BUAL5610/6610/6616-Predictive Modeling II

- **Instructor**: Dr. Pei Xu *(Pronunciation: |Pay| |She|)*
- E-mail: pzx0002@auburn.edu
- Office: Lowder 420

Syllabus

- **Class Time**
  - Section 001: TR 8:00 – 9:15 am @ *Lowder 20*
  - Section 002: TR 9:30 – 10:45 pm @ *Lowder 27*

- **Office Hours**: TR 1 – 2 pm; others by appt.
- **Graduate Assistant:** Lingxiao Wang (lzw0039@ auburn.edu)

Syllabus (cont.)

***Canvas*:** Lecture slides and additional reading materials will be given weekly through Canvas. Course announcements will also be published via Canvas. It is extremely important to check Canvas and your email regularly to stay informed about the class.

Syllabus (cont.)

- **Textbooks**
  - James, Witten, Hastie, and Tibshirani. "*An Introduction to Statistical Learning*. (Springer; 7th printing 2017)" ISBN-13: 978-1461471370; ISBN-10: 1461471370
  - Tan, Pang-Ning, Steinbach, Michael, Kumar, Vipin, Karpatne, Anuj. "*Introduction to Data Mining*", (Pearson, 2nd edition, 2018) ISBN-13: 9780133128901; ISBN-10: 0133128903
  - Wes McKinney. "*Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*" (2nd Edition). 2017. ISBN-13: 978-1491957660.

Syllabus (cont.)

- **SOFTWARE**

  - **Python 3**: https://www.anaconda.com/download/
    Anaconda is the most popular Python data science platform. It aims to simplify package management and deployment with a collection of over 1500 open source packages.

Syllabus (cont.)

GRADING

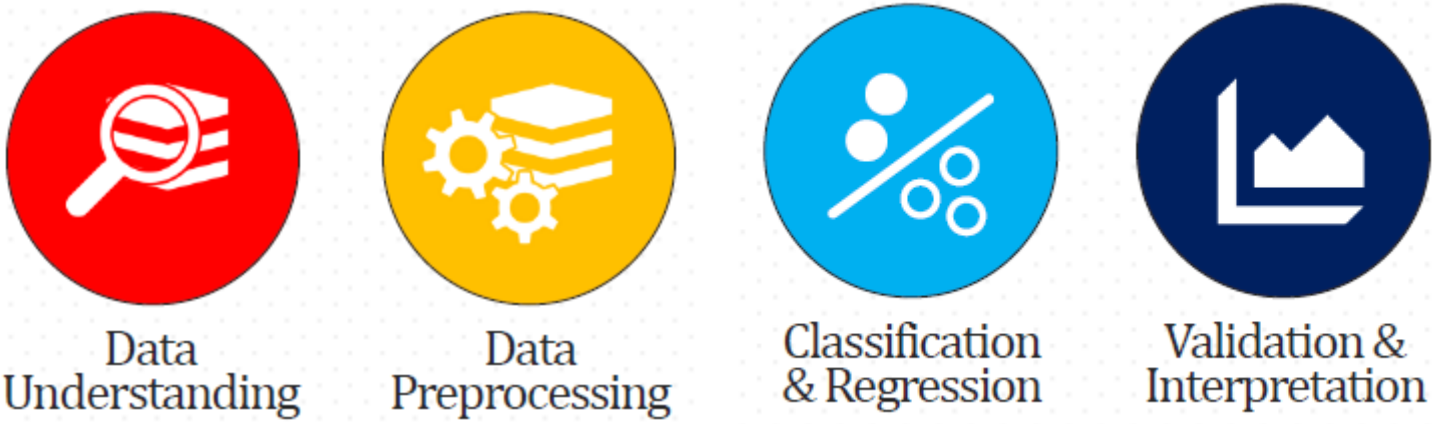
The grading policy for this course is point-based. Points can be earned from the following tasks.

1)	Homework Assignments	100 points	400 points in total
2)	In-class work/quizzes	50 points	
3)	Final Project	50 points	
4)	Take-Home Exam 1	100 points	
5)	Take-Home Exam 2	100 points	
6)	Attendance bonus (on-campus only)	≈ 10 points	

The numerical points you earned will be converted to letter grades as follows:

A:	>= 360	(90%)
B:	320– 359	(80%)
C:	280– 319	(70%)
D:	240– 279	(60%)
F:	0–239	

Course Topics



Syllabus (cont.)

Week	Date	Topics
1	Jan 9	Syllabus; Introduction to Business Analytics
2	Jan 14, 16	Introduction to Predictive Modelling; Introduction to Python
3	Jan 21, 23	Introduction to Pandas; Data Understanding
4	Jan 28, 30	Assessing Model Accuracy: Bias-Variance tradeoff
5	Feb 4, 6	Resampling Methods: Cross Validation & Bootstrap
6	Feb 11, 13	Tree-based Methods
7	Feb 18, 20	Ensemble Methods: Bagging & Boosting; Random Forest
8	Feb 25, 27	**Take home Exam 1**
9	Mar 3, 5	Nearest Neighbor; Naïve Bayes
10	**Mar 9-13**	**Spring Break**
11	Mar 17, 19	Support Vector Machines
12	Mar 24, 26	Neural Networks **[Decide a topic for final project]**
13	Mar 31, Apr 2	Model Comparison
14	Apr 7, 9	Sentiment Analysis
15	Apr 14, 16	Topic Mining
16	Apr 21, 23	Special Topic: The Frontiers of Predictive Modeling
17	**Apr 28-30**	**Take home Exam 2**

AUBURN
UNIVERSITY

Introduction to Business Analytics

Dr. Pei Xu
Auburn University
Tuesday, January 7, 2020

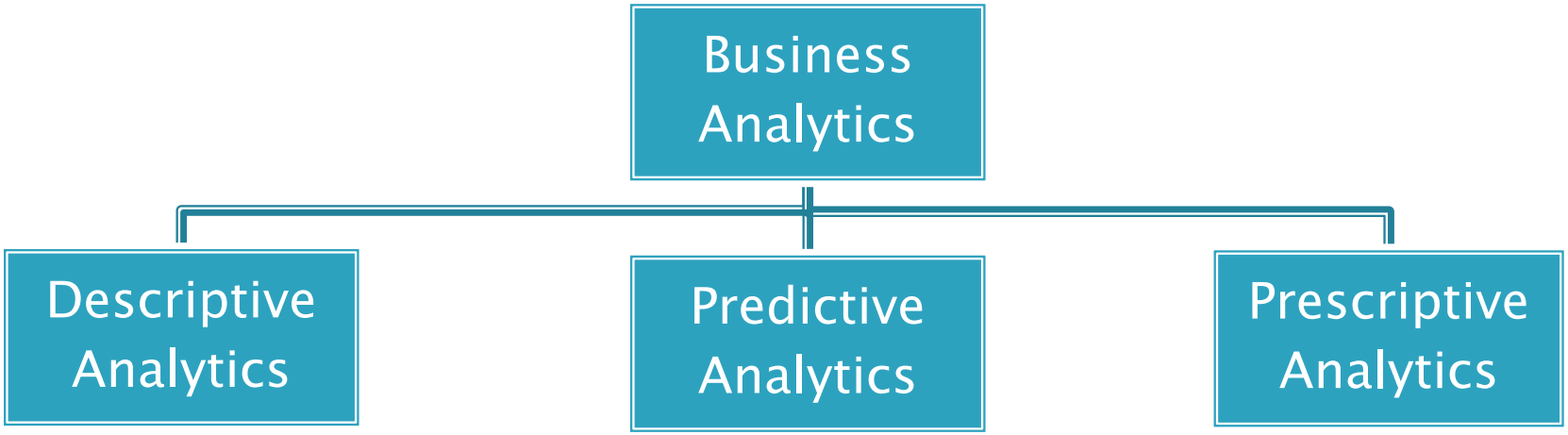
Content

- Definition of Business Analytics
- Unsupervised vs. Supervised learning
- The nature of this course: project-oriented

What is Business Analytics?

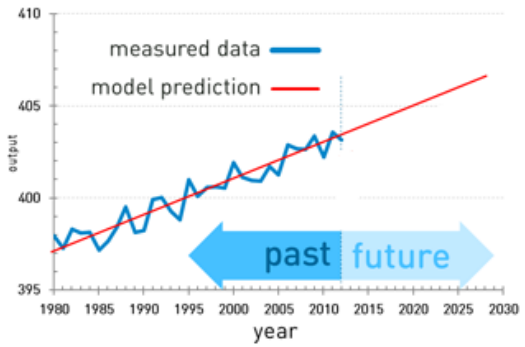
- The use of data analysis and computer technology to make better business decisions
  - Discover unknown unknowns in data
  - obtain actionable insight
  - communicate business data stories

Scope of Business Analytics

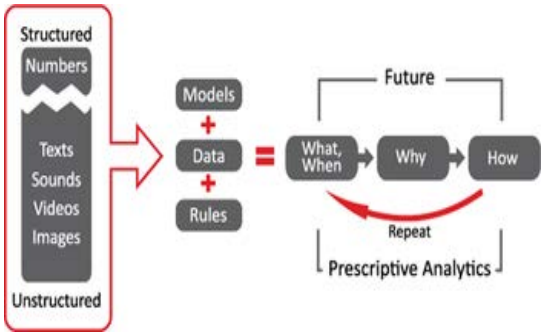


- Dashboard
- Graphs/Charts/Visual Display

- Models to forecast/predict



- Guidelines for optimal decision



Scope of Business Analytics

- Descriptive analytics
- Predictive analytics
- Prescriptive analytics

- Descriptive analytics
  - uses data to understand past and present
  - Summarize data into meaningful charts and reports
  - Typical questions:

    How much did we sell in each region?

    What was our revenue and profit last quarter?

    Visualization of Data Science Degree

Scope of Business Analytics

- Descriptive analytics
- Predictive analytics
- Prescriptive analytics

- Predictive analytics
  - analyzes past performance in an effort to predict the future
  - Detecting patterns or relationships, and then extrapolating these relationships forward in time
  - Typical questions:

    What will happen if demand falls by 10% or if supplier prices go up 5%?

    What do we expect to pay for fuel over the next several months?

    Target Pregnancy Prediction Program

Scope of Business Analytics

- Descriptive analytics
- Predictive analytics
- Prescriptive analytics

- Prescriptive analytics
  - uses optimization techniques to identify the best alternatives to minimize or maximize some objective
  - Typical question:

    What is the best way of shipping goods from our factories to minimize costs

Scope of Business Analytics

Example 1: Retail Markdown Decisions

Most department stores clear seasonal inventory by reducing prices. The key question is:

When to reduce the price and by how much, to meet inventory goal and maximize revenue?

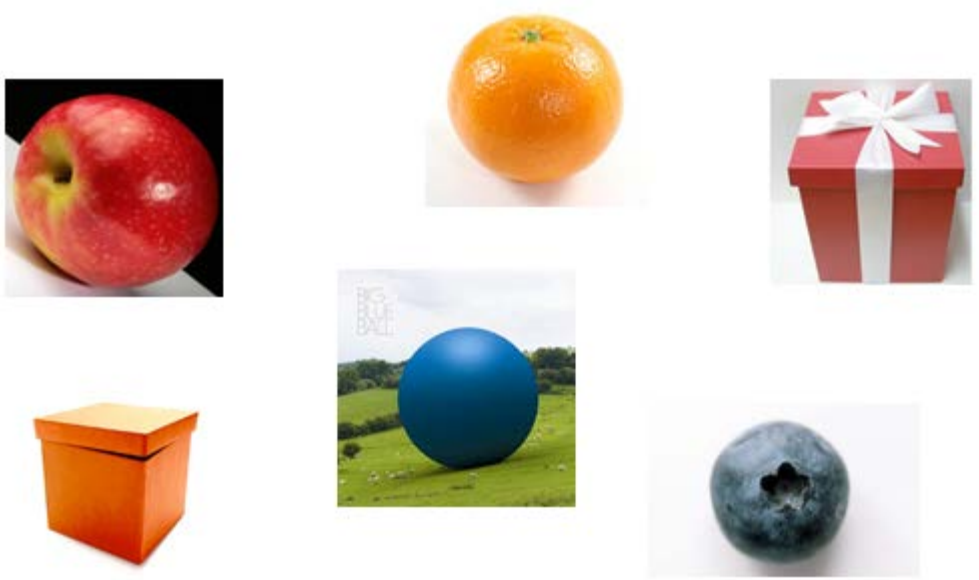
Descriptive analytics: examine historical data for similar products (prices, units sold, advertising, …)

Predictive analytics: predict sales based on pricing decisions

Prescriptive analytics: find the best sets of pricing and advertising to maximize sales revenue

Unsupervised and Supervised learning (1)

- **Unsupervised Learning**
  - The model is not provided with the correct results during the training.
    - Different Types of Clustering



How many clusters do you expect?

Unsupervised and Supervised learning (2)

- **Supervised Learning**
  - Training data includes both the input and the desired results.
  - For some examples the correct results (targets) are known and are given in input to the model during the learning process.
    - Neural Networks
    - Decision Trees

Unsupervised and Supervised learning (3)

- **Supervised Learning**

  Divide the whole dataset to two parts:

  - A **training set** is a set of data used to discover potentially predictive relationships.
  - A **test set** is a set of data used to assess the strength and utility of a predictive relationship.

Unsupervised and Supervised learning (4)

- **Supervised Learning**

SampleRNN trained on all 32 of Beethoven's piano sonatas.
https://soundcloud.com/samplernn/samplernn-music-1

How a Japanese cucumber farmer is using deep learning
https://cloud.google.com/blog/big-data/2016/08/how-a-japanese-cucumber-farmer-is-using-deep-learning-and-tensorflow

Visualizing a Self-Driving Future
https://www.youtube.com/watch?v=HJ58dbd5g8g

9 Applications of Machine Learning from Day-to-Day Life
https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

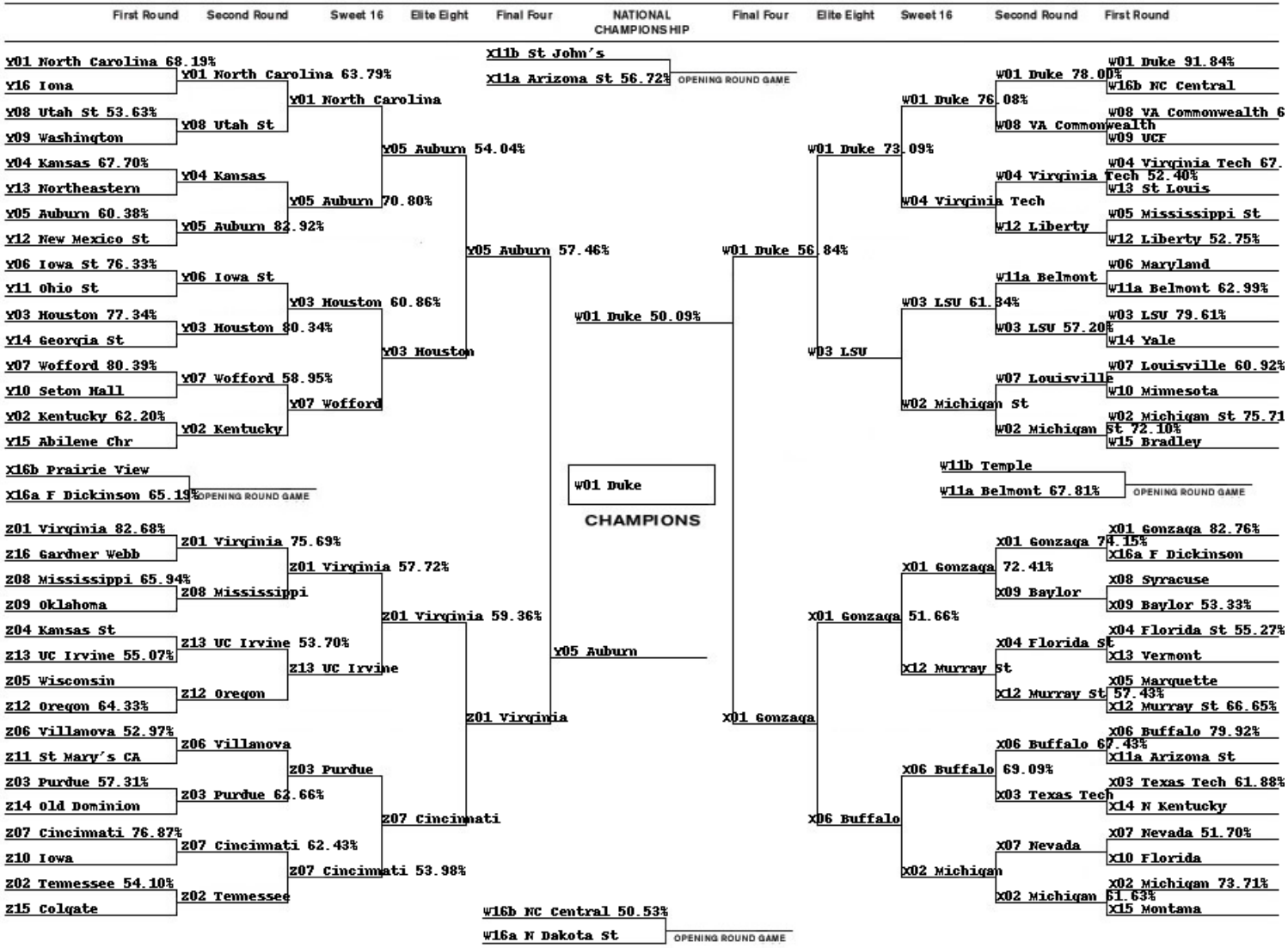
What is this course about?

- Know-how: project oriented 
- Model Selection 
- Result Interpretation 
- Business Implication 

- In-depth discussion of Algorithm 

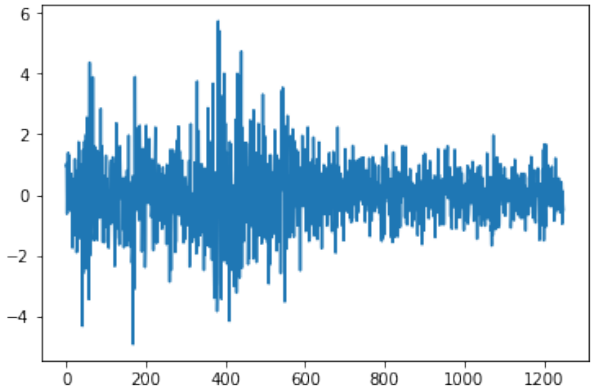
Sample Datasets

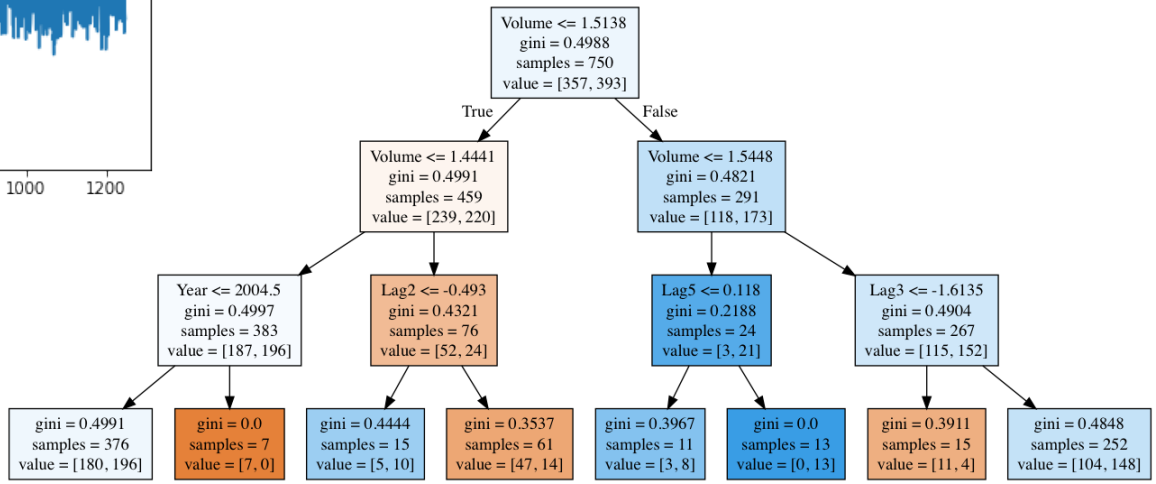
Sport Analytics: NCAA Men's Basketball



Stock Market Analytics

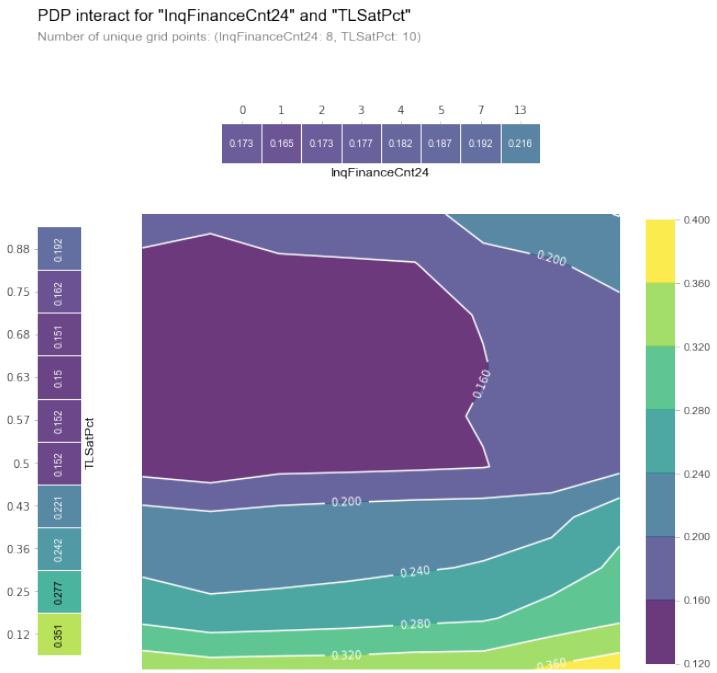
- This data set consists of percentage returns for the S&P 500 stock index over 1,250 days, from the beginning of 2001 until the end of 2005.





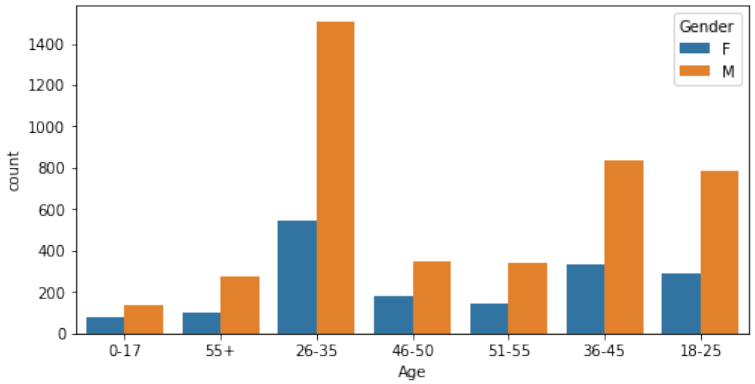
Credit Default

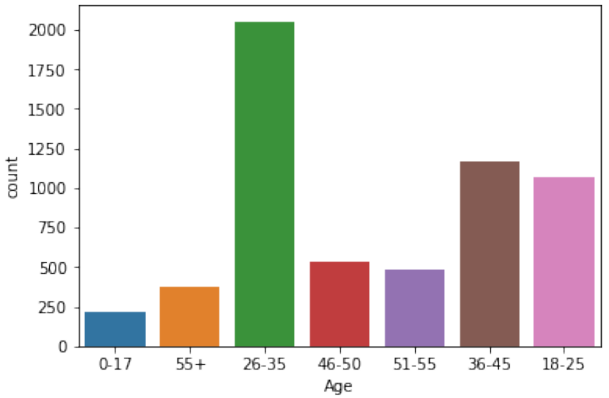
https://www.kaggle.com/c/GiveMeSomeCredit

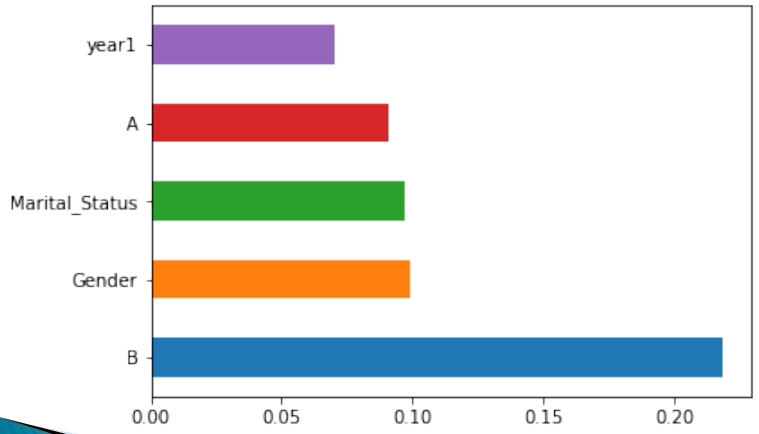


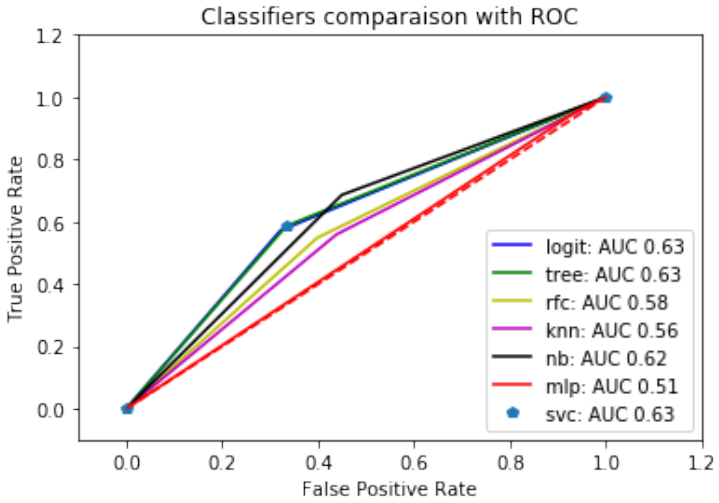
Variable	Type	Len	Label
BanruptcyInd	Num	8	Bankruptcy Indicator
CollectCnt	Num	8	Number Collections
DerogCnt	Num	8	Number Public Derogatories
InqCnt06	Num	8	Number Inquiries 6 Months
InqFinanceCnt24	Num	8	Number Finance Inquires 24 Months
InqTimeLast	Num	8	Time Since Last Inquiry
TARGET	Num	8	0 = Paid off; 1= Bad debt
TL50UtilCnt	Num	8	Number Trade Lines 50 pct Utilized
TL75UtilCnt	Num	8	Number Trade Lines 75 pct Utilized
TLBadCnt24	Num	8	Number Trade Lines Bad Debt 24 Months
TLBadDerogCnt	Num	8	Number Bad Dept plus Public Derogatories
TLBalHCPct	Num	8	Percent Trade Line Balance to High Credit
TLCnt	Num	8	Total Open Trade Lines
TLCnt03	Num	8	Number Trade Lines Opened 3 Months
TLCnt12	Num	8	Number Trade Lines Opened 12 Months
TLCnt24	Num	8	Number Trade Lines Opened 24 Months
TLDel3060Cnt24	Num	8	Number Trade Lines 30 or 60 Days 24 Months
TLDel60Cnt	Num	8	Number Trade Lines Currently 60 Days or Worse
TLDel60Cnt24	Num	8	Number Trade Lines 60 Days or Worse 24 Months
TLDel60CntAll	Num	8	Number Trade Lines 60 Days or Worse Ever
TLDel90Cnt24	Num	8	Number Trade Lines 90+ 24 Months
TLMaxSum	Num	8	Total High Credit All Trade Lines
TLOpen24Pct	Num	8	Percent Trade Lines Open 24 Months
TLOpenPct	Num	8	Percent Trade Lines Open
TLSatCnt	Num	8	Number Trade Lines Currently Satisfactory
TLSatPct	Num	8	Percent Satisfactory to Total Trade Lines
TLSum	Num	8	Total Balance All Trade Lines
TLTimeFirst	Num	8	Time Since First Trade Line
TLTimeLast	Num	8	Time Since Last Trade Line

Black Friday purchases





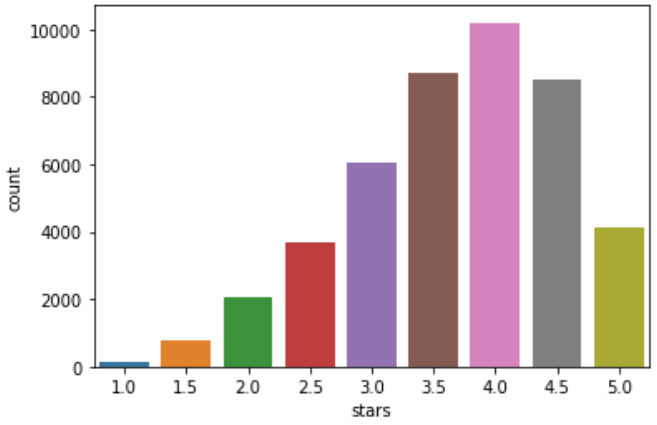




Yelp Review Ratings

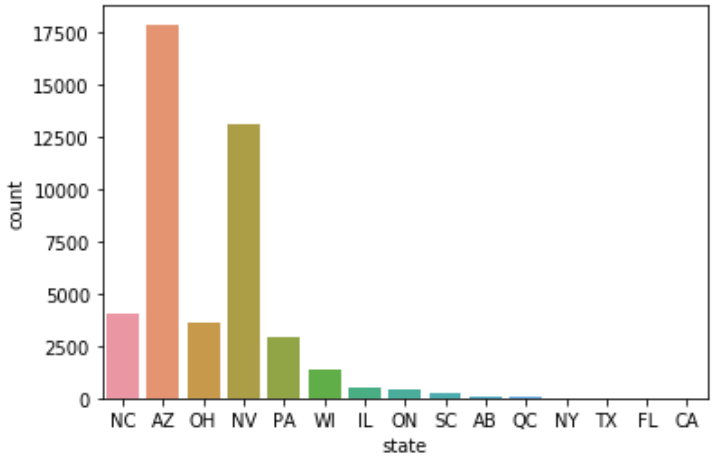
```
1 #number of each star rating in data
2 sns.countplot(rating2['stars'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x28e2b994320>
```



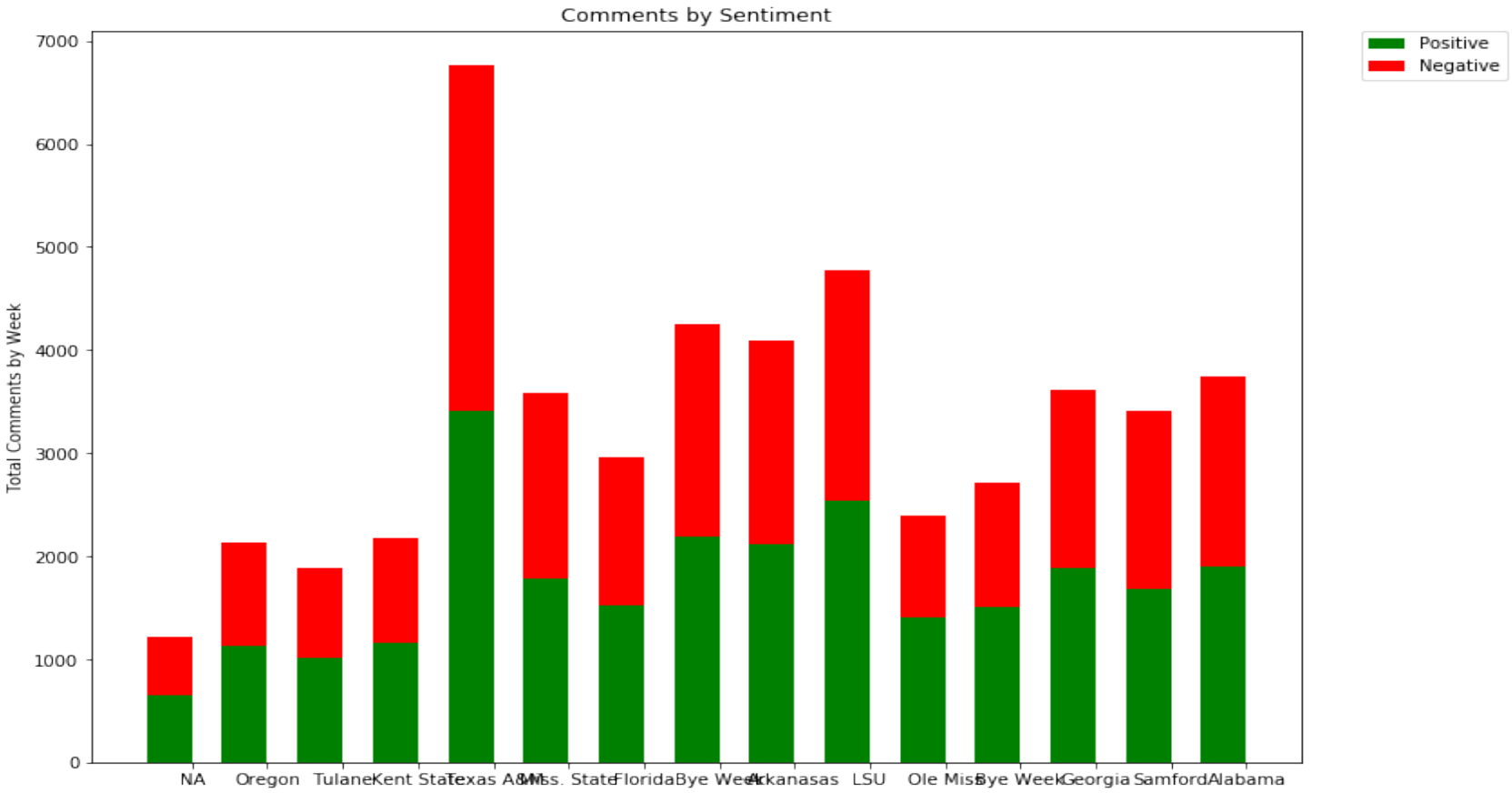

```
1 #number of business in data set for each state
2 sns.countplot(rating2['state'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x28e295b94e0>
```



business_id	city	is_open	latitude	longitude	name	review_count	stars	...	NC	NV	NY	OH	ON	PA	QC	SC	TX	WI
gnKjwL_1w79qoiV3IC_xQQ	Charlotte	1	35.092564	-80.859132	Musashi Japanese Restaurant	170	4.0	...	1	0	0	0	0	0	0	0	0	0
1Dfx3zM-rW4n-31KeC8sJg	Phoenix	1	33.495194	-112.028588	Taco Bell	18	3.0	...	0	0	0	0	0	0	0	0	0	0
fweCYi8FmbJXHCqLnwuk8w	Mentor-on-the-Lake	1	41.708520	-81.359556	Marco's Pizza	16	4.0	...	0	0	0	1	0	0	0	0	0	0
nh_kQ16QAoXWwqZ05MPfBQ	Las Vegas	1	36.116549	-115.088115	Myron Hensel Photography	21	5.0	...	0	1	0	0	0	0	0	0	0	0
KWywu2tTEPWmR9JnBc0WyQ	Las Vegas	1	36.080168	-115.182756	Hunk Mansion	107	4.0	...	0	1	0	0	0	0	0	0	0	0

Customized Twitter Sentiment



https://docs.google.com/spreadsheets/d/1vV61g7eiYjXINTj0gqtbDHA8m5knEcdsZrQHkFU-8Ro/edit?usp=sharing