The Last Time I Teach You

●●●

At Least Formally, In Class

Today’s Lecture

Final Project Chatter

Guest Speakers

End-to-End Data Science Project Recap … to the tune of …

Steve’s Presentation

GANs!

Final Notes, Resources

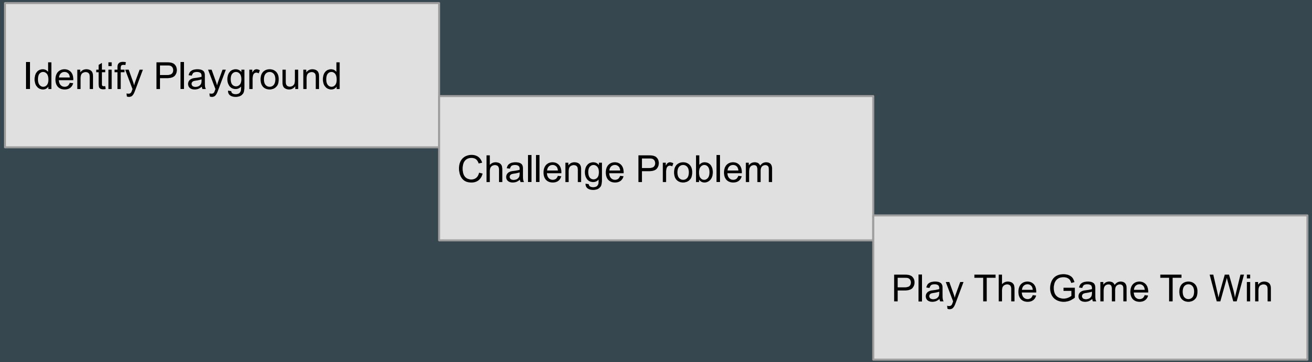
5 Minutes

What did you get out of this class? 3 Credits?

What do I owe you? Probably not money but...

Final Project

Questions, Format, Checkpoints



Identify Playground

Challenge Problem

Play The Game To Win

Project Tips

PTS? TBD.	Low	Average	High	X Factor
Coding	Little to no data wrangling, little to no feature manipulation	Some derivative data created, wrangling attempted and succeeded, some feature generation	Several creative feature generation techniques used, appropriate scaling and manipulation	Did you have to? Comparison? Roadblocks conquered?
Math / Algorithms	No algorithms or minimal out of the box data flow	Good use of an algorithm, some tuning	Baseline comparisons, algorithm tuning, appropriate usage of pipeline	Did you use these ‘correctly’ and ‘interestingly’ or just use them to use them?
Analysis	Tell us about the above	Opinions and insights on why you used this algorithm and what it is meant to do	You really want to win the game you are playing and can fully explain why you are succeeding or failing	Can you back up hypothesis with results? Did you try everything?
X Factor	Basic plots of EDA, no visualization of algorithms	The right plots, appropriate and informative information within, telling the story with visuals	Incredibly insightful and visually appealing plots, comparison plots, lecture slides-esque	Comparison? Take away story? Was there a surprise?

End to End Data Science Project

A company which is active in Big Data and Data Science wants to hire data scientists among people who successfully pass some courses which conduct by the company. **Many people signup for their training.** Company wants to know which of these candidates are really wants to work for the company after training or looking for a new employment because it helps to reduce the cost and time as well as the quality of training or planning the courses and categorization of candidates. Information related to demographics, education, experience are in hands from candidates signup and enrollment.

This dataset designed to understand the factors that lead a person to leave current job for HR researches too. By model(s) that uses the current credentials,demographics,experience data you will predict the probability of a candidate to look for a new job or will work for the company, as well as interpreting affected factors on employee decision.

Presenting The Analysis

Steve Schmidt

Stealing Data Scientists

Data Details

Features & Label

FEATURES

enrollee_id : Unique ID for candidate

city: City code

city_ development _index : Development index of the city (scaled)

gender: Gender of candidate

relevent_experience: Relevant experience of candidate

enrolled_university: Type of University course enrolled if any

education_level: Education level of candidate

major_discipline :Education major discipline of candidate

experience: Candidate total experience in years

company_size: No of employees in current employer's company

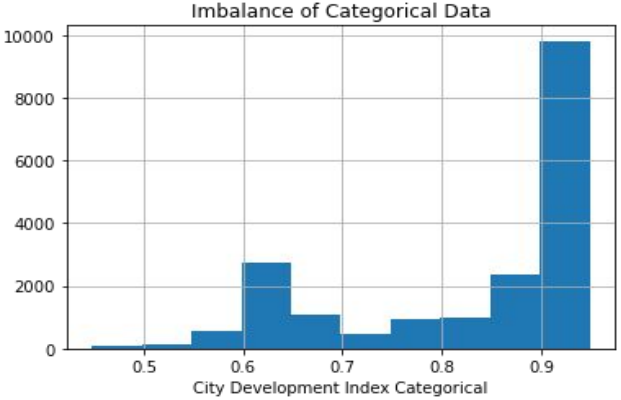
company_type : Type of current employer

lastnewjob: Difference in years between previous job and current job

training_hours: training hours completed

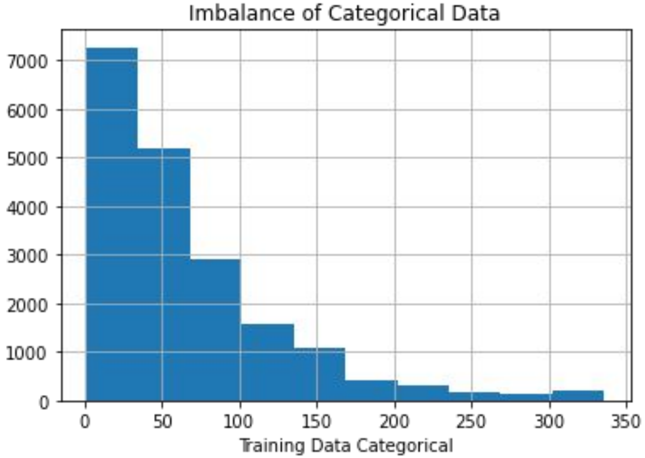
target: 0 – Not looking for job change, 1 – Looking for a job change

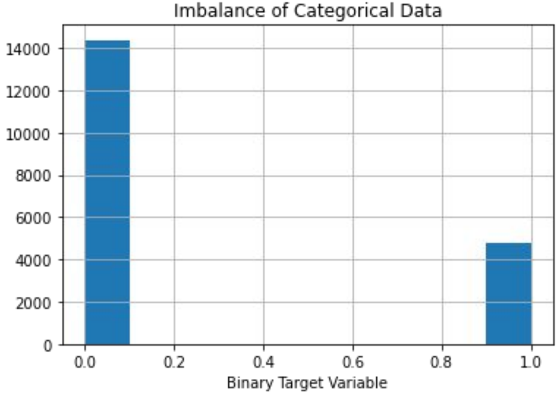
Exploratory Data Analysis



The numeric data left for us looks like this…

As noted, the imbalance in ‘yes’ responses can be seen below



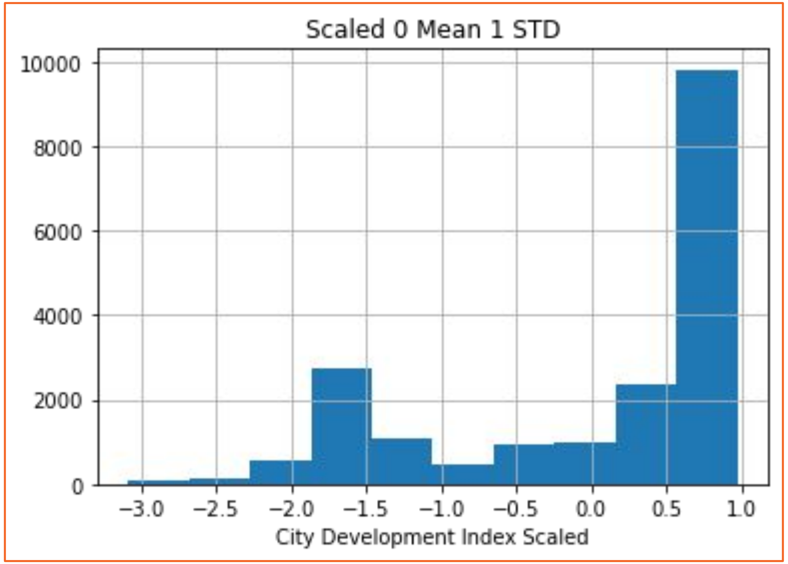


EDA for ML

City Development feature was already normalized so let’s try and scale it

After dealing with categorical data, attention goes to numeric data

```
Binned Traing Hours
(19.0, 37.0]      3949
(0.999, 19.0]    3849
(101.0, 336.0]  3830
(58.0, 101.0]    3772
(37.0, 58.0]      3758
```



‘qcut’ based off the eye test seems to give us decent binning and what we might expect from a training session

Machine Learning Plan

Outline, Pipeline, Baseline

Encode Categorical Variables

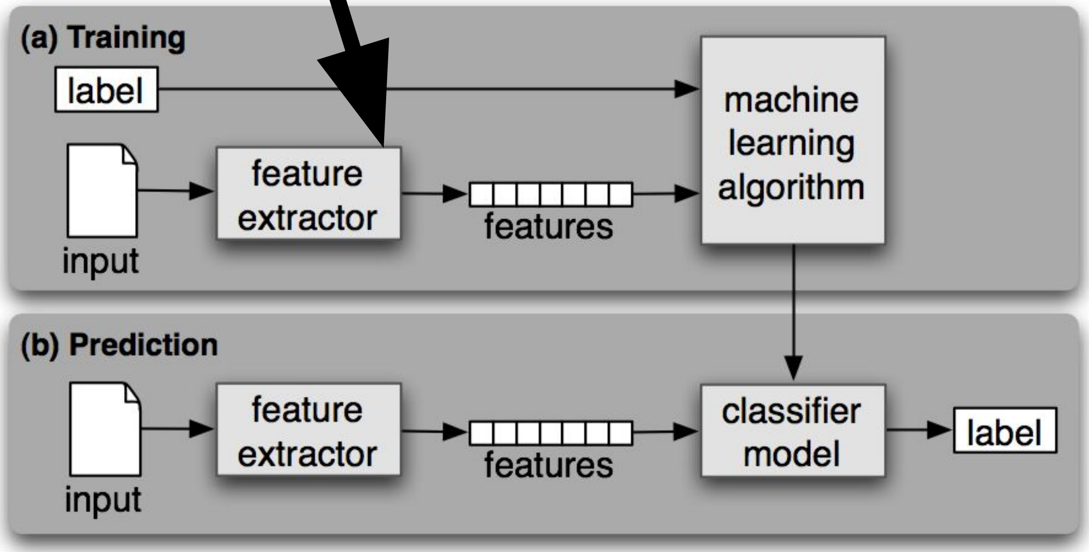
Determine ‘Expert Features’

Attempt Logistic Regression

Compare to Decision Trees

Compare to ...

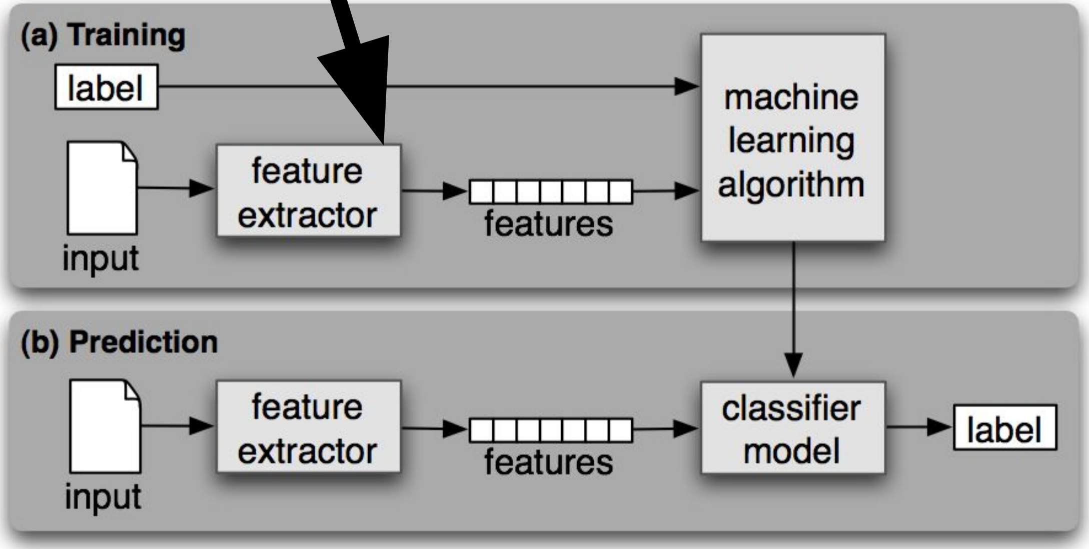
Baseline data - One Hot Encoding of Categorical Variables



Baseline Formulation & Results

Expert Formulation & Results

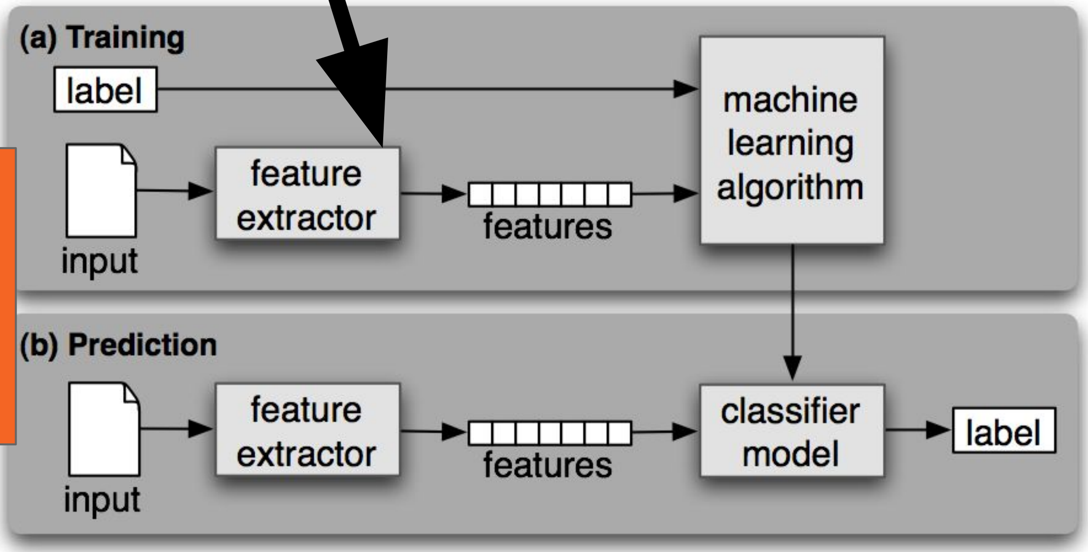
1) Scale & Bin Imbalanced Data
2) One Hot Encoding of Categorical Variables



Balanced Expert Formulation & Results

1) Scale & Bin Imbalanced Data
2) One Hot Encoding of Categorical Variables

1) Sample lightly from ‘0’ targets
2) Sample highly from ‘1’ targets

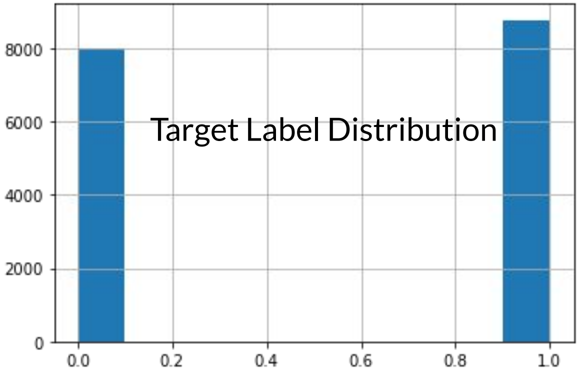


Logistic Regression - 77%
Decision Tree - 81%

Results Comparison

Decision Tree

Balanced Expert Data


Target Label Distribution

```
extra_1 =
data_expert_formatted[data_expert_formatted.target == 1]

extra_0 =
data_expert_formatted[data_expert_formatted.target == 0]

data_balanced = extra_0.sample(8000)

data_balanced = pd.concat([data_balanced, extra_1])

data_balanced = pd.concat([data_balanced,
extra_1.sample(4000)])
```

~ 83% accuracy on some runs...

Neural Network

Expert Data

```
  (fc1): Linear(in_features=192, out_features=100, bias=True)

  (fc2): Linear(in_features=100, out_features=100, bias=True)

  (fc3): Linear(in_features=100, out_features=50, bias=True)

  (fc4): Linear(in_features=50, out_features=1, bias=True)

  (dropout): Dropout(p=0.3, inplace=False)
```

Quickly gets to 79%...but this simple Network does no better:(

Human Learning...

- Encoding was necessary but not sufficient for learning
- Inconsistencies within imbalance features may have been the difference
- Was not much of an algorithmic trade off (after Logistic Regression)
- Bias may be inherent to this dataset due to data collection
    - Who responds to a survey anyway?
    - Who doesn’t want ‘free’ training?

Grading My Presentation

PTS? TBD.	Low	Average	High	X Factor
Coding	Little to no data wrangling, little to no feature manipulation	Some derivative data created, wrangling attempted and succeeded, some feature generation	Several creative feature generation techniques used, appropriate scaling and manipulation	Did you have to? Comparison? Roadblocks conquered?
Math / Algorithms	No algorithms or minimal out of the box data flow	Good use of an algorithm, some tuning	Baseline comparisons, algorithm tuning, appropriate usage of pipeline	Did you use these ‘correctly’ and ‘interestingly’ or just use them to use them?
Analysis	Tell us about the above	Opinions and insights on why you used this algorithm and what it is meant to do	You really want to win the game you are playing and can fully explain why you are succeeding or failing	Can you back up hypothesis with results? Did you try everything?
X Factor	Basic plots of EDA, no visualization of algorithms	The right plots, appropriate and informative information within, telling the story with visuals	Incredibly insightful and visually appealing plots, comparison plots, lecture slides-esque	Comparison? Take away story? Was there a surprise?