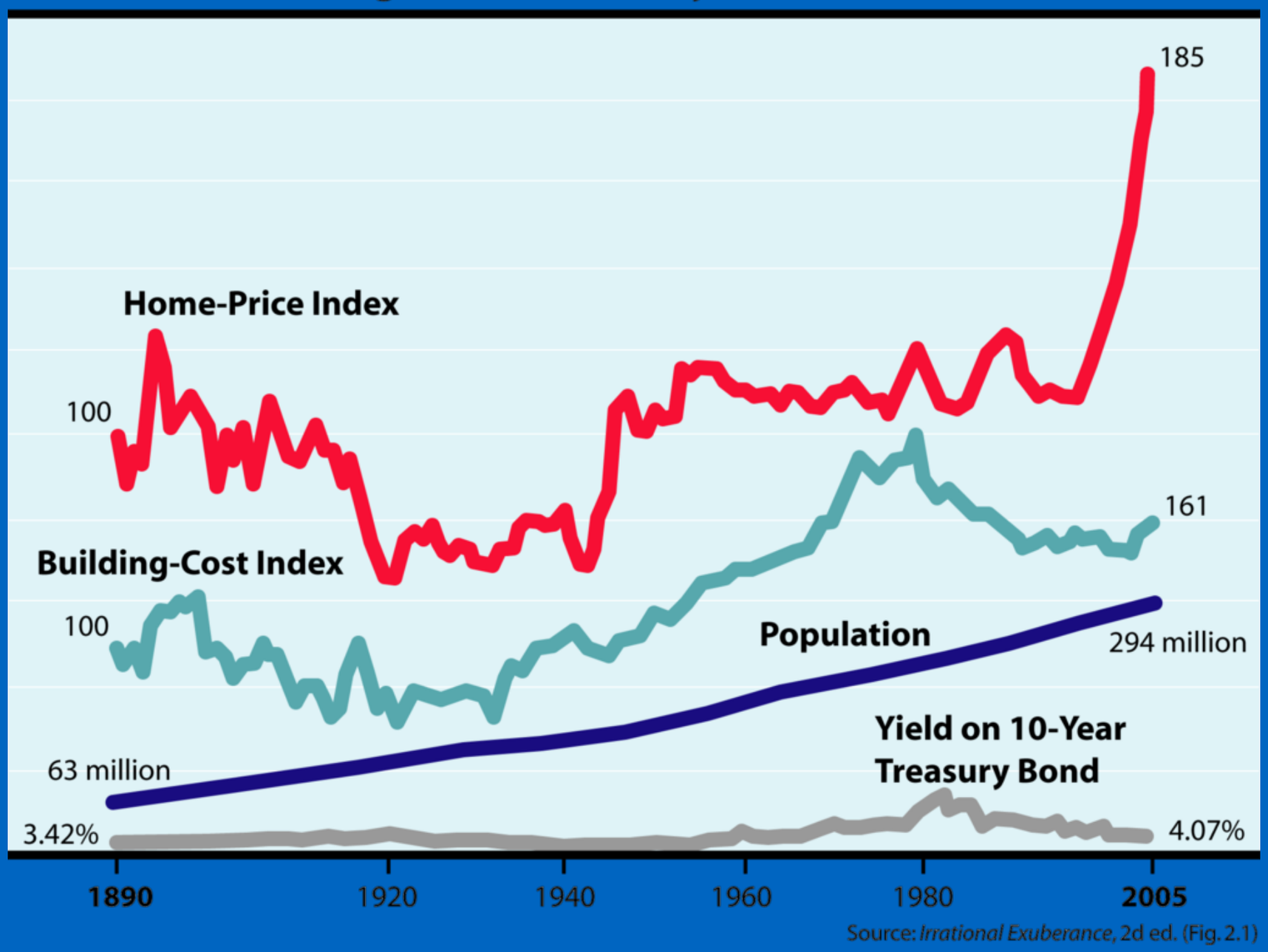
- **Description**
- **Evaluation**
- **Prizes**
- **Timeline**
- **Competition Overview**

Zillow's Zestimate home valuation has shaken up the U.S. real estate industry since first released 11 years ago.



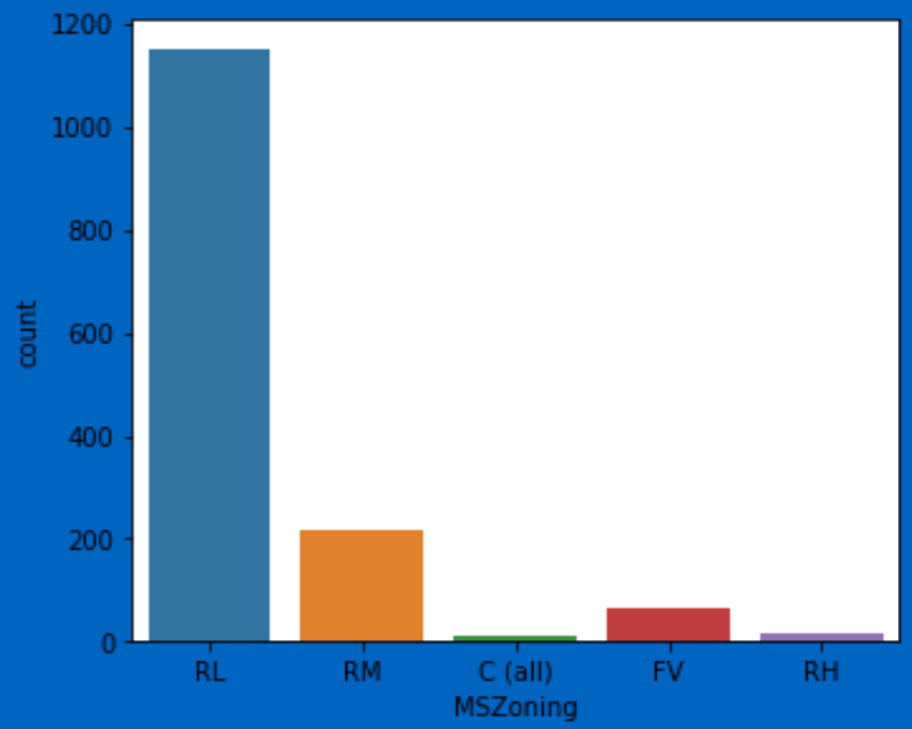
A home is often the largest and most expensive purchase a person makes in his or her lifetime. Ensuring homeowners have a trusted way to monitor this asset is incredibly important. The Zestimate was created to give consumers as much information as possible about homes and the housing market, marking the first time consumers had access to this type of home value information at no cost.

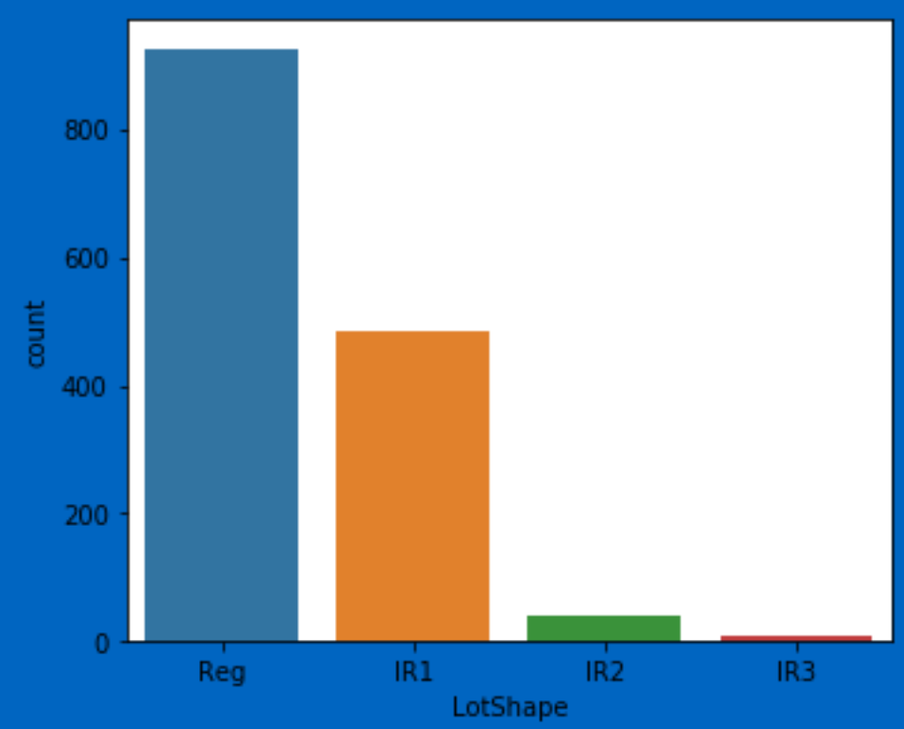
"Zestimates" are estimated home values based on 7.5 million statistical and machine learning models that analyze hundreds of data points on each property. And, by continually improving the median margin of error (from 14% at the onset to 5% today), Zillow has since become established as one of the largest, most trusted marketplaces for real estate information in the U.S. and a leading example of impactful machine learning.

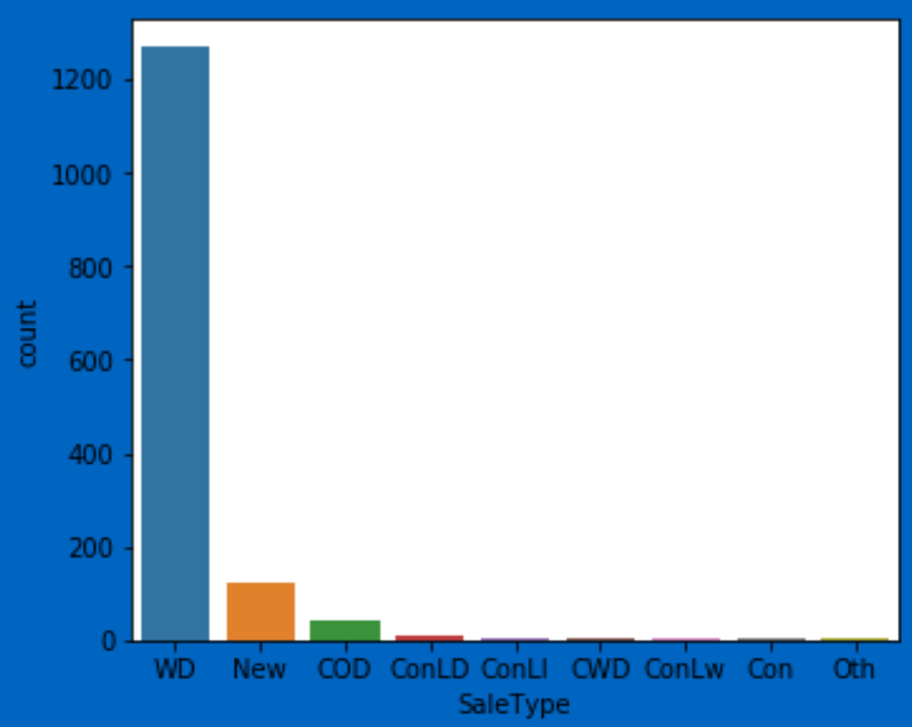

Inflation-adjusted U.S. home prices, Population,
Building costs, and Bond yields (1890–2005)
Home-Price Index
100
185
Building-Cost Index
100
161
Population
294 million
63 million
Yield on 10-Year
Treasury Bond
3.42%
4.07%
1890
1920
1940
1960
1980
2005
Source: Irrational Exuberance, 2d ed. (Fig. 2.1)

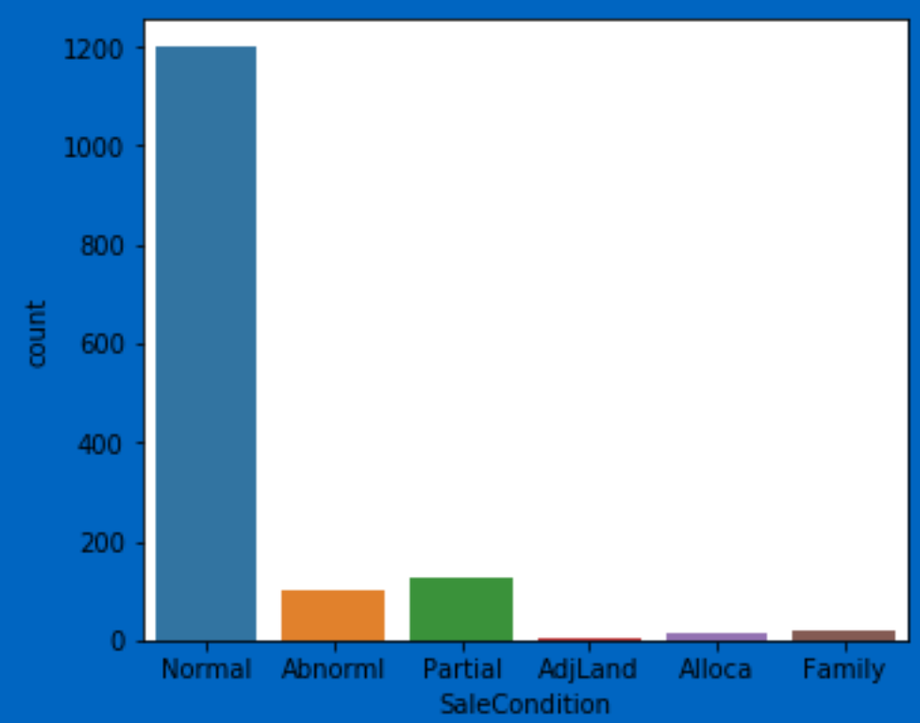
What's with a feature?

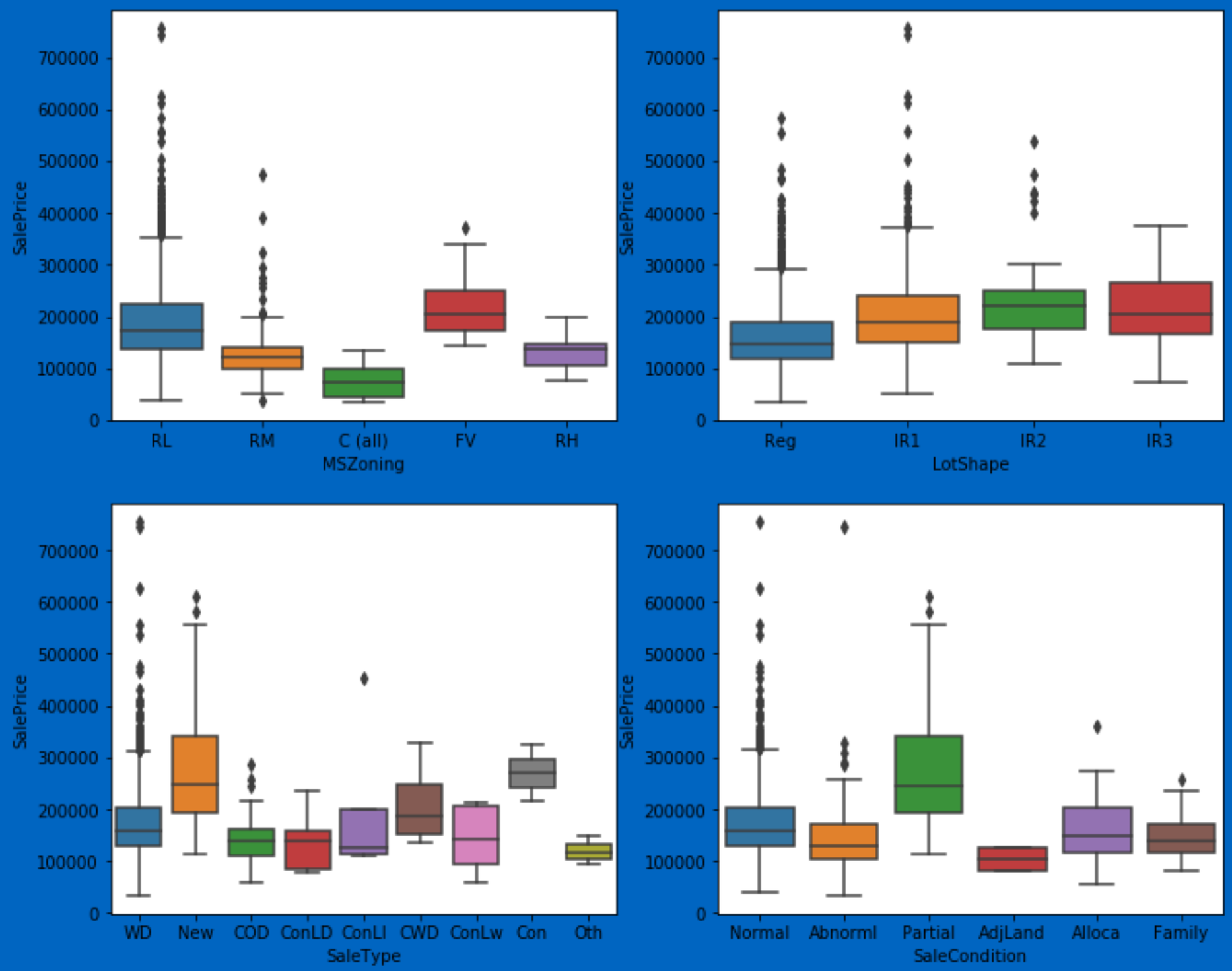
Continous


count
1200
1000
800
600
400
200
0
RL
RM
C (all)
FV
RH
MSZoning

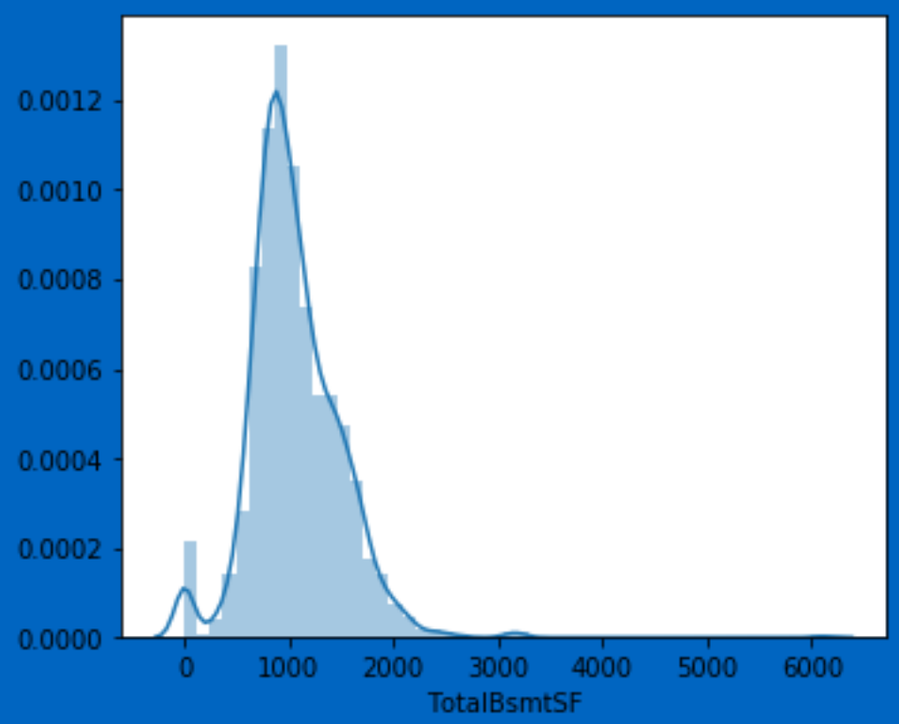

count
800
600
400
200
0
Reg
IR1
IR2
IR3
LotShape

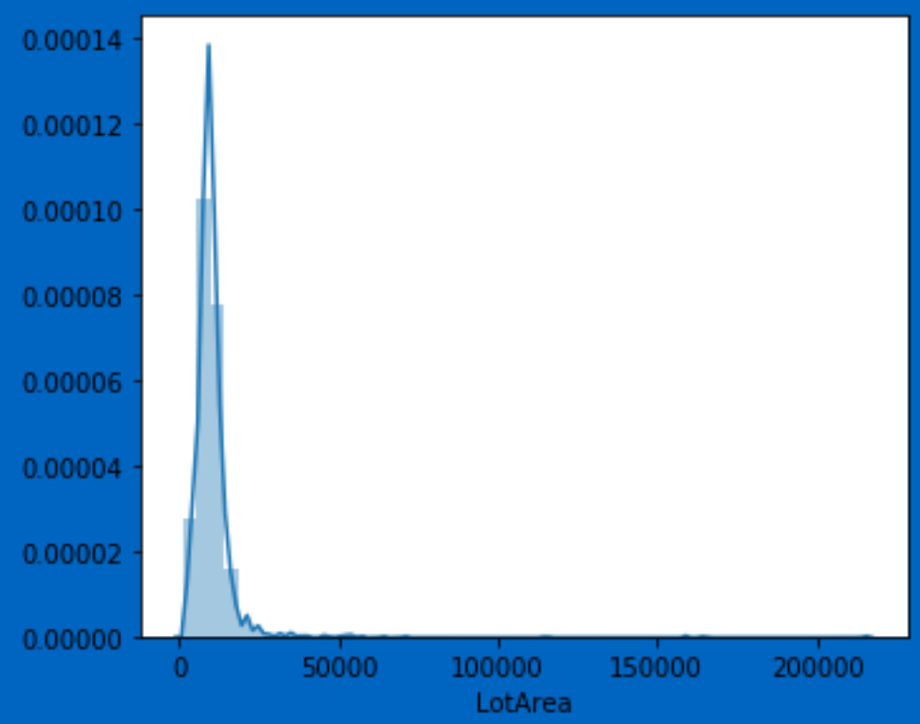

count
1200
1000
800
600
400
200
0
WD
New
COD
ConLD
ConLI
CWD
ConLw
Con
Oth
SaleType

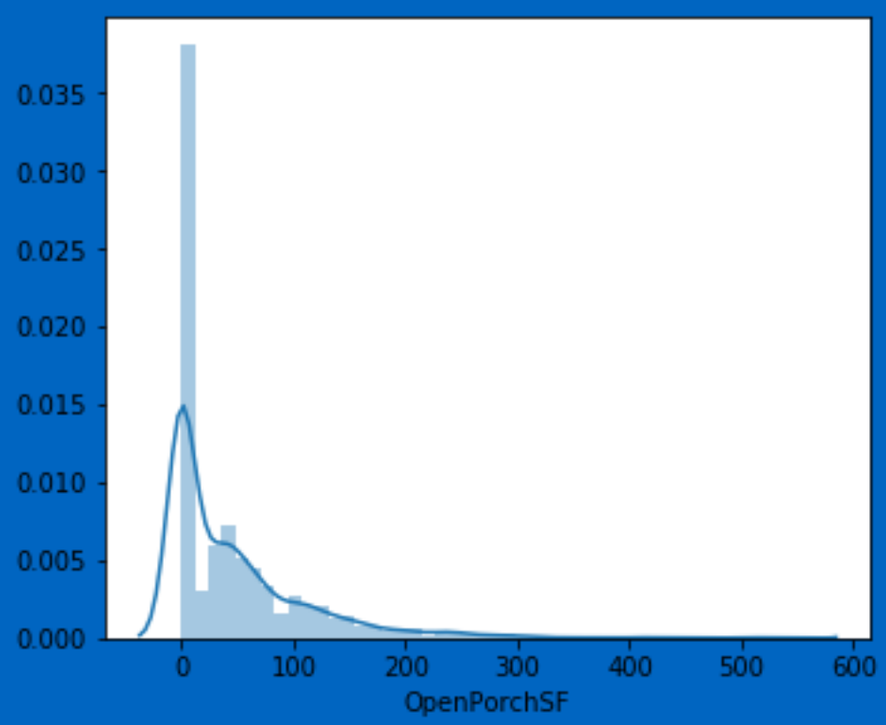

count
1200
1000
800
600
400
200
0
Normal
Abnorml
Partial
AdjLand
Alloca
Family
SaleCondition

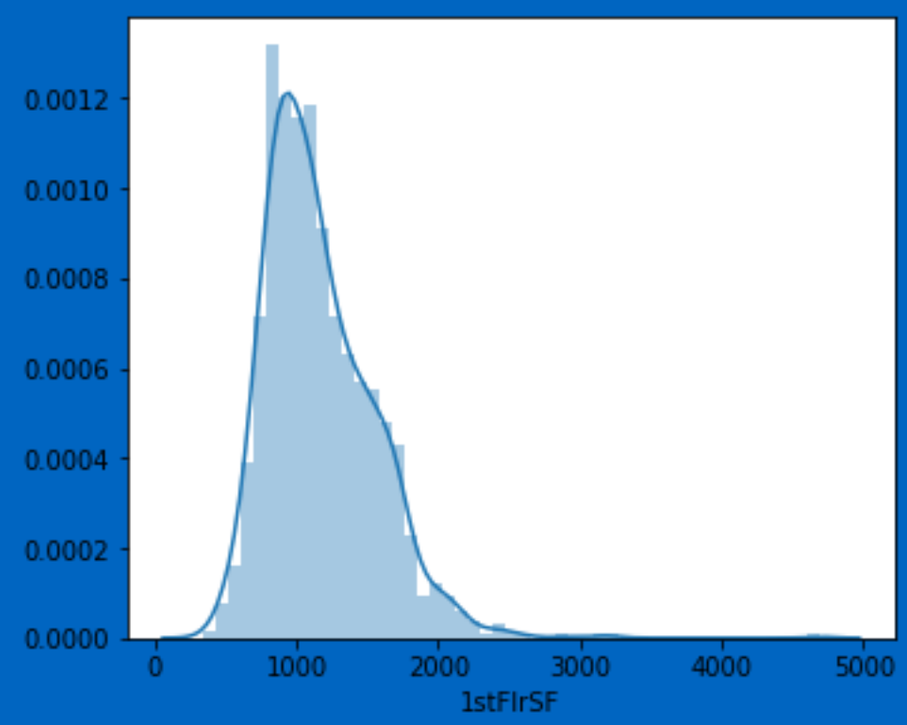

SalePrice
MSZoning
RL
RM
C (all)
FV
RH
LotShape
Reg
IR1
IR2
IR3
SaleType
WD
New
COD
ConLD
ConLI
CWD
ConLw
Con
Oth
SaleCondition
Normal
Abnorml
Partial
AdjLand
Alloca
Family

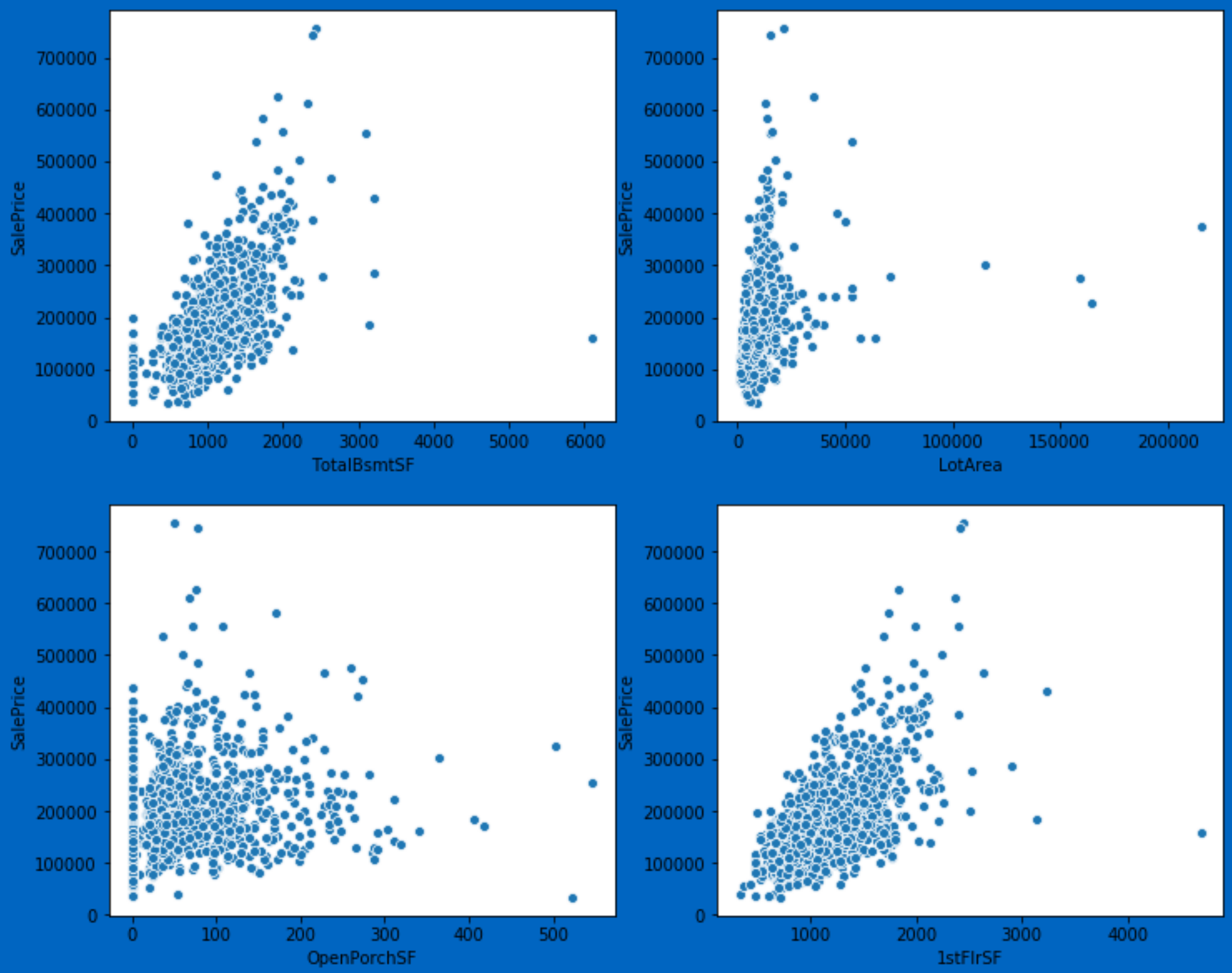
Categorical



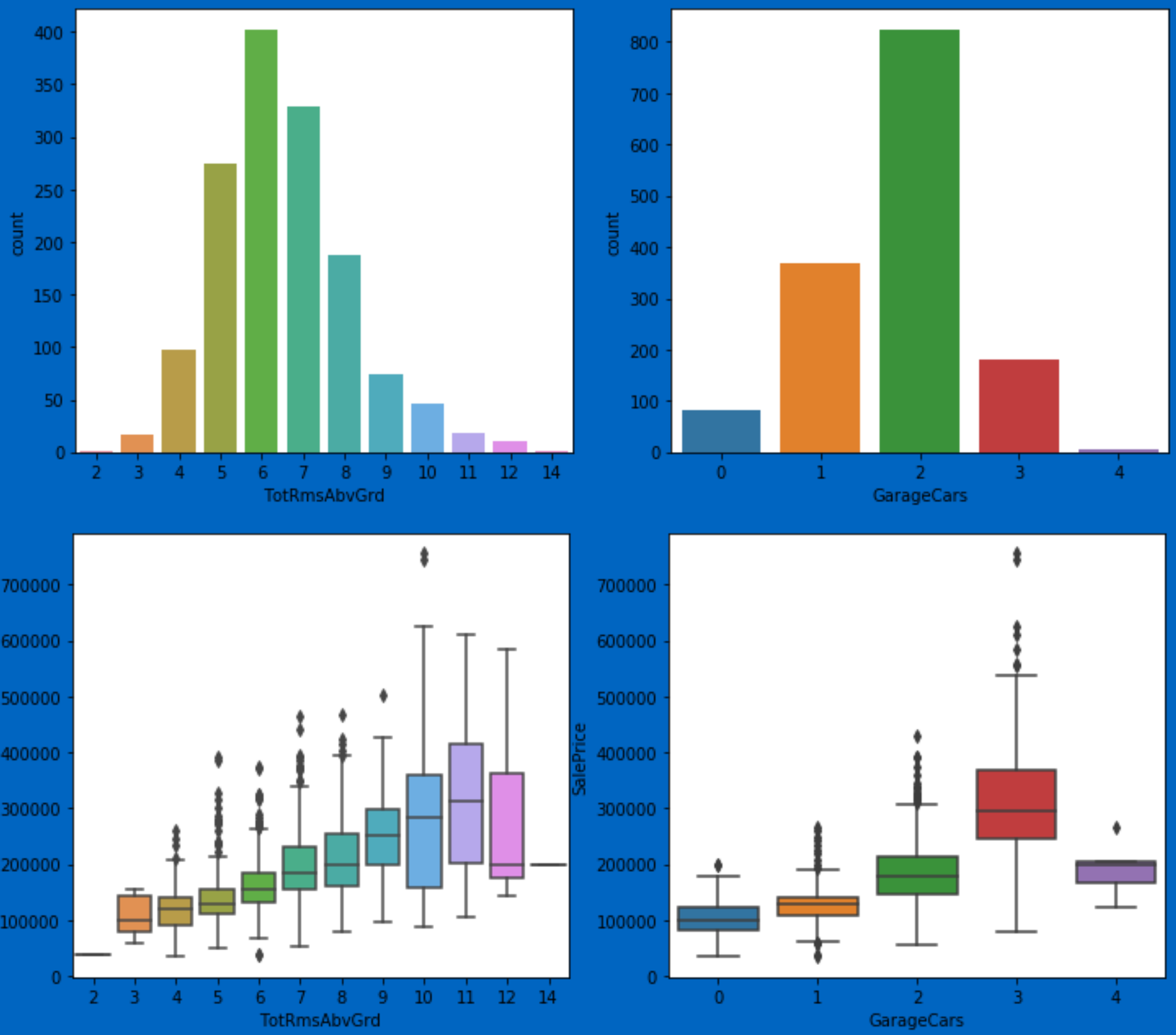




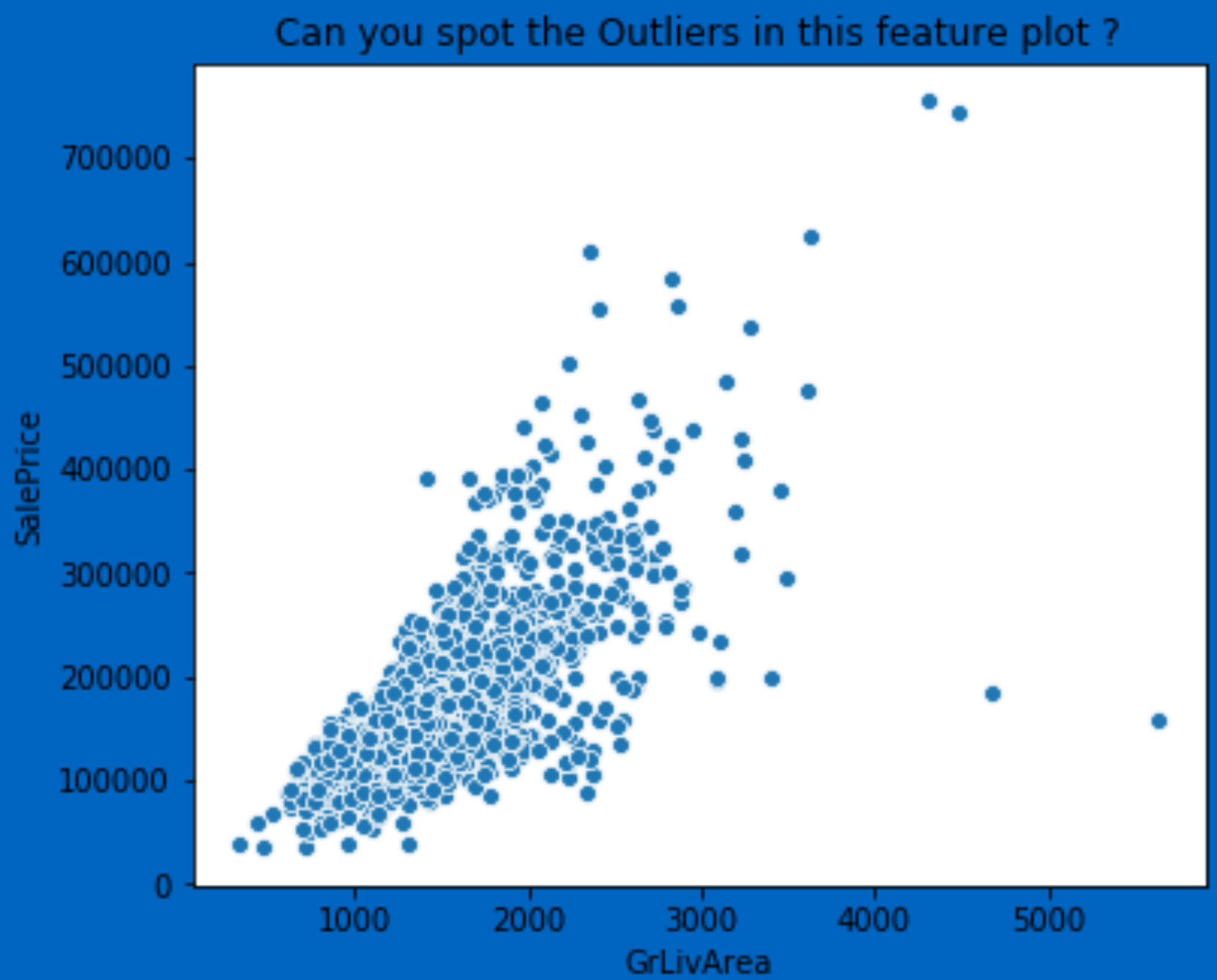



SalePrice
TotalBsmtSF
LotArea
OpenPorchSF
1stFlrSF

Ordinal


count
TotRmsAbvGrd
count
GarageCars
SalePrice
TotRmsAbvGrd
SalePrice
GarageCars

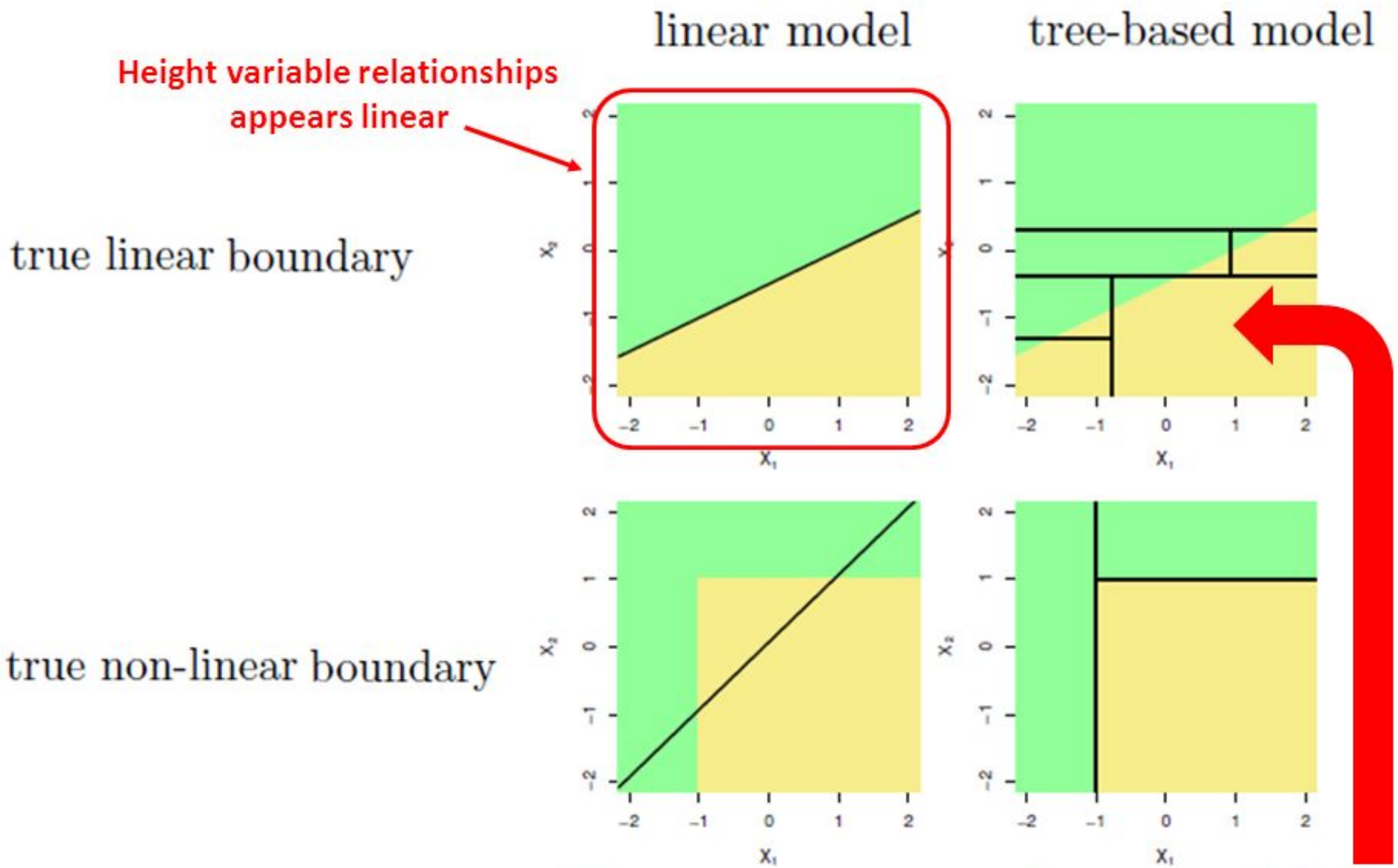
Outliers


Can you spot the Outliers in this feature plot ?
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
1000
2000
3000
4000
5000
GrLivArea

Part 2/5

What's in a Model ?

Linear Models vs. Decision Trees

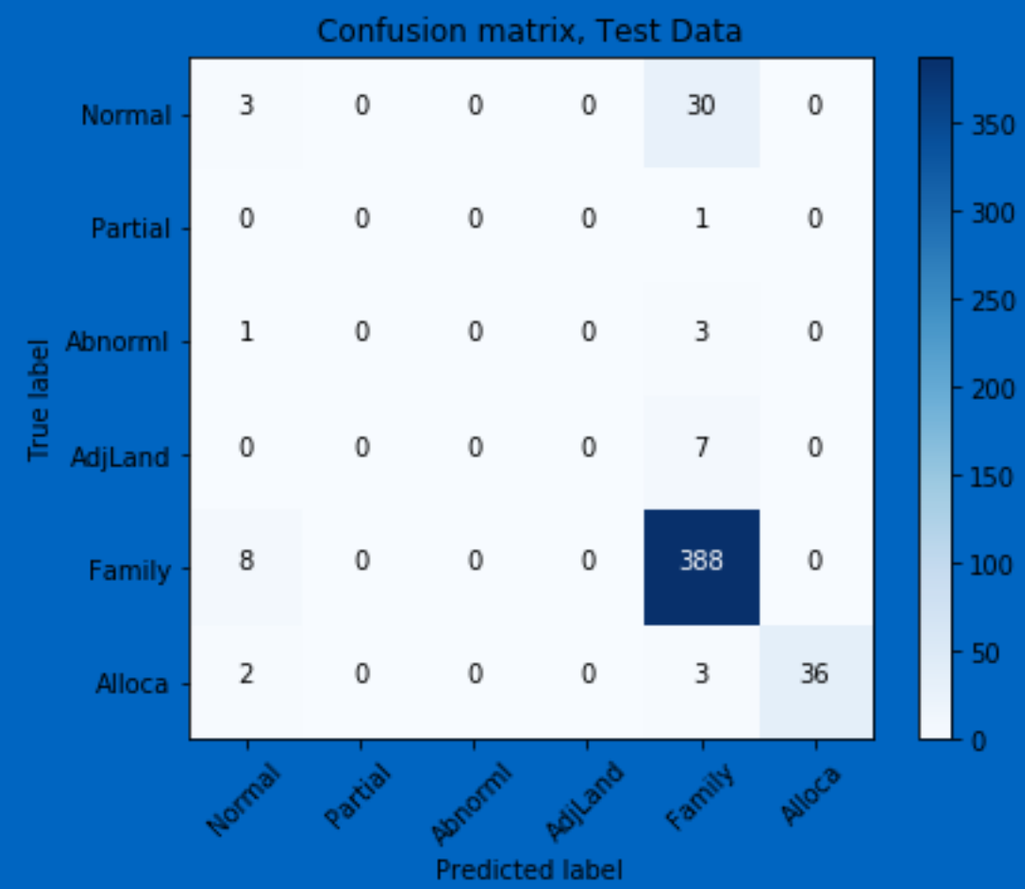


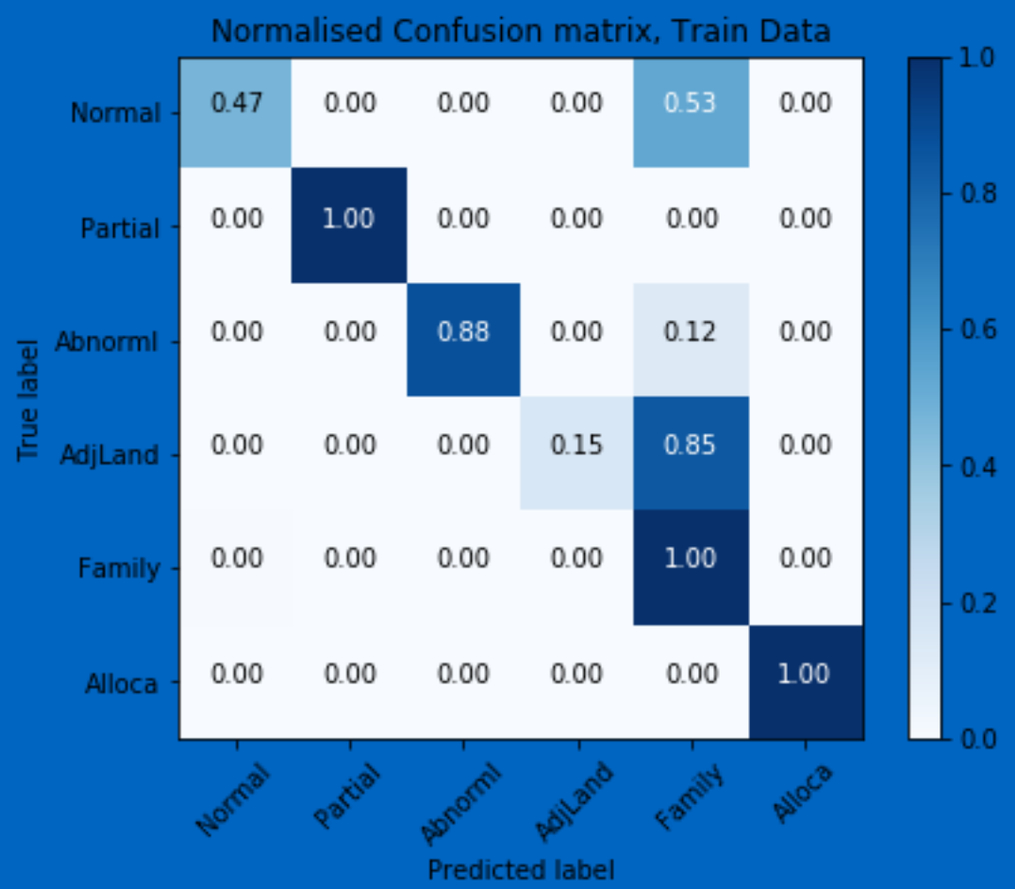
Decision Trees Would Not Appear to be as Helpful

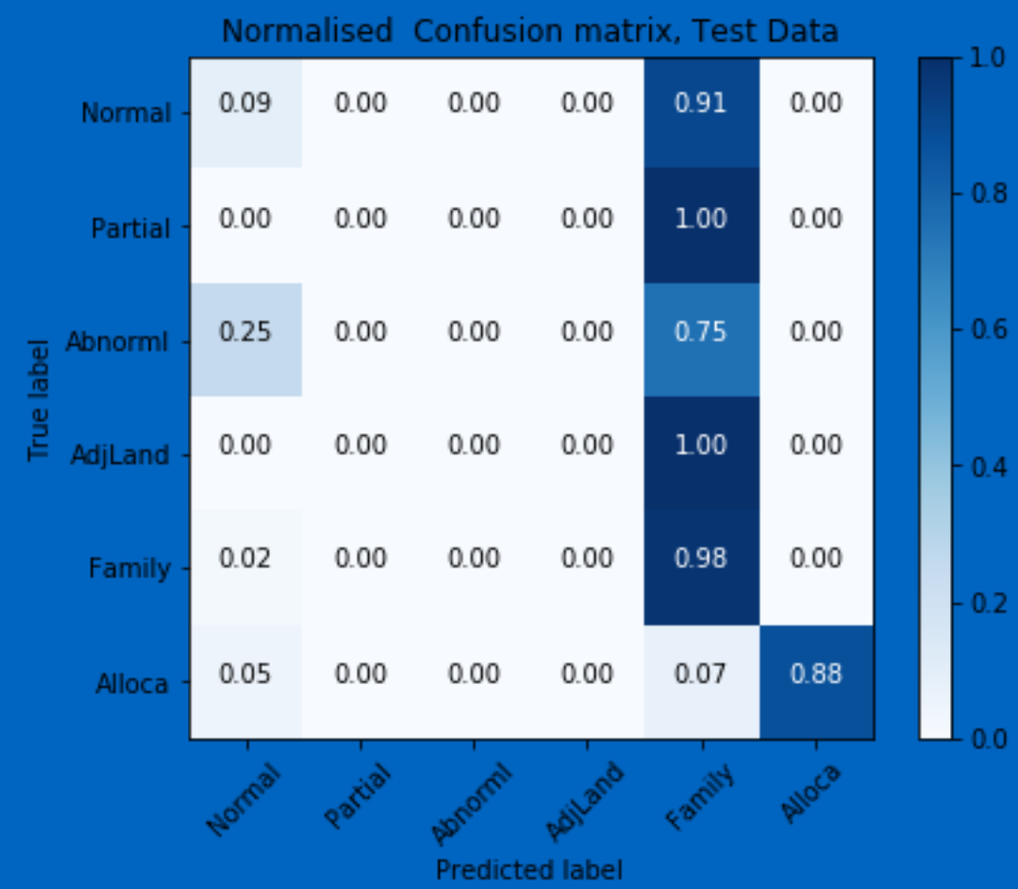
Source: https://slideplayer.com/slide/8334889/

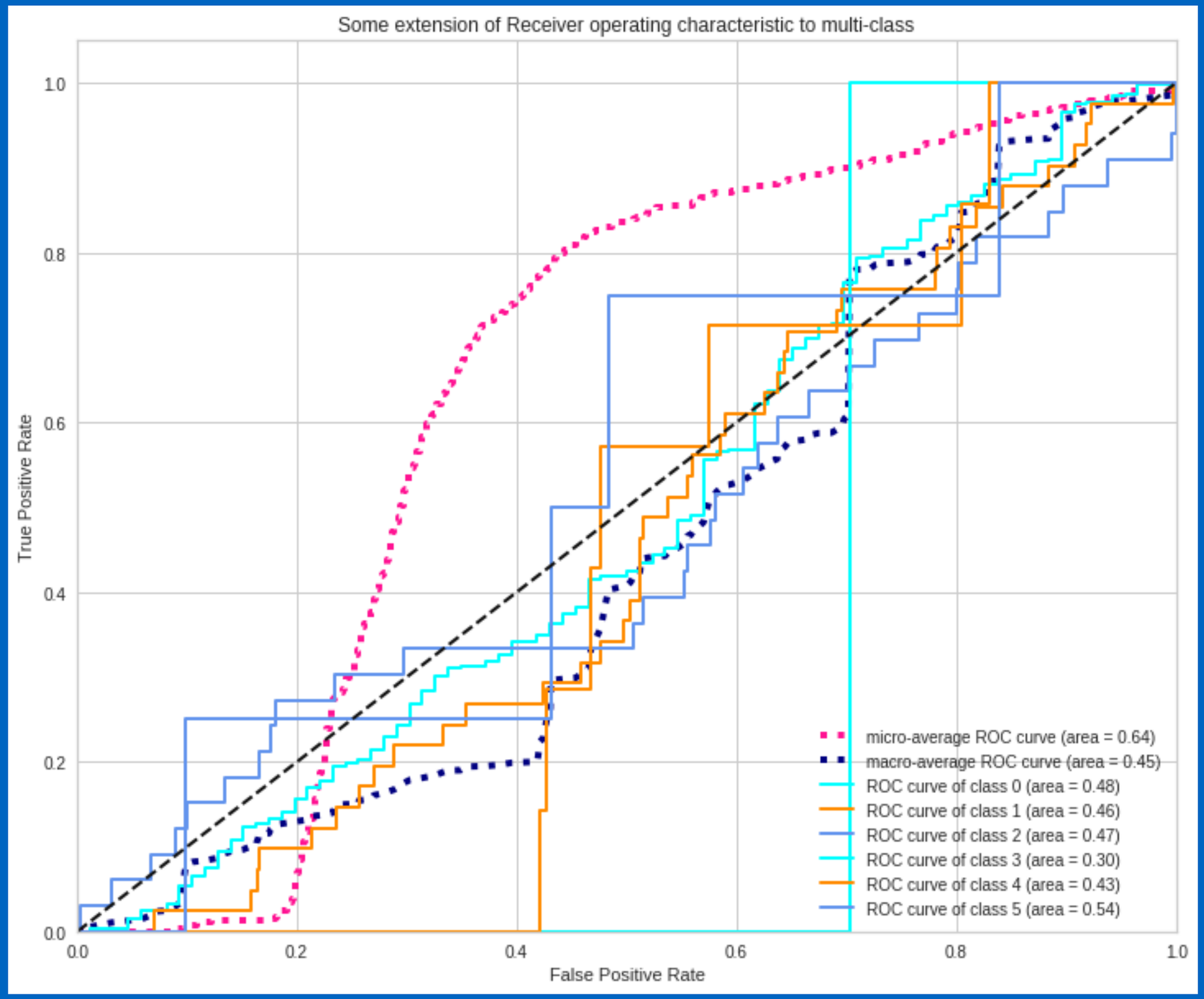
Part 2/5
Room for Improvement.

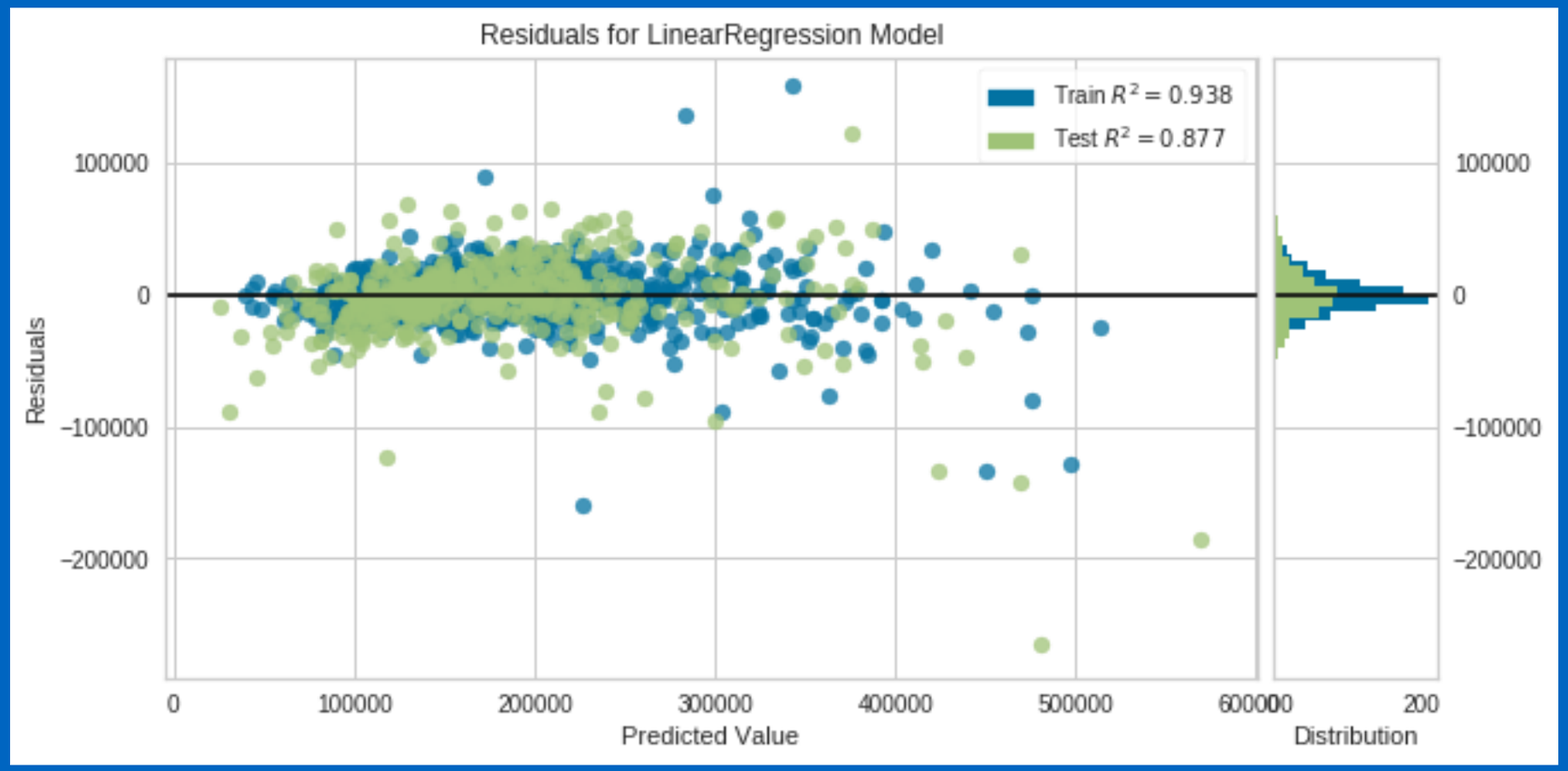

Confusion matrix, Train Data
Normal 32 0 0 0 36 0
Partial 0 3 0 0 0 0
Abnorml 0 0 7 0 1 0
AdjLand 0 0 0 2 11 0
Family 4 0 0 0 798 0
Alloca 0 0 0 0 0 84
True label
Normal Partial Abnorml AdjLand Family Alloca
Predicted label
0 100 200 300 400 500 600 700

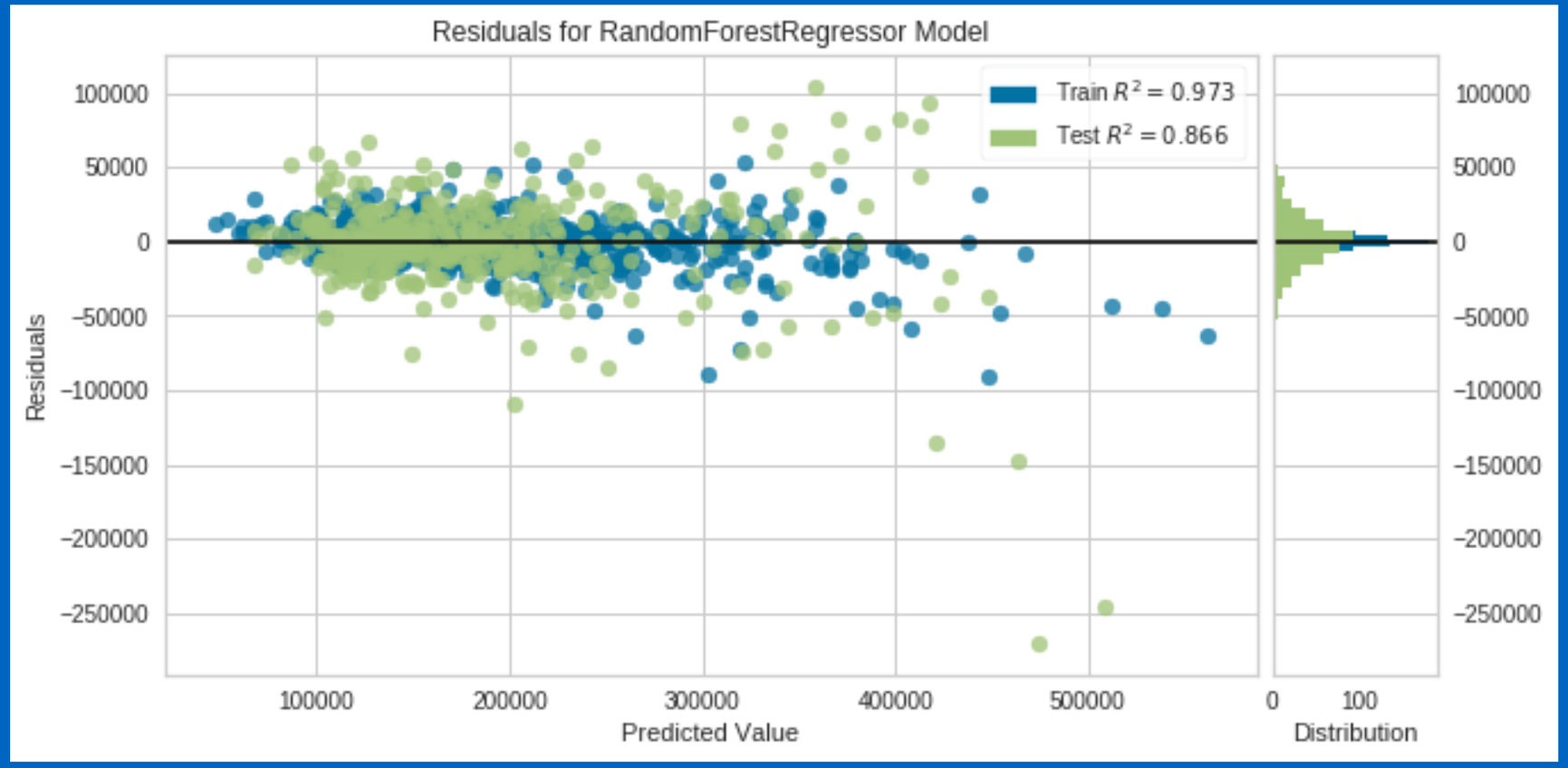

Confusion matrix, Test Data
Normal 3 0 0 0 30 0
Partial 0 0 0 0 1 0
Abnorml 1 0 0 0 3 0
AdjLand 0 0 0 0 7 0
Family 8 0 0 0 388 0
Alloca 2 0 0 0 3 36
True label
Normal Partial Abnorml AdjLand Family Alloca
Predicted label
0 50 100 150 200 250 300 350


Normalised Confusion matrix, Train Data
Normal 0.47 0.00 0.00 0.00 0.53 0.00
Partial 0.00 1.00 0.00 0.00 0.00 0.00
Abnorml 0.00 0.00 0.88 0.00 0.12 0.00
AdjLand 0.00 0.00 0.00 0.15 0.85 0.00
Family 0.00 0.00 0.00 0.00 1.00 0.00
Alloca 0.00 0.00 0.00 0.00 0.00 1.00
True label
Normal Partial Abnorml AdjLand Family Alloca
Predicted label
0.0 0.2 0.4 0.6 0.8 1.0


Normalised Confusion matrix, Test Data
Normal 0.09 0.00 0.00 0.00 0.91 0.00
Partial 0.00 0.00 0.00 0.00 1.00 0.00
Abnorml 0.25 0.00 0.00 0.00 0.75 0.00
AdjLand 0.00 0.00 0.00 0.00 1.00 0.00
Family 0.02 0.00 0.00 0.00 0.98 0.00
Alloca 0.05 0.00 0.00 0.00 0.07 0.88
True label
Normal Partial Abnorml AdjLand Family Alloca
Predicted label
0.0 0.2 0.4 0.6 0.8 1.0

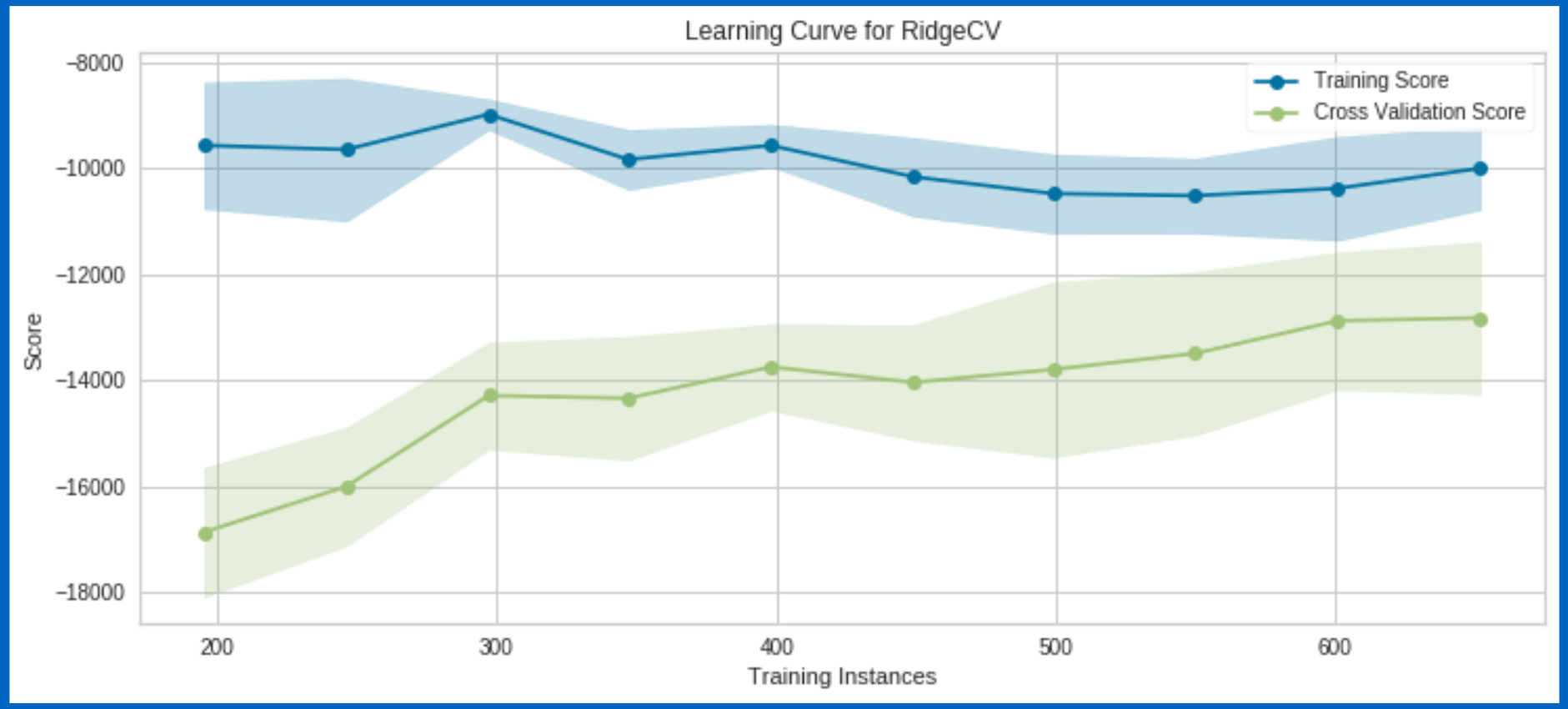

Some extension of Receiver operating characteristic to multi-class
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
micro-average ROC curve (area = 0.64)
macro-average ROC curve (area = 0.45)
ROC curve of class 0 (area = 0.48)
ROC curve of class 1 (area = 0.46)
ROC curve of class 2 (area = 0.47)
ROC curve of class 3 (area = 0.30)
ROC curve of class 4 (area = 0.43)
ROC curve of class 5 (area = 0.54)

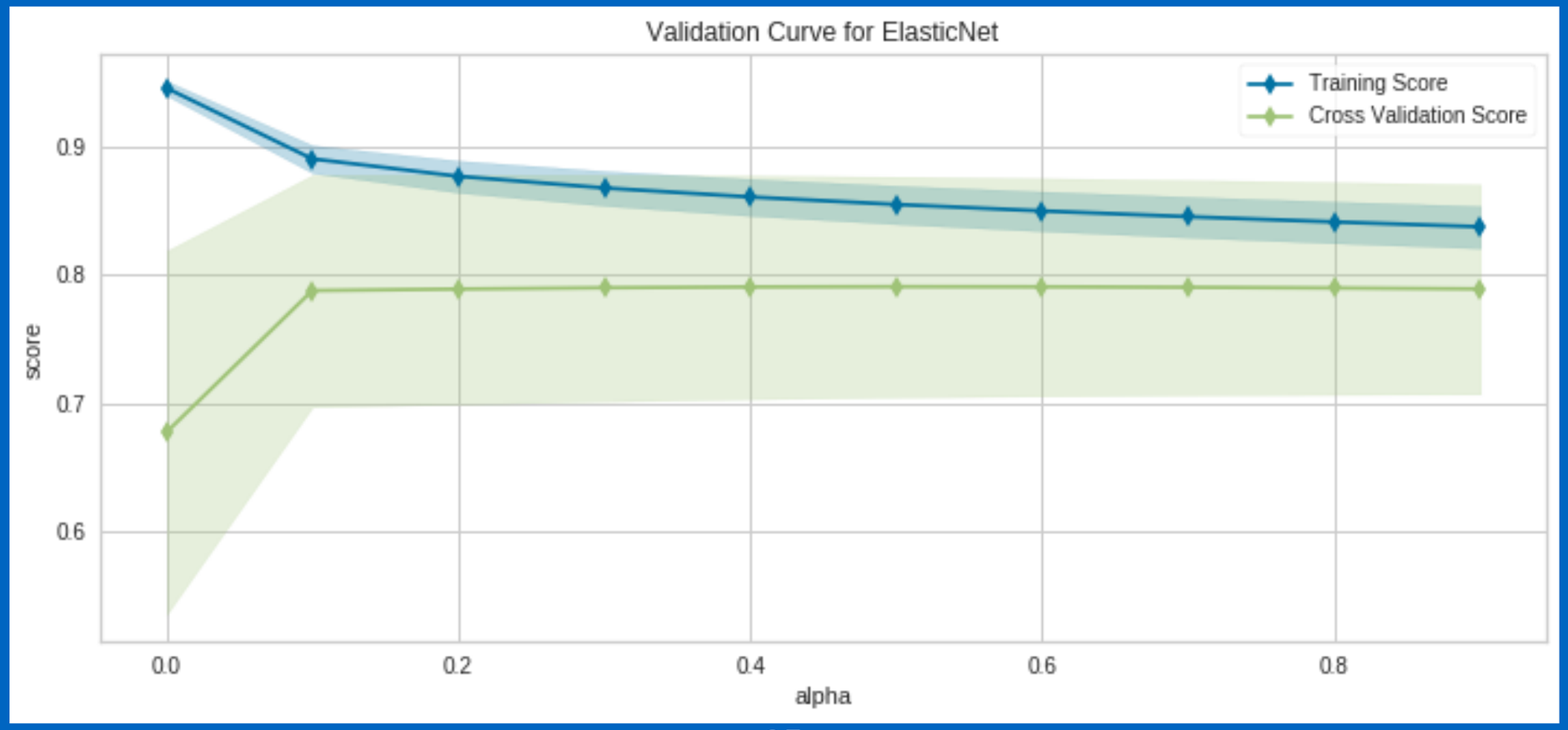

Residuals for LinearRegression Model
Train $R^2 = 0.938$
Test $R^2 = 0.877$
Residuals
Predicted Value
Distribution


Residuals for RandomForestRegressor Model
Train $R^2 = 0.973$
Test $R^2 = 0.866$
Residuals
Predicted Value
Distribution

Part 4/5

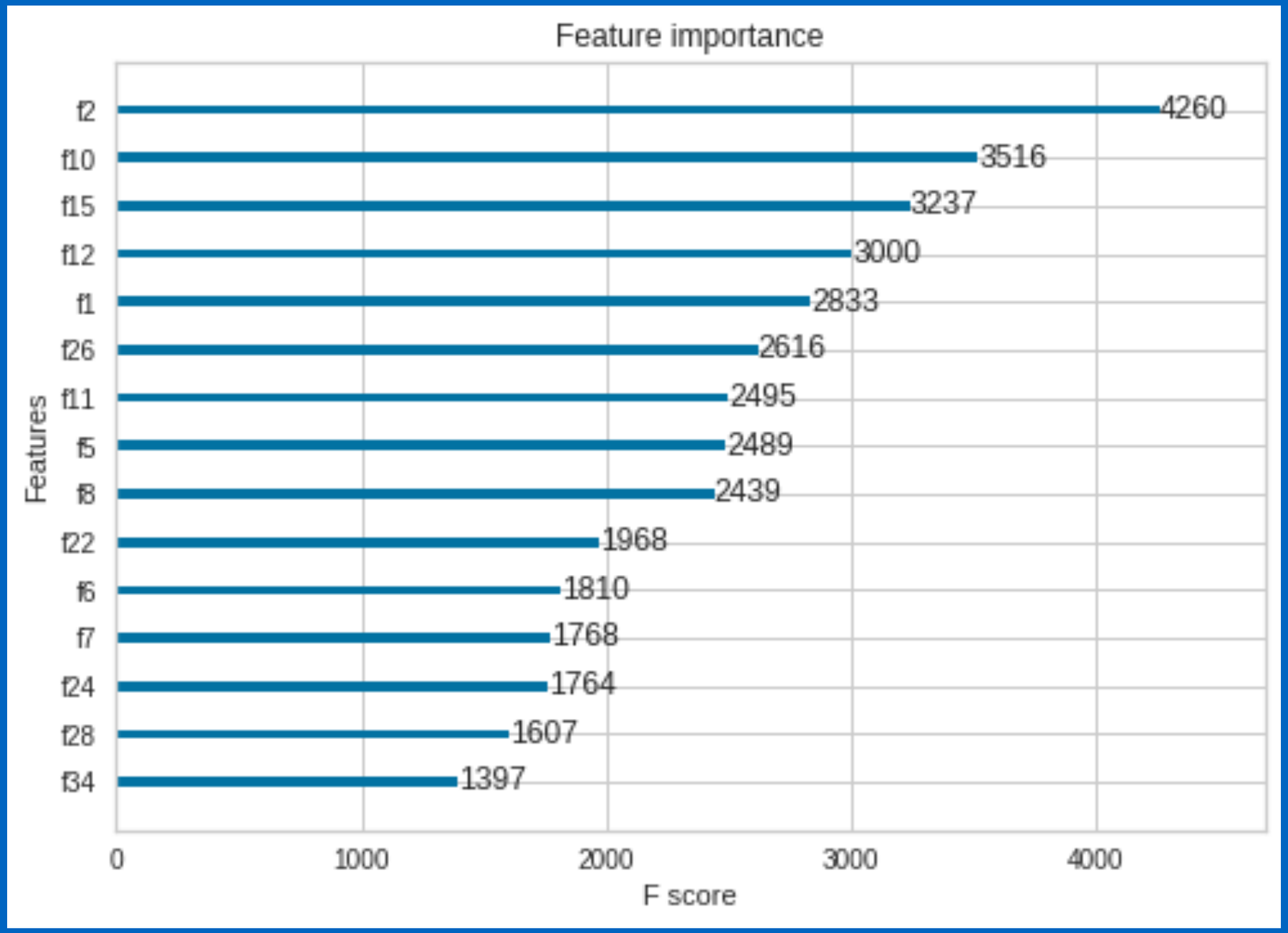
ML is so easy. I am getting 99.5% accuracy. Wait.. something is wrong !


Learning Curve for RidgeCV
Training Score
Cross Validation Score
−8000
−10000
−12000
−14000
−16000
−18000
Score
200
300
400
500
600
Training Instances


Validation Curve for ElasticNet
Training Score
Cross Validation Score
0.9
0.8
0.7
0.6
score
0.0
0.2
0.4
0.6
0.8
alpha

Part 5/5

Explain me what you learnt !


Feature importance
f2
f10
f15
f12
f1
f26
f11
f5
f8
f22
f6
f7
f24
f28
f34
4260
3516
3237
3000
2833
2616
2495
2489
2439
1968
1810
1768
1764
1607
1397
Features
0
1000
2000
3000
4000
F score