

Portfolio Milestone

MS Applied Data Science

- Ryan Timbrook
- NetID: RTIMBROO
- Term: January/Winter, 2020

Portfolio - Course - Alignment Overview

Course Projects included in Portfolio:

- IST 736 Text Mining
- IST 718 Big Data Analytics
- IST 707 Data Analytics
- IST 659 Database Administration Concepts and Database Management
- MBC 638 Data Analysis and Decision Making
- SCM 651 Business Analytics

Course Projects not included in Portfolio:

- ACC 652 Accounting Analytics
- MAR 653 Marketing Analytics
- IST 687 Introduction to Data Science
- IST 722 Data Warehouse
- IST 623 Intro to Information Security
- IST 600 Selected Topics – Transferred Credits – LING 570 Shallow Proc. for NLP
- CIS 668 Natural Language Processing – Transferred Credits – LING 571 Deep Proc. For NLP

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Project - Course - Alignment Overview

- IST 736 Text Mining
 - Public Sentiment Toward NFL Teams, Coaches, and Players (Team Project)
- IST 718 Big Data Analytics
 - IEEE-CIS Fraud Detection (Team Project)
- IST 707 Data Analytics
 - Real Estate Property Investment (Individual Project)
- IST 659 Database Administration Concepts and Database Management
 - A4B KPI BizOps Organization Database (Individual Project)
- MBC 638 Data Analysis and Decision Making
 - Purchase Order Process Improvement (Individual Project)
- SCM 651 Business Analytics
 - Recruiting Advertising Strategy (Team Assignment)

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Learning Objective - Project- Alignment Overview

- ❖ Collect and Organize Data
 - Public Sentiment Toward NFL Teams, Coaches, and Players
 - Real Estate Property Investment
 - Purchase Order Process Improvement

- ❖ Identify Patterns in data via visualization, statistical analysis, and data mining
 - Public Sentiment Toward NFL Teams, Coaches, and Players
 - IEEE-CIS Fraud Detection
 - Real Estate Property Investment

- ❖ Develop alternative strategies based on the data
 - Real Estate Property Investment
 - Recruiting Advertising Strategy

- ❖ Develop a plan of action to implement the business decisions derived from analysis
 - Purchase Order Process Improvement
 - A4B KPI BizOps Organization Schema

- ❖ Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals
 - Purchase Order Process Improvement
 - Public Sentiment Toward NFL Teams, Coaches, and Players
 - IEEE-CIS Fraud Detection

- ❖ Synthesize the ethical dimensions of data science practice
 - A4B KPI BizOps Organization Database
 - IEEE-CIS Fraud Detection

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA



IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players- Question and Problem to Solve

Business Driver

- NFL revenue grew an estimated \$900 million to \$14 billion in 2017, in 2018 it generated about \$15 billion.
- Fantasy football and the spread of legalized sports betting across the U.S. promises to lock in fans and keep them focused on the game.
- As a Fantasy football player, how can Data Science help me make the most intelligence selections when deciding my weekly roster? “How do I win more?”

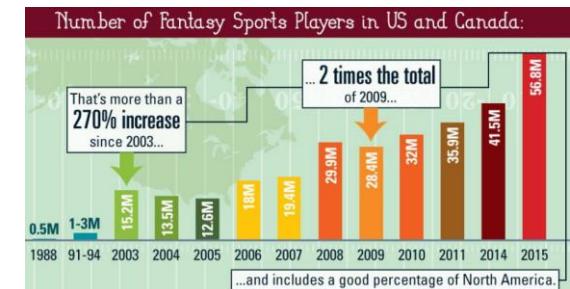
Fantasy Football is an \$18.6 BILLION Market.

About the Data:

- Text data will be mined from three primary sources
- 1) Fantasy football websites like Yahoo Sports:
 - 2) Twitter Social Media API live streams
 - 3) Twitter Social Media Historical API search

Problem to Solve

- Identify if real-time public popular opinion on NFL teams, players, and coaches is a predictor of their weekly fantasy football stats.



The left slide is titled "2019-1002 IST 736 Text Mining Final Project Public Sentiment Toward NFL Team, Coach, Player Can it predict weekly Fantasy Football outcomes? Ryan Timbrook, Dave Pescosolido, Diego Vales Course: IST 736 Text Mining Term: Fall 2019". The right slide is titled "Public Sentiment Toward the NFL Can it predict weekly Fantasy Football outcomes? Ryan Timbrook, Olego Vales, David Madsen School of Information Studies SYRACUSE UNIVERSITY".





IST 718 Big Data Analytics

-IEEE-CIS Fraud Detection-

Question and Problem to Solve

Business Driver

- Improve the efficacy of fraudulent transaction alerts, helping hundreds of thousands of businesses reduce their fraud losses and increase their revenue; while securing consumer's peace of mind and wallets!



kaggle 

Problem to Solve

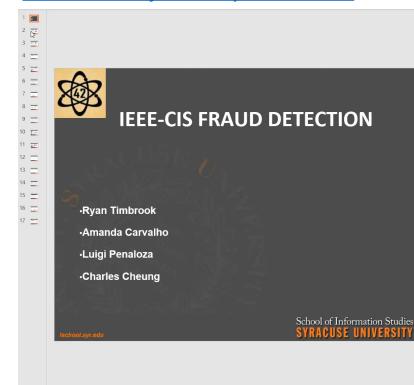
- Identify real-time fraudulent e-commerce transactions, using advanced Machine Learning algorithms, by automating alerts that block highly suspicious activities.



About the Data:

- The core data set for this project is provided by VESTA, the worlds leading payment service company, and is a kaggle competition being facilitated by the [IEEE Computational Intelligence Society](#).
- The data is broken into two files **identity** and **transaction**, which are joined by **TransactionID**. *Not all transactions have corresponding identity information.*

***[Link to Project Report Video](#)





IST 707 Data Analytics

-Real Estate Property Investment-

Question and Problem to Solve

Business Driver

- Targeting low risk property investment opportunities for property management firms or individual investors who buy-rent-sell single family homes throughout the United States.
- Given a base set of investment criteria, provide a predicted N-best list of US geolocation regions by zip code that offer the best ROI.



About the Data:

- Base datasets are provided by Zillow research data:
 - Home Values
 - Home Listings and Sales
 - Rental Values
- Additional data will be mined from:
 - Bureau of Labor Statistics and Census
 - Capital markets and economics



Problem to Solve

- How to predict a low risk / high yield return on property investment in a volatile market.
- Buy low, rent fair, sell high...
- Where and when to buy and sell that maximizes investment profits.
- Forecast future growth and decline of a region that yields Net Present Value (NPE) measurements significant enough to act on.

***[Link to Project Report Video](#)

The screenshot shows a slide titled "Real Estate Property Investments" with the subtitle "Invest with sound, objective data driven recommendations". It includes a brief description of the project's goal: to help users make informed decisions about property investments. The slide is divided into sections: "1. Introduction", "1.1 Problem Statement", "1.2 About the Data", and "1.3 Data Sources". Each section contains bullet points detailing specific goals or data sources. At the bottom right of the slide, there is a small "12" indicating it is the 12th slide in the presentation.



IST 659 Database Administration & Management

-A4B KPI BizOps Organization Database-

Question and Problem to Solve

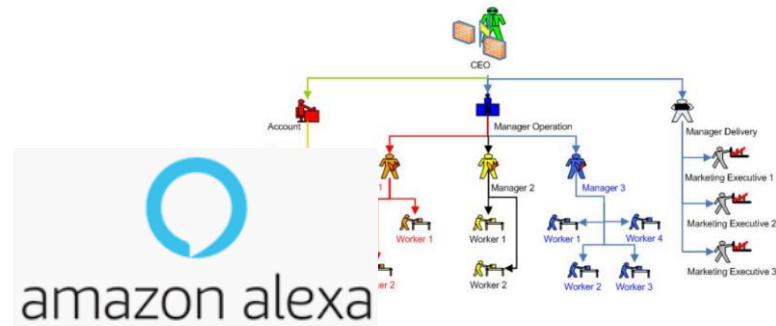
Business Driver

- Innovation of alternative Voice Channels for the Customer Service organization that are more cost effective than traditional telephony voice such as IVRs.
- Business and Operations KPIs are daily ‘must know’ factors that are challenging to get real-time updates on and take excessive resource time to pull reports together.



Problem to Solve

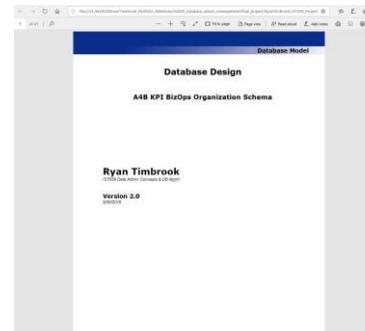
- No single source of truth exists for the Company’s Organization Hierarchy that includes application and business operation domains along with their KPI mapping.



About the Data:

- Company Organization data is pulled from internal LDAP system.
- Application Ownership and KPI sources are gathered from internal audit and team interviews.

***[Link to Project Report Video](#)





MBC 638 Data Analysis & Decision Making

- Purchase Order Process Improvement- Question and Problem to Solve

Business Driver

- 80% of vendor Purchase Orders (PO) for contingent staffing workforce are NOT being approved within the vendor offered 2% discount time period. On average, for each PO this is a missed opportunity of \$5,911 in cost savings.
- Lengthy approval cycle-times and the hours software development team's management spend validating the accuracy of vendors PO Invoices leads to decreased product output hurting the business's competitive 'Time to Market' strategy.



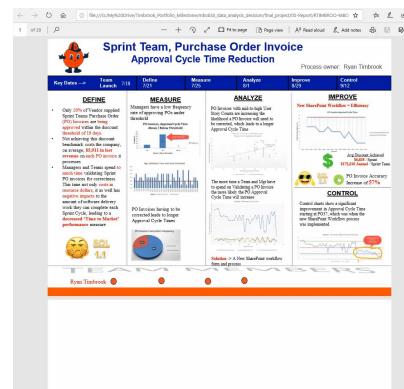
Problem to Solve

- Identify areas of waste within the PO approval workflow process.
- Devise a more efficient mechanism of collecting, organizing and reporting details of the work the vendor is submitting PO Invoices for.
- Implement a solution to be measured and establish control methods to ensure the process has improved the situation and is continually being enhanced.

About the Data:

- The data for this initiative comes from multiple disparate internal sources. It's generated from the Sprint Development Teams performing tasks associated with business deliverables. Systems like Rally and SharePoint capture the Sprint User Story tasks the Vendor submits PO Invoices for on a bi-weekly basis. PO Invoices are entered into SAP Ariba for internal processing including approvals for payment workflows.

***[Link to Project Report Video](#)





SCM 651 Business Analytics

-Recruiting Advertising Strategy-

Question and Problem to Solve

Business Driver

- Derive insights and knowledge from a Google ads advertisement recruitment campaign launched in 2011 for The Whitman School of Management in order to create a roadmap for 2020's new recruitment campaign.

Problem to Solve

- Identify time frames for each marketing campaign and the cost spent on each.
- Identify what the effectiveness was for each of the previous campaigns.
- Identify the key aspects of a United States campaign for 2020.
- Identify the costs for the advertising campaigns.
- Specify how to best measure performance of the decisions made after implementation.
- Specify other factors or considerations that are important.
- For 2020's campaign, stay within a \$100,000 budget.



About the Data:

- Witman.syr.edu website user traffic analytics collected by google analytics from March 2011 to October 2012.

United States Campaign Measurements									
Campaign	Start Date	End Date	Cost	Cost per Click	Users	New Users	Total Sessions	Bounce Rate	Page per Session
whitman.syr.edu	2/15/2011	8/15/2011	\$ 37,851.96	\$ 3.93	133,372	98,379	392,336	48.08%	1.76
MBA Marketing - Full-time	Jan/15/2012	2/15/2012	\$ 16,459.90	\$ 4.64	1,347	1,348	1,468	97.07%	1.05
MBA Marketing - IMBA	3/1/2012	10/15/2012	\$ 144,971.70	\$ 7.46	2,753	2,748	3,079	88.90%	1.14
Delta	9/1/2013	11/1/2013	\$ 10,000.00	\$ -	-	-	284	28.17%	3.17

***[Link to Project Report Video](#)

2019-0703 SCM 651 Business Analytics
Homework Assignment 2 (week 6)

Team Name: Team 5
Team Members:

- Ryan Timbrook
- Christopher Webster
- David Boni

Assignment Topic: Recruiting Advertising Strategy
Due Date: 8/8/19

Learning Outcome Objectives

- ❖ Collect and Organize Data
- ❖ Identify Patterns in data via visualization, statistical analysis, and data mining
- ❖ Develop alternative strategies based on the data
- ❖ Develop a plan of action to implement the business decisions derived from analysis
- ❖ Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals
- ❖ Synthesize the ethical dimensions of data science practice

-Collect and Organize Data- Overview

- Public Sentiment Toward NFL Teams, Coaches, and Players
- Real Estate Property Investment
- Purchase Order Process Improvement

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Learning Objective

-Collect and Organize Data-

➤ Public Sentiment Toward NFL Teams, Coaches, and Players

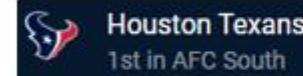
The Bishop @BillBishopKHOU · Dec 4, 2019
Good morning. The NFL has named **Houston Texans QB Deshaun Watson** the AFC Offensive Player of the Week following his performance in the team's 28-22 victory over the New England Patriots in Week 13. #khou #HTownRush

DreamTeamApple @DreamTeamApple1 · Oct 28, 2019
Replying to @NFLFilms @deshawnwatson and 2 others
I've been disappointed with the **Texans** offensive line taking a step back with their pass protection since Tyus Howard's injury, but **Deshawn Watson** has been making sensational plays even when the pocket collapses! A lot like he did last year. **Deshawn Watson** is Michael Jordan!

AdvoSports @advosports · Dec 1, 2019
HOUSTON (AP) — **Deshawn Watson** threw three touchdown passes and had the first TD reception of his career, and the **Houston Texans** frustrated Tom Brady in a 28-22 victory over the New England Patriots on Sunday night. victoriaadvocate.com/ap/sports/wats...

FleaFlickerFFB @FleaFlickerFFB · Nov 17, 2019
I'm taking **Deshawn Watson** and the **Houston Texans** to beat the Lamar Jackson led Baltimore Ravens.

Fun game in store between these two young QB's



PRESEASON				SEPTEMBER				OCTOBER				NOVEMBER				DECEMBER				
AT	VS	AT	VS	WK 1 • AT	WK 2 • VS	WK 3 • AT	WK 4 • VS	WK 5 • VS	WK 6 • AT	WK 7 • AT	WK 8 • VS	WK 9 • AT	WK 11 • AT	WK 12 • VS	WK 13 • VS	WK 14 • VS	WK 15 • AT	WK 16 • AT	WK 17 • VS	
THURSDAY AUG 8 7:00 PM ABC 13	SATURDAY AUG 17 7:00 PM ABC 13	SATURDAY AUG 24 6:00 PM ABC 13	THURSDAY AUG 29 7:00 PM ABC 13	MONDAY SEPT 9 8:10 PM ESPN	SUNDAY SEPT 15 12:00 PM CBS	SUNDAY SEPT 22 3:25 PM CBS	SUNDAY SEPT 29 12:00 PM FOX	SUNDAY OCT 6 12:00 PM FOX	SUNDAY OCT 13 12:00 PM CBS	SUNDAY OCT 20 12:00 PM CBS	SUNDAY OCT 27 12:00 PM CBS	SUNDAY NOV 3 8:30 AM NFLN + LONDON	SUNDAY NOV 17 12:00 PM CBS	THURSDAY NOV 21 7:20 PM FOX/AMAZON	SUNDAY DEC 1 7:20 PM NBC	SUNDAY DEC 8 12:00 PM CBS	SUNDAY DEC 15 7:00 PM CBS	TBD TBD	DEC 21/22 TBD	SUNDAY DEC 29 12:00 PM CBS

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

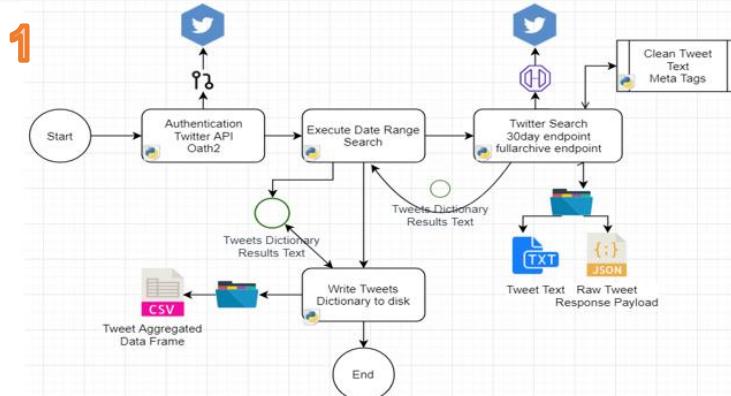
Collect and Organize Data

IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players-

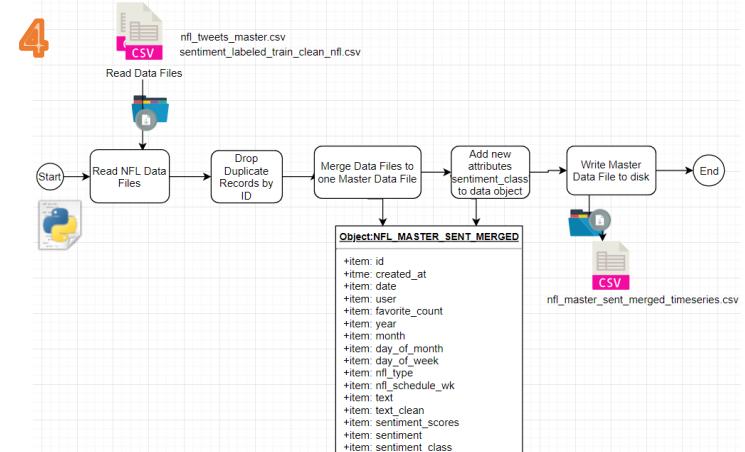
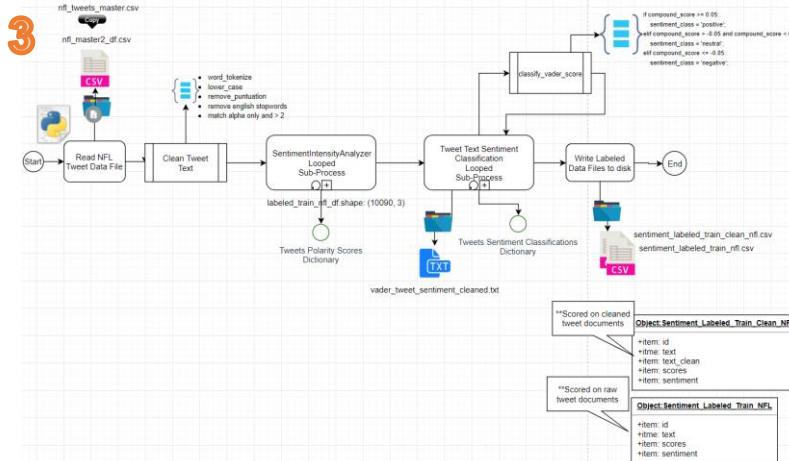
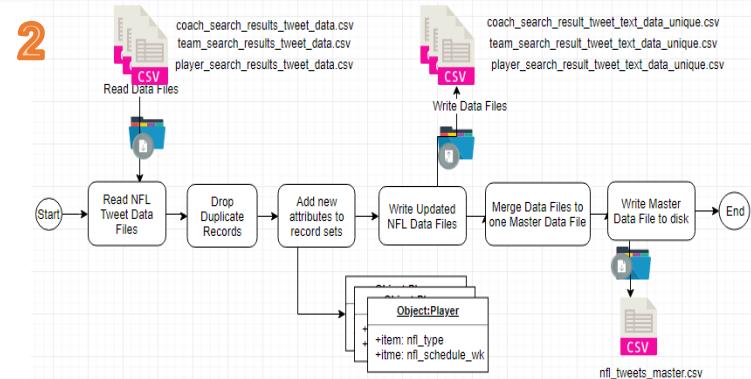
Collection Methods:

- Internet Web Scraping
- Twitter API Live Stream Capture
- Twitter API Historical Search Capture



Organization Methods:

- NFL type object modeling
- Local file system, text corpus, storage and retrieval
- Data engineering pipeline



Collect and Organize Data

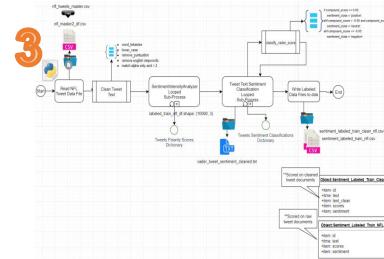
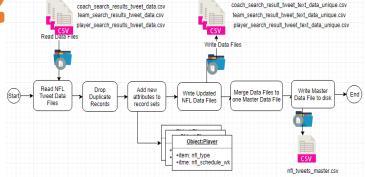
IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players-

Organization Methods:

- NFL type object modeling
- Local file system, text corpus, storage and retrieval
- Data engineering pipeline

2



Format Tweets JSON Code Overview

```

jupyter format_raw_twitter_data (advanced)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 Logout
In [1]: # Objective:
# This function of the data engineering step reads in the sentiment_labeled_train_clean .csv data file generated by the process executed in section 2.1.6 and the raw_tweets .csv data file generated by the process executed in section 2.1.8, and outputs a merged data file to be used for modeling, visualization, and analysis.

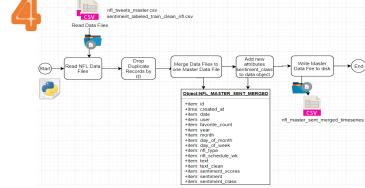
In [2]: # import pandas as pd
# import numpy as np
# import os
# import json
# import re
# import string
# import math
# from datetime import date
# from datetime import timedelta
# from datetime import datetime
# In [3]: # custom python packages
# import rtweet
# import rtweet_util as br

In [4]: # set global properties
# notebook_file_name = 'rtweet_raw_twitter_data'
# report_file_name = 'rtweet_raw_twitter_data'
# app_name = 'rtweet_raw_twitter_data'
# log_level = 10 # 10=DEBUG, 20=INFO, 30=WARNING, 40=ERROR, 50=CRITICAL
# set working directory structure
# data_dir = 'data'
# log_dir = 'logs'
# in_dir = 'in'
# out_dir = 'out'
# temp_dir = 'temp'

# create base output directories (if they don't exist)
# if not os.path.exists(logDir): os.mkdir(logDir)
# if not os.path.exists(logDir): os.makedirs(logDir)
# if not os.path.exists(dataDir): os.makedirs(dataDir)
# 
```

Detailed description: This screenshot shows a Jupyter Notebook cell containing Python code for processing tweets. The code imports necessary libraries like pandas, numpy, os, and json, and sets global properties for the notebook and report. It defines variables for file names, app names, log levels, and directory structures. It also handles the creation of log and data output directories if they do not already exist.

4



Sentiment Analysis Merge

```

jupyter nfl_sentiment_analysis_data_merge_master (advanced)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 Logout
In [1]: # Objective:
# This function of the data engineering step reads in the sentiment_labeled_train_clean .csv data file generated by the process executed in section 2.1.7 and the nfl_tweets_master .csv data file generated by the process executed in section 2.1.8, and outputs a merged data file to be used for modeling, visualization, and analysis.

In [2]: # import pandas as pd
# import numpy as np
# import os
# import json
# import re
# import string
# import math
# from datetime import date
# from datetime import timedelta
# from datetime import datetime
# In [3]: # custom python packages
# import rtweet
# import rtweet_util as br

In [4]: # set global properties
# notebook_file_name = 'nfl_sentiment_analysis_data_merge_master'
# report_file_name = 'nfl_sentiment_analysis_data_merge_master'
# app_name = 'nfl_sentiment_analysis_data_merge_master'
# log_level = 10 # 10=DEBUG, 20=INFO, 30=WARNING, 40=ERROR, 50=CRITICAL
# set working directory structure
# data_dir = 'data'
# log_dir = 'logs'
# in_dir = 'in'
# out_dir = 'out'
# temp_dir = 'temp'

# create base output directories (if they don't exist)
# if not os.path.exists(logDir): os.mkdir(logDir)
# if not os.path.exists(logDir): os.makedirs(logDir)
# if not os.path.exists(dataDir): os.makedirs(dataDir)
# 
```

Detailed description: This screenshot shows a Jupyter Notebook cell containing Python code for merging sentiment analysis data. The code imports pandas, numpy, os, and json, and sets global properties for the notebook and report. It defines variables for file names, app names, log levels, and directory structures. It also handles the creation of log and data output directories if they do not already exist.

Learning Objective

-Collect and Organize Data-

➤ Real Estate Property Investment



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Collect and Organize Data

IST 707 Data Analytics

-Real Estate Property Investment-

Collection Methods:

- Real Estate datasets pulled from Zillow Research web site.
- U.S. Economic indicators datasets pulled from [Datahub.io](#)

Organization Methods:

- Exploratory data analysis was performed on each of the datasets, five Zillow real estate and fourteen U.S. National Yearly Economic reports
- All datasets were reorganized and structured into timeseries format
- Data scrubbing, cleaning, transformation, and merging of disparate datasets was performed for analyses

** Merged all disparate Economic dataset features to one common source for analysis

Year	FF_Target_Rate_Avg	Inflation	GDP	GDP_Percent_Change	Education_Budget	Population	Investor_Flow_Avg	National_HPI_Avg	House_Hold_Income	Employed	Employed_Percent	Unemployed	Unemployed_Percent	CPI_Index_Avg	Cash_Surp_Def	Bond_Yield_10y_Avg	
1982	9.392857	6.203740	3345.0	8.8	15374.0	231664000.0		NaN	NaN	83918	99526	57.8	10678	9.7	96.500000	-2.735732	13.001667
1983	9.053125	3.948367	3638.1	11.1	15267.0	233792000.0		NaN	NaN	85290	100834	57.9	10717	9.6	99.600000	-4.819500	11.105000
1984	10.150000	3.548237	4040.7	7.6	15336.0	235825000.0		NaN	NaN	86789	105005	59.5	8539	7.5	103.883333	-3.676845	12.438333
1985	8.044643	3.199612	4346.7	5.6	18952.0	237924000.0		NaN	NaN	88458	107150	60.1	8312	7.2	107.566667	-3.927316	10.623333
1986	6.740132	2.017624	4590.2	6.1	17750.0	240133000.0		NaN	NaN	89479	109597	60.7	8237	7.0	109.608333	-4.248440	7.682500

Final Merged Dataset - Real Estate Combined with Economic Data Features

Time range - 1997 - 2017 (that was the cleanest that could be achieved at this time...)

*Train classifiers on Feature 'Price_Point_Class'

- 0: means observation's price value is < 25% of the State Price Average
- 1: means observations fall within the normal (average) range of the State Price Average
- 2: means observations falls above the 75% range of the State Price Average

--Determine if classifiers can identify future home value classes based on prior date, location and economic features that have the most impact on both positive and negative price value swings...

- Dataset Shape: (88452, 16)

Date	ZipCode	log_Price	log_Price_Monthly_Avg	log_Price_diff	Price_Point_Class	CPI_Index_Avg_f	Interest_Rate_f	Housing_Price_Index_f	Bond_Yield_10y_f	Inflation_f	GDP_f	Population_f	House_Hold_Income_f	Employment_f	Cash_Surp_Def
1997-01-01	98052	12.342486	11.804569	-0.537918	0	157.959370	4.814434	83.076214	6.055381	2.812443	8577.554463	272639888.8	11.537125	11.767443	-3.733803
1997-02-01	98052	12.352806	11.808970	-0.543837	0	158.404761	4.731693	83.392929	6.015761	2.812443	8577.554463	272639888.8	11.537125	11.767443	-3.733803
1997-03-01	98052	12.357526	11.811044	-0.546482	0	158.947674	4.633084	82.836534	6.073349	2.812443	8577.554463	272639888.8	11.537125	11.767443	-3.733803
1997-04-01	98052	12.367903	11.815443	-0.552460	0	159.387626	4.378458	83.893792	6.079290	2.812443	8577.554463	272639888.8	11.537125	11.767443	-3.733803
1997-05-01	98052	12.378131	11.819842	-0.558289	0	159.765440	4.652420	85.006970	6.126489	2.812443	8577.554463	272639888.8	11.537125	11.767443	-3.733803

Collect and Organize Data

IST 707 Data Analytics

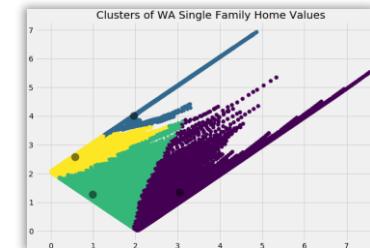
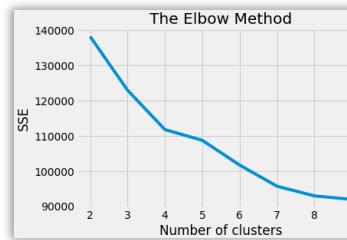
-Real Estate Property Investment-

REAL ESTATE DATASETS - ZILLOW

- Rename 'Region Name' Column to ZipCode
- Convert ZipCode field to string
- Remove columns of non-interest:
 - 'RegionID', 'SizeRank', 'City', 'Metro', 'CountyName'
 - '1996-04', '1996-05', '1996-06', '1996-07', '1996-08', '1996-09', '1996-10', '1996-11', '1996-12'
 - '2019-01', '2019-02', '2019-03', '2019-04', '2019-05', '2019-06', '2019-07', '2019-08', '2019-09'
- Fill NaN with median value

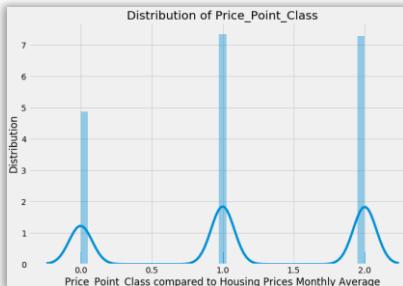
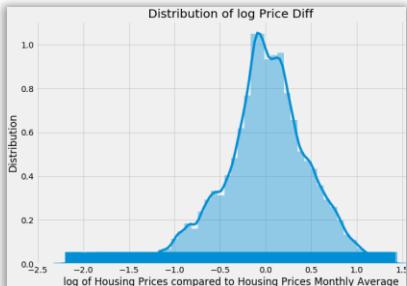
Clean the forecast dataset for clustering

- limite the features for clustering - and the observations to just the predition time (5 years) + one year observed
- remove additive terms and multiplicative terms as well as the datetimestamp
- save series objects for later re joining



- Transformation of the data's necessary to merge the datasets together after processed through prophet
- Look over the distribution of key features
- Set price thresholds for supervised learning classification
- Price_Point_Class is a generated feature for supervised classification. Details are shown below

Date	ZipCode	log_Price	log_Price_Monthly_Avg	log_Price_diff	Price_Point_Class
1997-01-01	98052	12.342486	11.804569	-0.537917	0
1997-02-01	98052	12.352806	11.808970	-0.543837	0
1997-03-01	98052	12.357526	11.811044	-0.546482	0
1997-04-01	98052	12.367903	11.815443	-0.552460	0
1997-05-01	98052	12.378131	11.819842	-0.558289	0



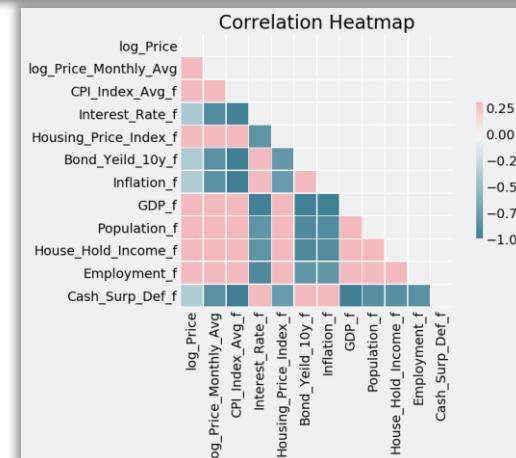
Final Merged Dataset - Real Estate Combined with Economic Data Features

Time range - 1997 - 2017 (that was the cleanest that could be achieved at this time...)
*Train classifiers on Feature 'Price_Point_Class'

- 0: means observation's price value is < 25% of the State Price Average
- 1: means observations fall within the normal (average) range of the State Price Average
- 2: means observations falls above the 75% range of the State Price Average

--Determin if classifiers can identify future home value classes based on prior date, location and economic features that have the most impact on both positive and negative price value swings...

- Dataset Shape: (88452, 16)



Collect and Organize Data

IST 707 Data Analytics

-Real Estate Property Investment-

Data Collection and Organization Code Overview

The screenshot shows a Jupyter Notebook interface with the title "Ryan_Timbrook_Project_Report" (autosaved). The notebook contains a section titled "Real Estate Property Investments" with the following content:

Real Estate Property Investments

Invest with sound, objective data driven recommendations

Syracuse Applied Data Science, IST-707 Data Analytics

Ryan Timbrook (RTIMBROO)
DATE: 9/8/2019 ASSIGNMENT: Final Project

1. Introduction

A real estate transaction can be an emotional time for everyone. The complexities between buyers and sellers are the result of different experiences and expectations. Success in today's market is guided by knowledge, communication, and partnership.

Buyers are waiting later in life to purchase their first home. They have very specific expectations on what they are looking for, and willing to take the time to get exactly what they want. To be successful, buyers will turn to experienced professionals to guide them through the buying process and to sift through the voluminous of data.

Sellers past experiences have been rooted in market conditions significantly different than we are seeing today. Many are resisting the realities of the market and are slow to react to the valuable feedback the data provides. To be successful, sellers will need to utilize skilled professionals to interpret the specifics of today's market and take swift action to adjust for changing trends.

1.1 Problem Statement:

- How to predict a low risk / high yield return on property investment in a volatile market.
- Where and when to buy and sell that maximizes investment profits.
- Forecast future growth and decline of a region in order to guide investors with optimized, data driven, recommendations.

1.2 About the Data

Base Real Estate data provided by: [Zillow](#)
Base Federal Reserve data provided by: [kaggle]<https://www.kaggle.com/federalreserve/interest-rates>

Base Economic data sets provided by: [datahub io]<https://datahub.io/core/cpi-usa/image.png>

Zillow Data: Timeseries Real Estate data by ZipCode U.S.
Zillow Home Value Index (ZHVI): A smoothed, seasonally adjusted measure of the median estimated home value across a given region and housing type. It is a dollar-denominated alternative to repeat-sales indices.

- Zip_Zhvi_SingleFamilyResidence.csv
- Zip_Zhvi_AllHomes.csv
- Zip_MedianRentalPricePerSqr_Fr.csv
- Zip_MedianRentalPrice_AllHomes.csv
- Zip_MedianListingPrice_AllHomes.csv

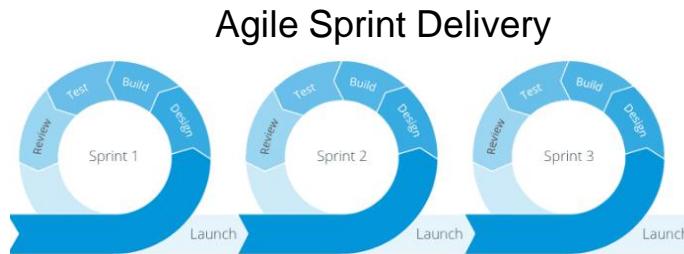
Datahub.io: U.S., National Yearly Economic Reports

- interest_rates.csv
 - Inflation, GDP deflator (annual %) and Inflation, consumer prices (annual %) for most countries in the world when it has been measured. Data The data comes from The World Bank (CPI), The World Bank (GDP) and is collected from 1973 to 2014. There are some values missing from data
- inflation-consumer.csv
- inflation-gdp.csv
- education_budget_data.csv
 - United States of America Education budget to GDP analysis Data Data comes from Office of Management and Budget, President's Budget from white house official
- population.csv

Learning Objective

-Collect and Organize Data-

➤ Purchase Order Process Improvement



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Collect and Organize Data

MBC 638 Data Analysis and Decision Making

-Purchase Order Process Improvement-

Collection Methods:

- Baseline data sources for this initiative are found in three primary internal business operations systems such as SAP Ariba, CA Rally, and Microsoft SharePoint. Each system was queried to extract the necessary data needed for analysis.

Organization Methods:

- After each of the collection cycles, the data is entered into a master spreadsheet for analyses.
- During the Define phase of the DMAIC lifecycle, Project Definitions & Terms along with Data Collection Attributes are defined.
- As a function of the process improvement workflow, data is entered into SharePoint forms that organizes and standardizes the data entry for consistency and automated workflow mechanisms.

Project Definitions & Terms

Defects

- A PO invoice not approved in the Ariaba procurement system within the discount threshold time
- A PO invoice needing correction after submission in Ariaba
- A Manager spending more than 2 hours validating and approving a PO invoice
- A Dev Team spending more than 3 hours validating and or correcting a PO invoice

Corporate Measurement Goals

- PO Discount Threshold is **19 days**
- Manager PO Invoice 'Validation Time Spent' Threshold is **2 hours**
- Team PO Invoice 'Validation Time Spent' Threshold is **3 hours**
- Less than 10%** of PO Invoices are rejected for needing corrections
- 80% or more** of PO Invoices are approved within the Discount Threshold

SQL Definitions

- Unit:** A Unit is a PO; 4 possible Defect Units have been identified per PO Invoice Submission
- Timeframe:** Two week Sprint Cycle
- Units per Timeframe:** 3 PO Invoices are created each Sprint Cycle based on the number of Sprint Teams managed by the sampled Dev Manager
- PO Approval Cycle Time:** Refers to the time from when the Vendor submits a PO invoice till the time when the Manager approves the invoice in Ariaba

The GOAL is to **REMOVE WASTE**
This will **improve EFFICIENCY and PRODUCTIVITY**

Define

Measure Analyze Improve Control

Data Collection Attributes

Column Definitions	
Sprint	Sprint sequence name for the EIT department
Sprint Start Date	Sprints are in two week cycles starting wednesdays
Sprint End Date	Sprint cycle end date, every other Tuesday
Sprint Team	Name of the Sprint Team the PO is billed to
User Stories Count	Number of Rally User Stories complete for given sprint
Unique Project Count	Number of Unique Projects the User Stories align to
PO Line Item Count	Number of line items displayed on Arabia PO invoice
PO Submission Date	Date the PO was submitted by the Vendor
PO Approved Date	Date the PO was approved by Dev Manager of Sprint Team
Needed Correction	Flag specifying if the PO had errors and needed correction
PO Approval Cycle Time	Days between PO submission date and the Dev Manager approving the PO
PO Cycle Time	Total cycle time to complete the PO billing process. Starts 1 day post sprint end date
Within Disc Threshold	Flag indicator specifying if the PO cycle time was within the discount threshold of 15 biz days
Team Validation Meetings	Cumulative team time to conduct meetings to validate PO line item data
Mgr Validation Time	DevManager time to pull User Story data from Rally and validate it against PO line items
PO Est. Cost (\$)	Vendor estimated calculated cost based on Sprint Team Resources needed to deliver User Stories for Sprint
PO Act. Cost (\$)	Vendor actual calculated cost based on Sprint Team Resources needed to deliver User Stories for Sprint
PO Est. LOE (hrs)	Vendor calculated sum of estimated level of efforts to deliver each User Story for a Sprint
PO Act. LOE (hrs)	Vendor calculated sum of actual level of efforts to deliver each User Story for a Sprint

Define

Measure Analyze Improve Control

Collect and Organize Data

MBC 638 Data Analysis and Decision Making

-Purchase Order Process Improvement-

Data Collection and Organization Code Overview

1 of 28 | file:///G:/My%20Drive/Timbrook_Portfolio_Milestone/mbc638_data_analysis_decision/final_project/05-Report/RTIMBRO-MBC638-ProssesImprovementProject.pdf

Sprint Team, Purchase Order Invoice Approval Cycle Time Reduction

Process owner: Ryan Timbrook



Key Dates -->	Team Launch	7/18	Define	7/21	Measure	8/25	Analyze	8/1	Improve	8/29	Control	9/12
---------------	-------------	------	--------	------	---------	------	---------	-----	---------	------	---------	------

DEFINE

- Only 20% of Vendor supplied Sprint Teams Purchase Order (PO) Invoices are being approved within the discount threshold of 19 days.
- Not achieving this discount benchmark costs the company, on average, \$5,911 in lost revenue on each PO invoice it processes.
- Managers and Teams spend to much time validating Sprint PO invoices for correctness. This time not only costs in resource dollars, it as well has negative impacts to the amount of software delivery work they can complete each Sprint Cycle, leading to a decreased 'Time to Market' performance measure.

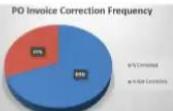
 

MEASURE

Managers have a low frequency rate of approving POs under threshold

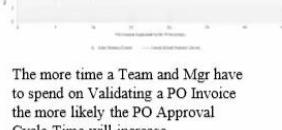
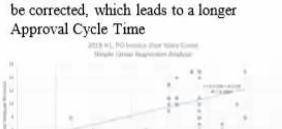


PO Invoices having to be corrected leads to longer Approval Cycle Times

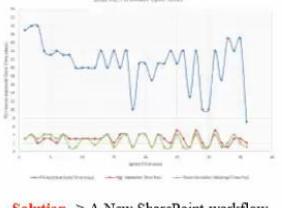


ANALYZE

PO Invoices with mid-to-high User Story Counts are increasing the likelihood a PO Invoice will need to be corrected, which leads to a longer Approval Cycle Time



The more time a Team and Mgr have to spend on Validating a PO Invoice the more likely the PO Approval Cycle Time will increase



Solution -> A New SharePoint workflow form and process

IMPROVE

New SharePoint Workflow = Efficiency



Avg Discount Achieved
\$6,635 / Sprint
\$172,510 Annual / Sprint Team

   PO Invoice Accuracy Increase of 57%

CONTROL

Control charts show a significant improvement in Approval Cycle Time starting at PO37, which was when the new SharePoint Workflow process was implemented



TEAM MEMBERS

Ryan Timbrook 

-Identify patterns in data via visualization, statistical analysis, and data mining- Overview

- Public Sentiment Toward NFL Teams, Coaches, and Players
- IEEE-CIS Fraud Detection
- Real Estate Property Investment

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



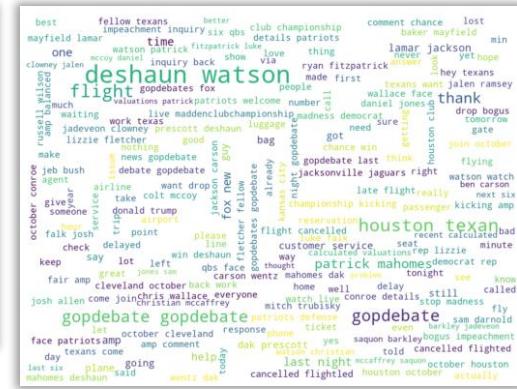
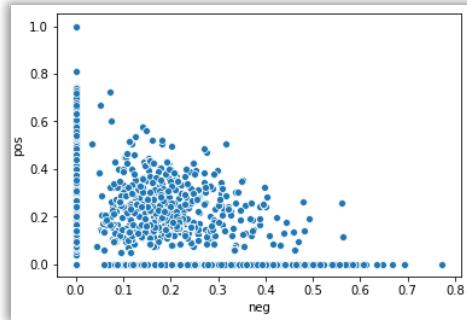
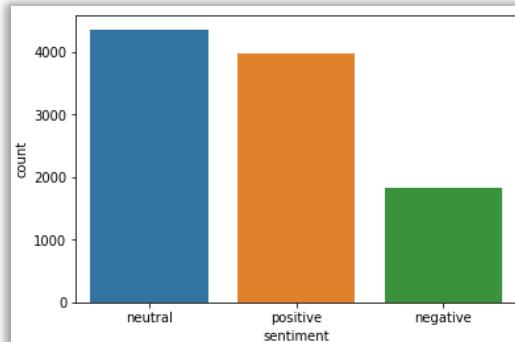
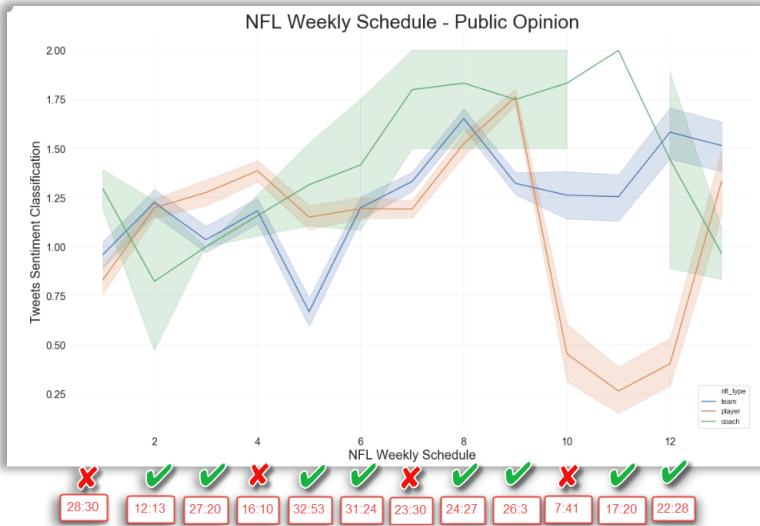
Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Learning Objective

-Identify patterns in data via visualization, statistical analysis, and data mining-

➤ Public Sentiment Toward NFL Teams, Coaches, and Players



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

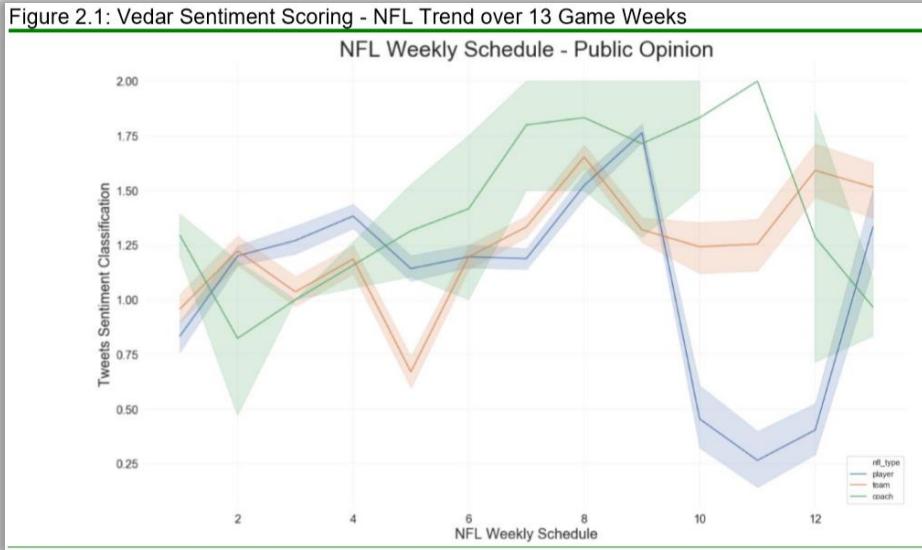
iSCHOOL.SYR.EDU/BIGDATA

Identify patterns in data via visualization, statistical analysis, and data mining

IST 736 Text Mining

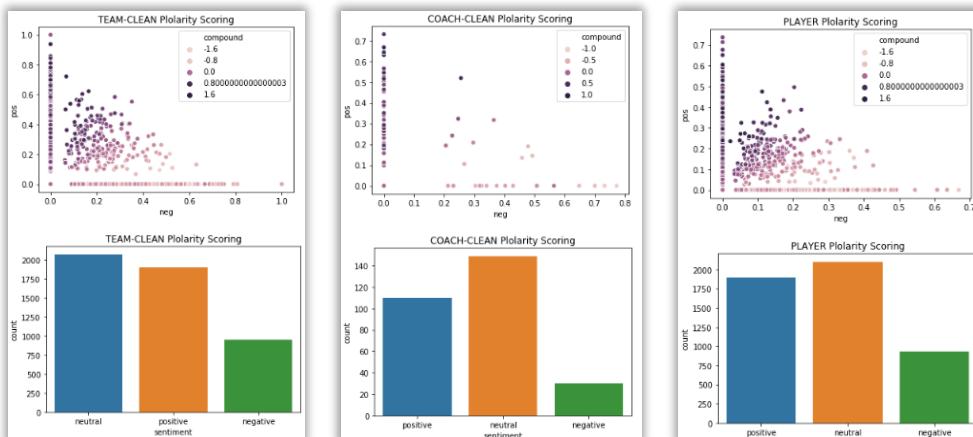
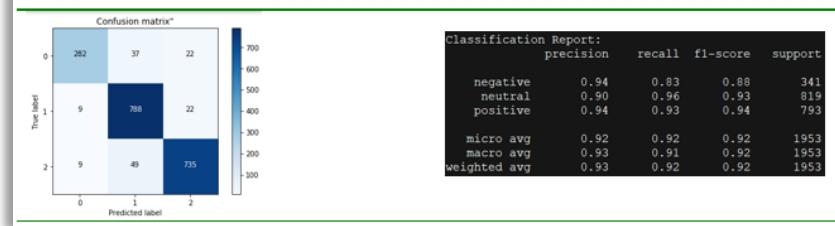
-Public Sentiment Toward NFL Teams, Coaches, and Players-

Figure 2.1: Vedar Sentiment Scoring - NFL Trend over 13 Game Weeks

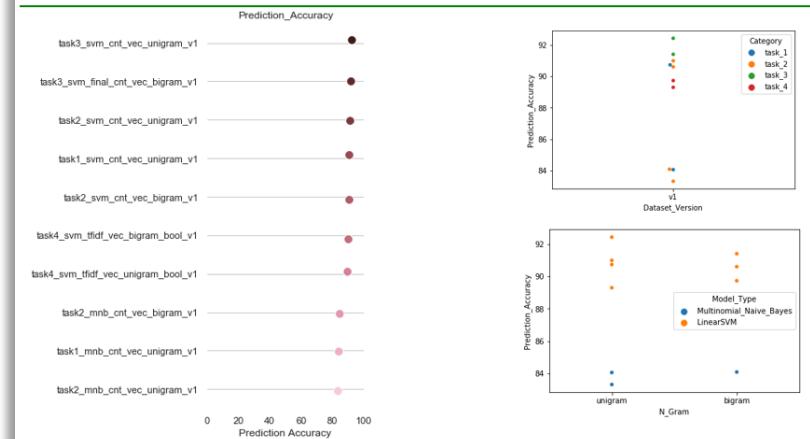


Most POSITIVE WORDS		Most NEGATIVE WORDS	
Top and Bottom 10 of Most POSITIVE Learned Words		Top and Bottom 10 of Most NEGATIVE Learned Words	
-1.3223 shit	1.9337 best	-1.6849 care	1.5858 axed
-1.2525 bad	1.9671 cheering	-1.4871 best	1.6238 offensive
-1.1760 anymore	1.9906 winners	-1.3031 wealth	1.6280 wrong
-1.1112 absolutely	2.0000 love	-1.2034 stay	1.6342 ill
-1.1050 usually	2.0000 easy	-1.0856 improve	1.6606 worst
-1.0301 kill	2.0000 easily	-1.0672 hope	1.6794 bad
-1.0247 hate	2.0239 win	-1.0495 fun	1.7560 problem
-1.0170 threat	2.0893 great	-1.0124 play	1.7898 ass
-1.0000 advertising	2.0930 better	-1.0000 comparing	1.8333 hurt
-1.0000 crushed	2.1714 bold	-0.9472 receiver	1.9542 injury

SVM Unigram Model Prediction Accuracy Results: 93%



3.1.1.5.1 Model Accuracy Comparison Summary



Identify patterns in data via visualization, statistical analysis, and data mining

IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players-

Patterns in Classification Modeling

The screenshot shows a Jupyter Notebook interface with the following content:

FINAL PROJECT
Ryan Timbrook (RTIMBROO)
DATE: 12/3/2019
Topic:

1. Objective

Task 1:

- Build a unigram MNB model and a unigram SVMs model.
- Print the top 10 indicative words for the most positive category and the most negative category from the MNB and SVMs models respectively.
- You can change other parameters to your preference. Report your choice and explain why.
- Report the confusion matrix, precisions, and recalls. Explain whether your models performed equally well on all categories, or some categories turn out to be easier or more difficult for MNB or SVMs.
- Submit your revised script along with your report.

Task 2:

Revise the script to build a MNB model and a SVMs model based on both unigram and bigram. For fair comparison, please keep the same 60% for training and the rest 40% for testing.

- Compare the confusion matrix and other evaluation measures (accuracy, precision, recall). Discuss whether adding bi-grams was helpful for sentiment classification, based on MNB and SVMs respectively.

Task 3:

Revise the script to build the best model by tuning parameters and using the entire training data set (changing from 60% to 100%). Report what parameters you used to train the model, and its cross validation accuracy.

FOR RUNNING IN GOOGLE COLAB ONLY

```
In [1]: 1 # toggle for working with colab
2 isColab = False

In [2]: 1 #ONLY RUN WHEN WORKING ON COLAB*
2 #####
3 # mount google drive for working in colab
4
5 #from google.colab import drive
6 #drive.mount('/content/gdrive', force_remount=True)
7
8 # working within colab, set base working directory
9 #base_dir = "./gdrive/My Drive/IST736_PRJ_Realestate/buy_rent_sell/"
10
11 # validate directory mapping
12 #ls f'{base_dir}'
13
14 # upload custom python files
15 #from google.colab import files
16 #uploaded_files = files.upload()
17
18 # print files uploaded
19 #for f in uploaded_files.keys():
20 #    print(f"file name: {f}")
21
22 #isColab = True
```

Identify patterns in data via visualization, statistical analysis, and data mining

IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players-

Patterns in Topic Modeling

The screenshot shows a Jupyter Notebook interface with the following details:

- Title Bar:** jupyter nfl_tweets_topic_modeling Last Checkpoint: 12/11/2019 (autosaved)
- Toolbar:** File Edit View Insert Cell Kernel Widgets Help
- Cell Header:** Trusted Python 3
- Section 1: Topic Modeling - Mallet LDA**
 - Author: Ryan Timbrook (RTIMBROO)
 - DATE: 12/7/2019
 - Topic: Perform Topic Modeling on NFL Tweet Text
- Section 2: 1. Objective**

Topic Modeling
LDA is an algorithm that can "summarize" the main topics of a text collection.
Topic modeling will be performed at the NFL Type level collected by NFL Game Schedule Week.

 - Coach
 - Team
 - Player
- Section 3: Topic Modeling References**
 - [sklearn LatentDirichletAllocation](#)
 - [Topic Modeling with SciKit Learn](#)
 - [Complete Guide to Topic Modeling](#)
 - [Topic Modelling with SciKit-learn -- Derek Greene University College Dublin](#)
- Section 4: Coding Environment Setup**

Import packages

```
In [1]: W 1 # import packages for analysis and modeling
2 import pandas as pd #data frame operations
3 import numpy as np #arrays and math functions
4 np.random.seed(42)
5
6 import matplotlib.pyplot as plt #2D plotting
7 %matplotlib inline
8 import seaborn as sns #
9 import os
10 import sys
11 import json
12 from os import path
13 import re
14 import random
15 import json
16 from datetime import date
17 from datetime import time
18 from datetime import datetime
19 import warnings
20 from timer import default_timer
21 import logging
22 warnings.filterwarnings('ignore')
23
24 # import nltk
25 nltk.download('wordnet')
26 from nltk import PorterStemmer
27 from nltk.stem import PorterStemmer
28 from nltk.stem import PorterStemmer
29 stemmer = PorterStemmer()
30
31 from nltk.corpus import stopwords
32 from nltk.tokenize import word_tokenize, wordpunct_tokenize
```

Identify patterns in data via visualization, statistical analysis, and data mining

IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players-

Patterns in Polarity Scoring

The screenshot shows a Jupyter Notebook interface with the title "jupyter classify_train_nfl_master (autosaved)". The notebook has a header section with author information: "Author: Ryan Timbroo (RTIMBROO)", "DATE: 12/3/2019", and "Topic:". Below this, a section titled "1. Objective:" is present. The main body of the notebook contains several code cells (In [1] through In [6]) containing Python code for data processing and analysis.

```
In [1]: 1 import pandas as pd
2 import numpy as np
3 import json
4 import os
5 from os import path
6 import fnmatch
7 import io
8 import re
9 import string
10 from datetime import date
11 from datetime import time
12 from datetime import datetime
13
14 import seaborn as sns
15 import matplotlib.pyplot as plt #2D plotting
16 %matplotlib inline
17 import warnings
18 warnings.filterwarnings('ignore')
19 #
20 from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

In [2]: 1 %set_env 'CATALYST_LOG_LEVEL'=15
env: 'CATALYST_LOG_LEVEL'=15

In [3]: 1 # custom python packages
2 import rtimbrou_utils as br

In [4]: 1 # set global properties
2 notebook_file_name = 'classify_train_nfl_master'
3 report_file_name = 'classify_train_nfl_master'
4 app_name = 'classify_train_nfl_master'
5 log_level = 10 # 10-DEBUG, 20-INFO, 30-WARNING, 40-ERROR, 50-CRITICAL
6
7 # setup working directory structure
8 # set global properties
9 dataDir = './data'
10 logOutDir = './logs',
11 imageDir = './images'
12 outputDir = './output'
13
14 # create base output directories if they don't exist
15 if not os.path.exists(logOutDir): os.mkdir(logOutDir)
16 if not os.path.exists(imageDir): os.mkdir(imageDir)
17 if not os.path.exists(dataDir): os.mkdir(dataDir)
18 if not os.path.exists(outputDir): os.mkdir(outputDir)

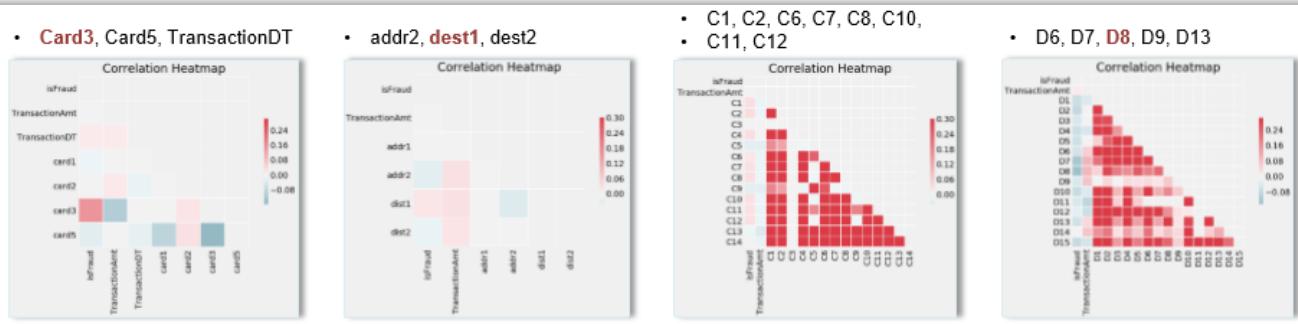
In [5]: 1 # get a logger for troubleshooting / data exploration
2 logger = br.getFileLogger(logoutDir+'/',app_name,level=log_level)

In [6]: 1 #train_nfl_master = pd.read_csv(f'{dataDir}/train_nfl_master.csv', encoding='latin')
2
3 # the nfl_master2_df is a copy of the nfl_tweets_master.csv created from the 'merge_datasets_to_master' notebook
4 # it's attributes have been reduced to just the 'id' and 'text' fields for classification
5 #train_nfl_master = pd.read_csv(f'{dataDir}/train/nfl_master2_df.csv', encoding='latin')
6
7 train_nfl_master = pd.read_csv(f'{dataDir}/nfl_tweets_master.csv', encoding='latin')
8
9 # nfl roster list to use for custom stop word removal
10 nfl_roster = pd.read_csv(f'{dataDir}/nfl_teams_roster_data.csv', encoding='utf8')
```

Learning Objective

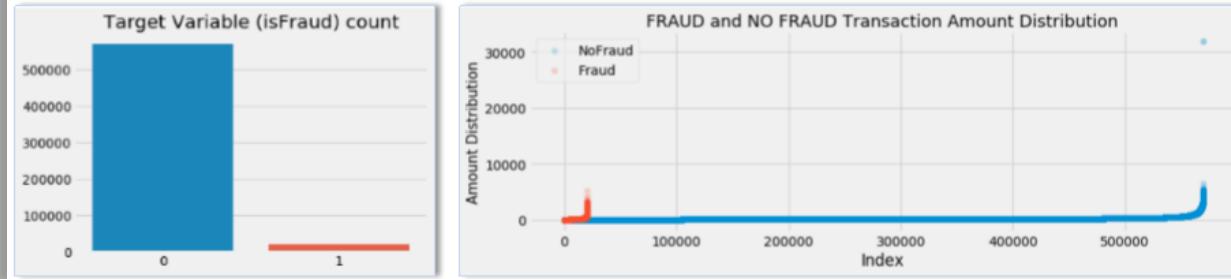
-Identify patterns in data via visualization, statistical analysis, and data mining-

➤ IEEE-CIS Fraud Detection



Representations of Class Imbalance Target Variable

- Challenges...



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Identify patterns in data via visualization, statistical analysis, and data mining

IST 718 Big Data Analytics

-IEEE-CIS Fraud Detection-

Base Neural Network

Base Model Hyperparameters:

- Kernel_initializer: normal
- Dense Input Layer:
 - Activation Function: '**relu**'
- Dense Output Layer:
 - Activation Function: '**softmax**'
- Compiler:
- loss: '**categorical_crossentropy**'
- optimizer: '**adam**'
- Built on:
- epochs: **300**
- batch_size: **1000**

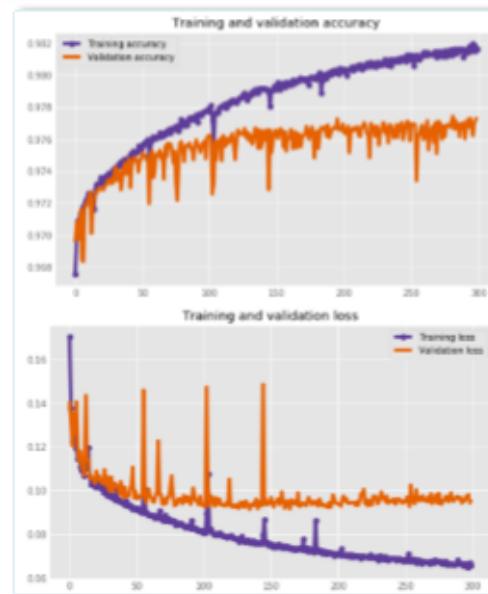
```
X_train shape: (330702, 205)
y_train shape: (330702, 2)
X_test shape: (118108, 205)
y_test shape: (118108, 2)
X_val shape: (141730, 205)
y_val shape: (141730, 2)
```



Model Summary:

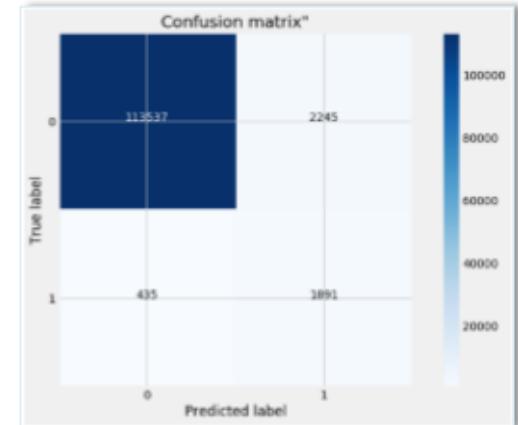
```
None
254520
Model: "sequential_1"

Layer (type)          Output Shape         Param #
dense_1 (Dense)      (None, 503)           253512
dense_2 (Dense)      (None, 2)              1008
Total params: 254,520
Trainable params: 254,520
Non-trainable params: 0
```



```
Test loss: 0.0959736775085113
Test accuracy: 0.9773089037192809
Baseline Error: 2.2691096280719023
```

	precision	recall	f1-score	support
Class0	0.98	1.00	0.99	113972
Class1	0.81	0.46	0.59	4136
accuracy			0.90	118108
macro avg	0.90	0.73	0.79	118108
weighted avg	0.97	0.98	0.97	118108



Identify patterns in data via visualization, statistical analysis, and data mining

IST 718 Big Data Analytics

-IEEE-CIS Fraud Detection-

Patterns in Exploratory Data Analysis and Neural Networks

The screenshot shows a Jupyter Notebook interface with the title "jupyter Timbrook_Fraud_ExploreCleanTransform (autosaved)". The notebook content is as follows:

Final Project

Syracuse Applied Data Science, IST-718 Big Data Analytics

Team: AUQ-42 Team Members:

- Ryan Timbrook
- Amanda Carvalho
- Luigi Penalosa
- Chikeung Cheung

DATE:
ASSIGNMENT: IEEE-CIS Fraud Detection (kaggle competition)

Business Question

Improve the efficacy of fraudulent transaction alerts, helping hundreds of thousands of businesses reduce their fraud loss and increase their revenue, while securing consumer's peace of mind and wallets!

Problems to solve

Identify real-time fraudulent e-commerce transactions, using advanced Machine Learning algorithms, by automating alerts that block highly suspicious activities.

Why the problem is important

Everyone who uses e-commerce technology and modern banking systems are at risk of being a victim of fraud. It costs both the individual as well as the merchant who offers refunds for fraudulent transactions, and not all scenarios are covered, leaving many individuals having to pay.

Chargebacks are a growing costly burden for merchants. By eliminating chargebacks, fines, and fees related to third-party fraud and unauthorized charges, the client, VESTA, is able to significantly reduce the operational costs and resources associated with complex chargeback management solutions and the specialized staff necessary for rapid, scalable business growth. This leaves all the cost risk on the client. Improving automated fraudulent detection technology will greatly reduce this cost.

About the Data

The core data set for this project is provided by VESTA, the world's leading payment service company, and is a kaggle competition being facilitated by the [IEEE Computational Intelligence Society](#).

VESTA

Predicting the probability that an online transaction is fraudulent, as denoted by the binary target `isFraud`.

The data is broken into two files `identity` and `transaction`, which are joined by `TransactionID`. *Not all transactions have corresponding identity information.*

Categorical Features - Transaction

- ProductCD
- card1 - card6
- addr1, addr2
- P_emaildomain
- R_emaildomain
- M1 - M9

Categorical Features - Identity

- DeviceType
- DeviceInfo

Learning Objective

-Identify patterns in data via visualization, statistical analysis, and data mining-

➤ Real Estate Property Investment



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

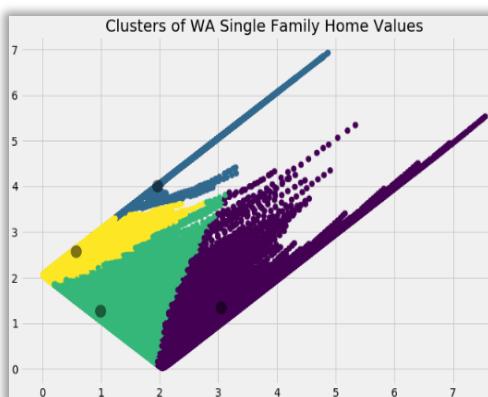
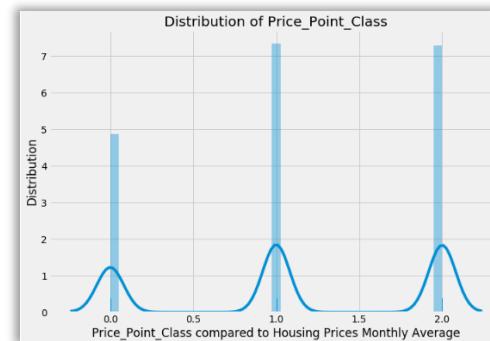
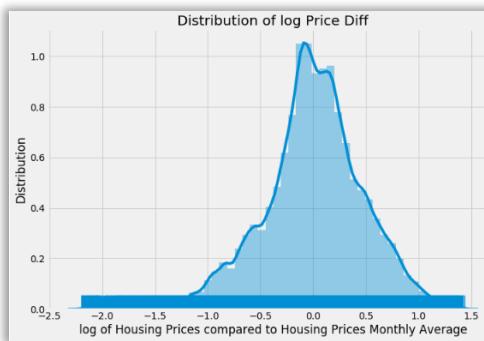
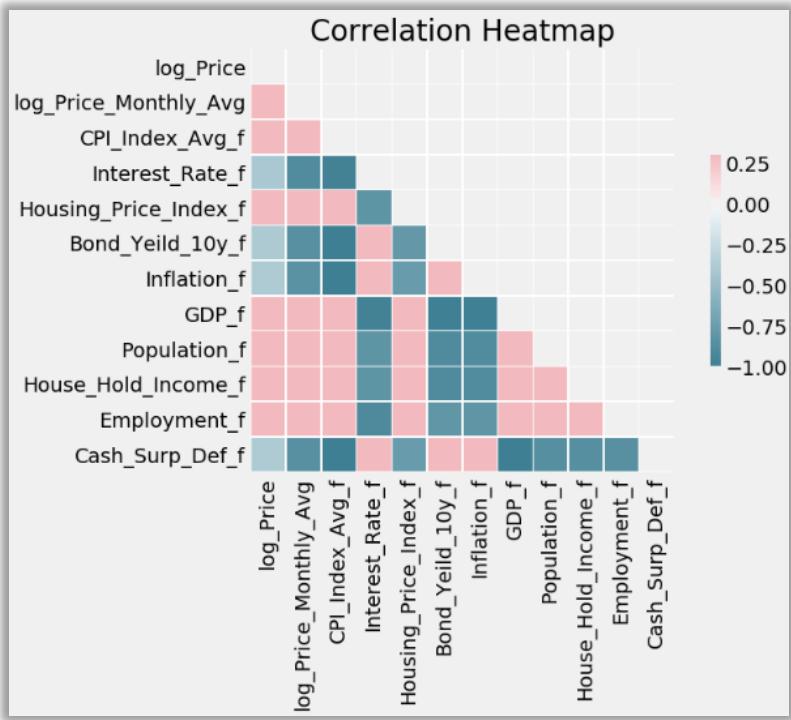
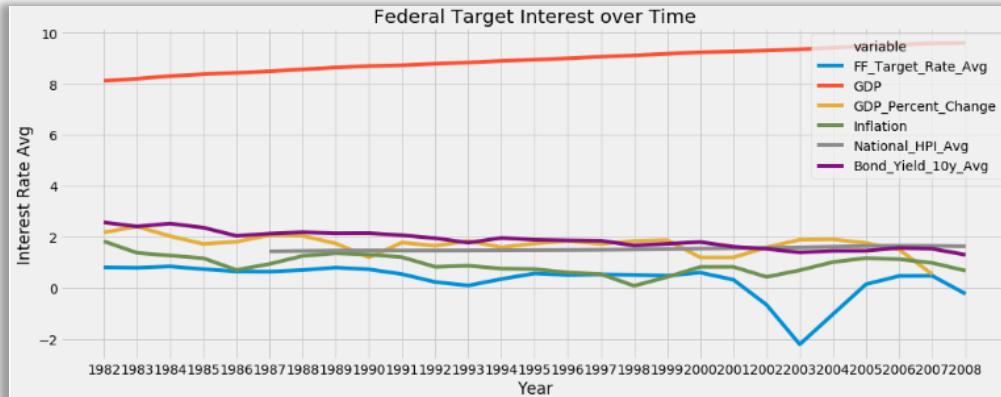
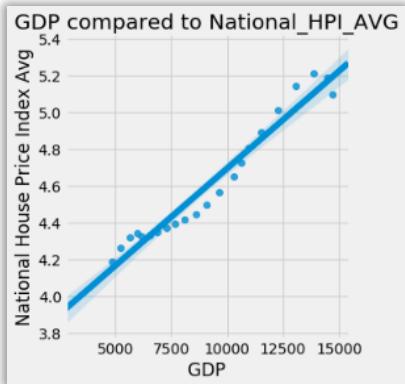
Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Identify patterns in data via visualization, statistical analysis, and data mining

IST 707 Data Analytics

-Real Estate Property Investment-

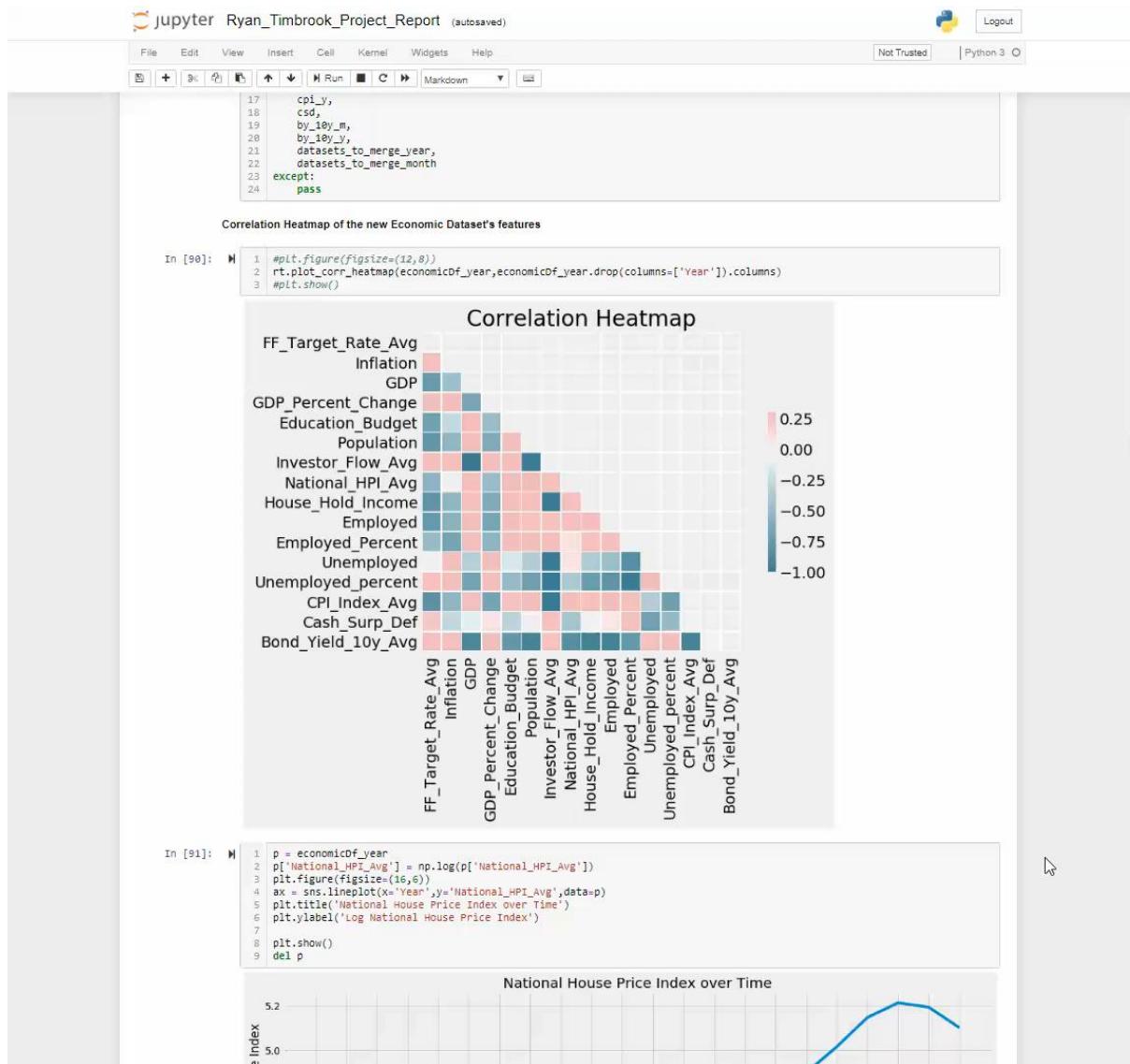


Identify patterns in data via visualization, statistical analysis, and data mining

IST 707 Data Analytics

-Real Estate Property Investment-

Patterns in Exploratory Data Analysis



-Develop alternative strategies based on the data- Overview

- Real Estate Property Investment
- Recruiting Advertising Strategy

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Learning Objective

- Develop alternative strategies based on the data-

➤ Real Estate Property Investment

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



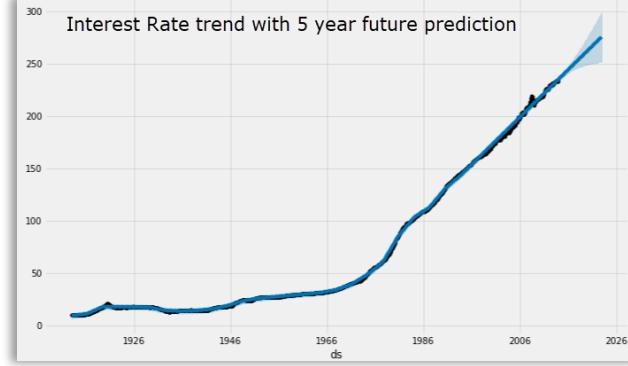
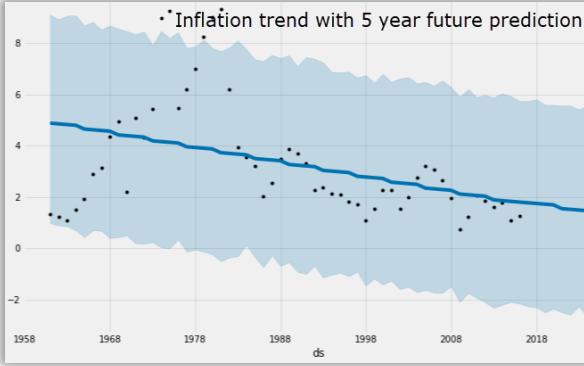
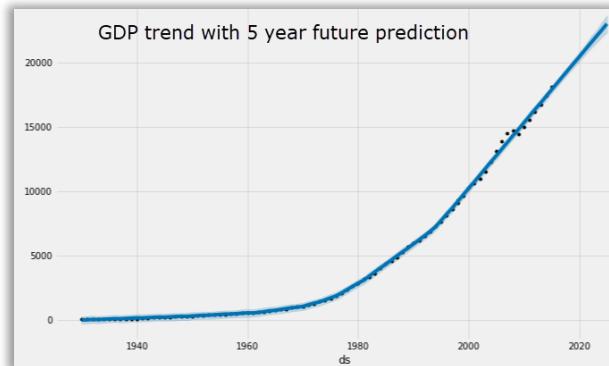
Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Develop alternative strategies based on the data

IST 707 Data Analytics
-Real Estate Property Investment-

- Original **Economic Reports** dataset **incomplete**, unusable – **Alternative** found at **Datahub.io**
- Time series date ranges **short of target year** – Future **predictive** values created (Facebook Prophet)
 - **GDP Year:** Forecasted from 2016, 2017, 2018
 - **Inflation:** Forecasted for 2017, 2018
 - **Interest Rates:** Forecasted for 2016, 2017, 2018



Learning Objective

- Develop alternative strategies based on the data-

➤ Recruiting Advertising Strategy

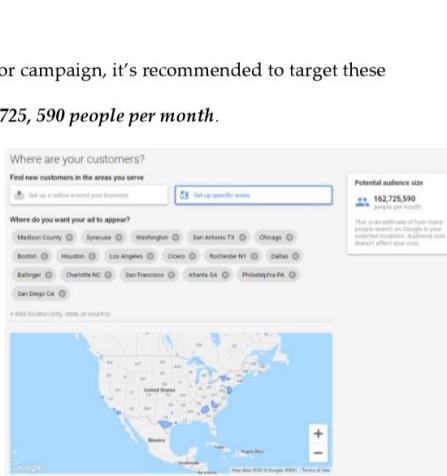
3.1 Recommendations

Advertising Regions:

- Taking the top ten US states by region for each prior campaign, it's recommended to target these regions for 2020.

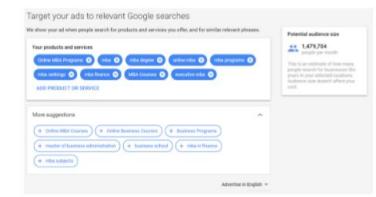
Google Ads estimates these US regions have a reach of 162, 725, 590 people per month.

- Syracuse, NY
- New York, New York
- Washington, District of Columbia
- San Antonio, TX
- Chicago, IL
- Boston, MA
- Huston, TX
- Los Angeles, CA
- Cicer, NY
- Rochester, NY
- Ballinger, TX
- Charlotte, NC
- San Francisco, CA
- Atlanta, GA
- Philadelphia, PA
- San Diego, CA



Keywords to use:

- Online MBA Programs
- Mba
- Mba degree
- Online mba
- Mba programs
- Mba rankings
- Mba finance
- Mba Courses
- Execute mba



4.1.1 Proposed 2020 Budget Strategy

Campaign	2020 Total Duration	Ads Runtime Iterations	Provider	Budget
Whitman.syr.ed	9 months	1 month trial, then 4x2mo. iterations	Google Ads	\$ 14,778.00
		1 month trial, then 4x2mo. iterations	Facebook	\$ 11,666.67
		1 month trial, then 4x2mo. iterations	GMASS	\$ 6,666.67
MBA Marketing - Full Time	9 months	1 month trial, then 4x2mo. iterations	Google Ads	\$ 14,778.00
		1 month trial, then 4x2mo. iterations	Facebook	\$ 11,666.67
MBA Marketing - IMBA	9 months	1 month trial, then 4x2mo. iterations	Google Ads	\$ 14,778.00
		1 month trial, then 4x2mo. iterations	Facebook	\$ 11,666.67
		1 month trial, then 4x2mo. iterations	GMASS	\$ 6,666.67
				\$ 99,334.00

9

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Develop alternative strategies based on the data

IST 707 Data Analytics

-Recruiting Advertising Strategy-

Based on the historical data and insights gained from analyzing the google analytics platform, using the following metrics: **Cost per Click, Users, New Users, Bounce Rate, Pages per session**, the **most effective** campaign identified was the **Whitman.syr.edu**.

United States Campaign Measurements											
Campaign	Start Date	End Date	Cost	Cost per Click	Users	New Users	Total Sessions	Bounce Rate	Page per Session		
whitman.syr.edu	2/15/2011	8/15/2011	\$ 37,851.36	\$ 3.93	133,732	98,379	192,134	48.08%	3.76		
MBA Marketing – Full-time	Jan/15/2012	2/15/2012	\$ 16,459.90	\$ 4.64	1,347	1,348	1,468	97.07%	1.05		
MBA Marketing – iMBA	2/1/2012	10/15/2012	\$ 144,971.70	\$ 7.46	2,753	2,748	3,079	88.96%	1.14		
Delta	9/1/2013	11/1/2013	\$ 10,000.00	\$ -			284	28.17%	3.37		

2020 Advertising Recruitment Recommendations:

- The 2020 Campaign recommendations include continued use of **Google Ad Marketing**, with the additional recommendations to incorporate **Facebook Ad Business** and **GMASS Targeting marketing**.
- The recommendations included in the report detail the **advertising regions** to focus their **ad campaigns** to, **keywords** to use in **search analytics**, the **best days of the week** and **time of the day** for the **advertising** to be **published**, the **expected advertising costs** which includes a **budget plan**, a post-implementation success measurement plan, and a **list of other factors** we found essential to the **success** of the upcoming **campaign**.

-Develop a plan of action to implement the business decisions derived from analysis-

Overview

- Purchase Order Process Improvement
- A4B KPI BizOps Organization Database

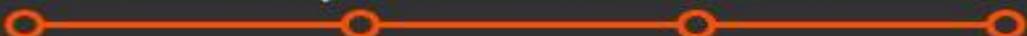
DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



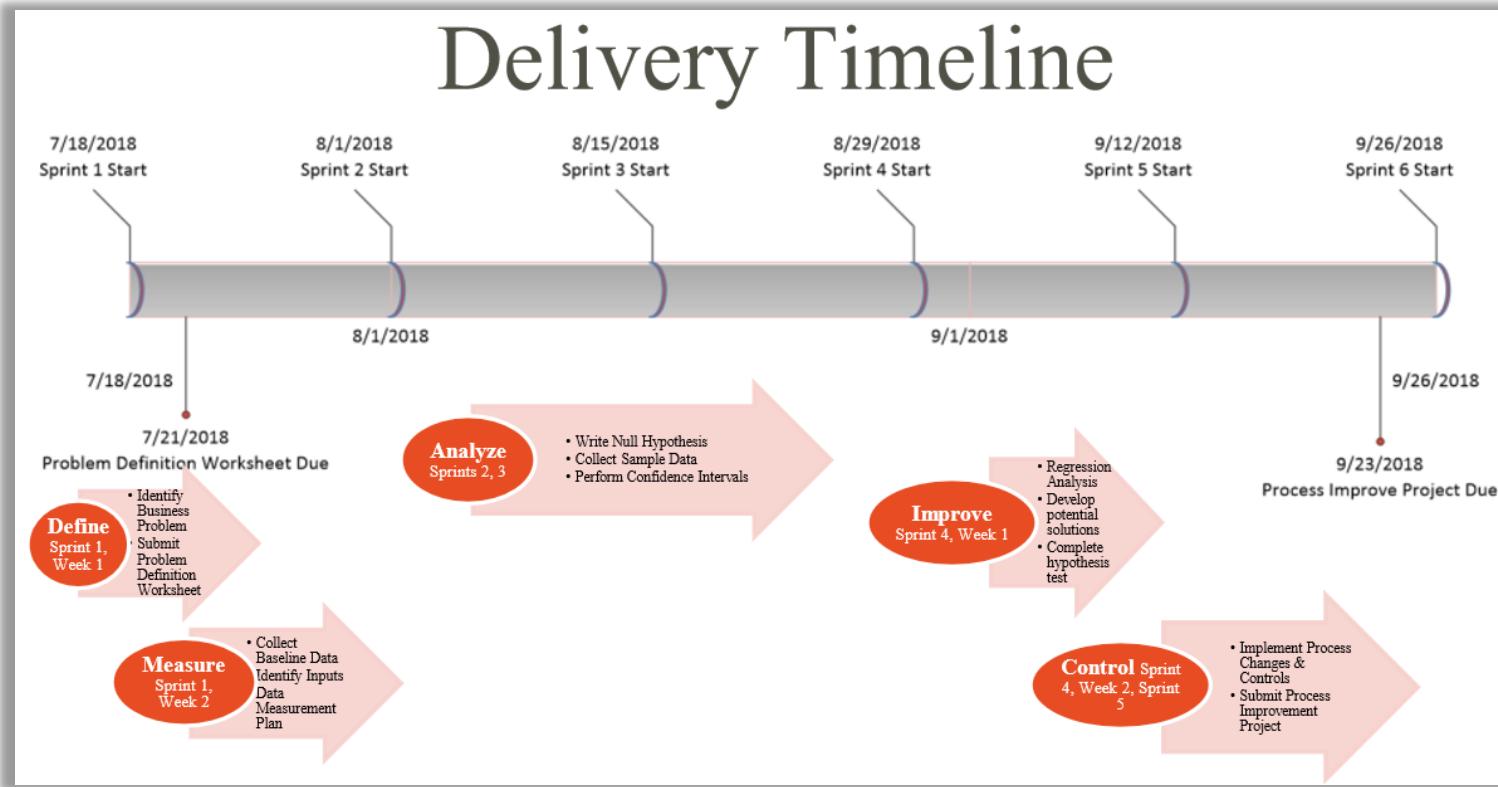
Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Learning Objective

- Develop a plan of action to implement the business decisions derived from analysis-

➤ Purchase Order Process Improvement



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Develop a plan of action

MBC 638 Data Analysis and Decision Making

-Purchase Order Process Improvement-

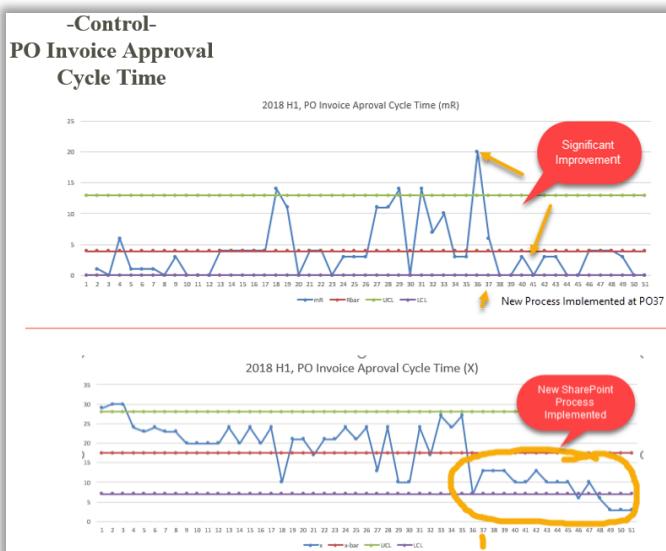
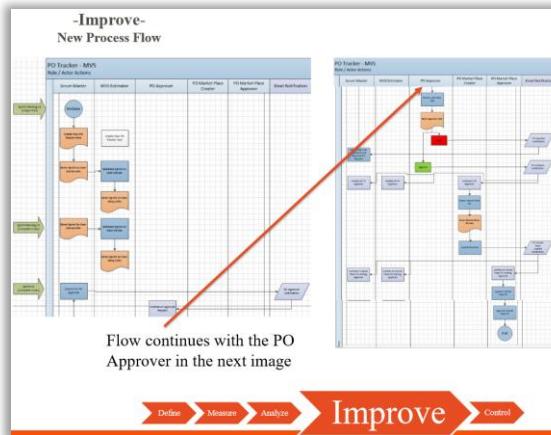
In the 'Improve' phase, a **plan was executed** that included the **design** and **pilot** of a **new process workflow** that incorporated automation through Microsoft SharePoint tooling. By designing a standard user interface for entering the necessary attributes needed for an accurate and timely PO validation and approval process along with a detailed cross-functional workflow diagram for role/resource training, an **improved cycle time efficiency of 70%** was observed during the **pilot of the new process**.

-Improve-
New SharePoint Workflow and Data Input Form

This form represents all of the attributes and data elements which need to be collected for timely validation of Purchase Orders submitted by the Vendor.

Prior to this new process, all of this data had to be manually tracked down by the Dev Manager before approving a PO in the Ariba Procurement System.

Define > Measure > Analyze > **Improve** > Control



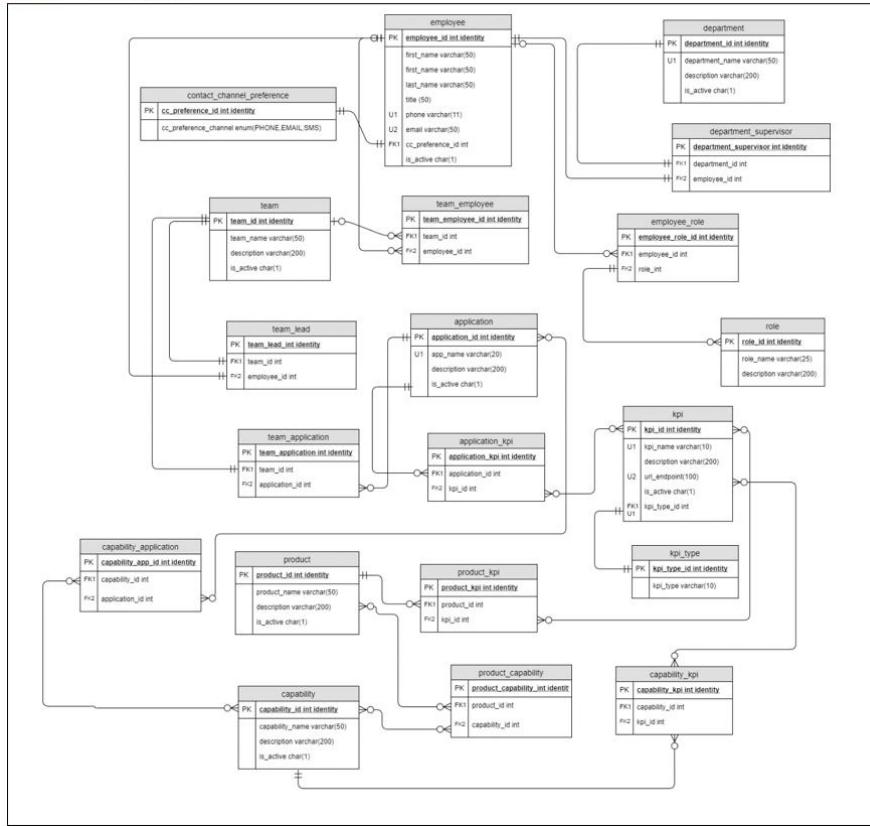
To **measure on-going success** and determine the business value of pushing this new process to other teams within the organization, **controls** were **implemented** that would **continue to measure and report cycle-times** along with POs being **approved** within the given discount thresholds.

Learning Objective

-Develop a plan of action to implement the business decisions derived from analysis-

➤ A4B KPI BizOps Organization Database

3.1.1 Diagram



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Develop a plan of action

IST 659 Database Admin and Management

-A4B KPI BizOps Organization Database-

- **Project Initiative Name:** VXPD A4B ‘KPI Buddy’ Alexa Skill Pilot
- **Team:** VXPD DevOps
- **Project Type:** Agile Sprint, 2-Week Cycles
- **Project Duration:** Three Sprint Cycles (6 Weeks)
- **Team Roles:** Product Owner, Scrum Master, Voice Experience Designer, DBA, Voice Experience Developers, DevOps SRE
- Taking the design and implementation strategy laid out in this project, I ran an internal Pilot initiative with my DevOps team at work to prototype this database with new Alexa KPI Skills my team was building. Leveraging our AWS Cloud Environment, I set up an Amazon RDS for SQL Server instance and implemented this BizOps Organization Database on AWS.
- The Pilot was a success, offering our internal customers (Product and Technology Organization) a simple voice interface, customized to their level of KPI needs.

-Demonstrate communication skills regarding data and its analysis-

Overview

- Purchase Order Process Improvement
- Public Sentiment Toward NFL Teams, Coaches, and Players
- IEEE-CIS Fraud Detection

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Learning Objective

-Demonstrate communication skills regarding data and its analysis-

➤ Purchase Order Process Improvement

Sprint Team, Purchase Order Invoice Approval Cycle Time Reduction

Process owner: Ryan Timbrook

Key Dates -->	Team Launch	7/18	Define	7/21	Measure	7/25	Analyze	8/1	Improve	8/29	Control	9/12
DEFINE												
Only 20% of Vendor supplied Sprint Teams Purchase Order (PO) Invoices are being approved within the discount threshold of 19 days.												
Not achieving this discount benchmark costs the company, on average, \$5,911 in lost revenue on each PO invoice it processes												
Managers and Teams spend to much time validating Sprint PO invoices for correctness. This time not only costs in resource dollars, it also has negative impacts to the amount of lead time delivery work they can complete each Sprint Cycle, leading to a decreased "Time to Market" performance measure												
MEASURE												
Managers have a low frequency rate of approving POs under threshold												
ANALYZE												
PO Invoices with mid-to-high User Story Counts are increasing the likelihood a PO Invoice will need to be corrected, which leads to a longer Approval Cycle Time												
IMPROVE												
New SharePoint Workflow = Efficiency												
CONTROL												
PO Invoices having to be corrected leads to longer Approval Cycle Times												
Solution → A New SharePoint workflow form and process												

Analyze - PO Invoice Approval Cycle Time

Hypothesis Statements

PO Invoices not approved within discount threshold
Is my average PO Approval process cycle time (avg = 25 days, std dev = 5.5) performing well versus goal (avg less than 19 days)?
Type of Test: One-Tail Test
Sample Size: n = 36
Test Statistic:
 $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
 $Z = \frac{25 - 19}{5.5/\sqrt{36}} = 2.17 > 0.89$

Test Statistic

One-Tail Test
Lower/left-tail
Sample Size: n = 36
Test Statistic:
 $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
 $Z = \frac{25 - 19}{5.5/\sqrt{36}} = 2.17 > 0.89$

Simple Linear Regression on PO Approval Cycle time output compared to User Stories Count

Hypothesis Statements

Number of User Stories vs. PO Invoice Approval Within Threshold
Is the number of User Stories for a given PO associated with PO invoices not being approved within our threshold time?

Sample Size: n = 36
Is the PO Approval Cycle Time is effected by the number of User Stories in the given PO Invoice
Is the PO Approval Cycle Time is effected by the number of User Stories in the given PO Invoice

...>>> p value is lower than our confidence alpha of .05, we reject the Null Hypothesis
...>>> PO Approval Cycle Time is effected by the number of User Stories in the given PO Invoice. The linear trend line shows a positive increase

Regression Statistics

Multiple R: 0.818630104
R Square: 0.670833333
Adjusted R Square: 0.634931347
Standard Error: 5.0372285
Observations: 36

Dependent Variable	Independent Variable	t Stat	P-value
User Stories Count	Intercept	2.34710482	2.79862412
User Stories Count	User Stories Count	4.27700000	0.00017616

2018 H1, PO Invoice User Story Count Simple Linear Regression Analysis

Number of User Stories vs. PO Invoice Approval Cycle Time (days)

User Stories Count

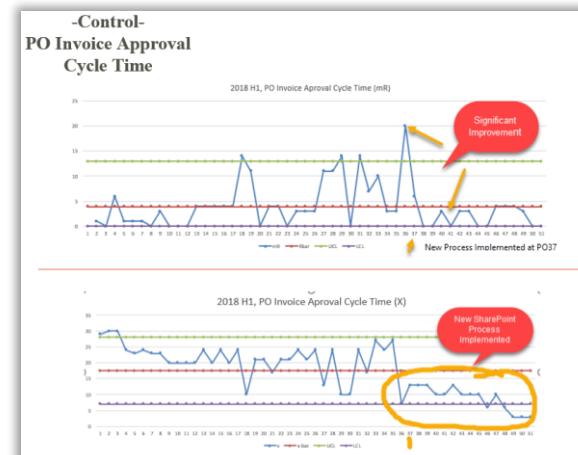
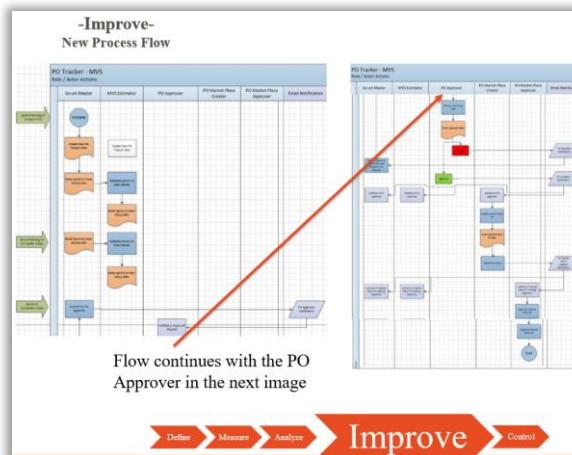
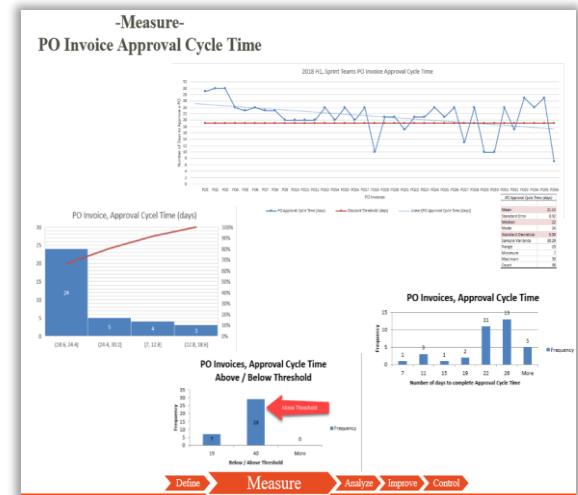
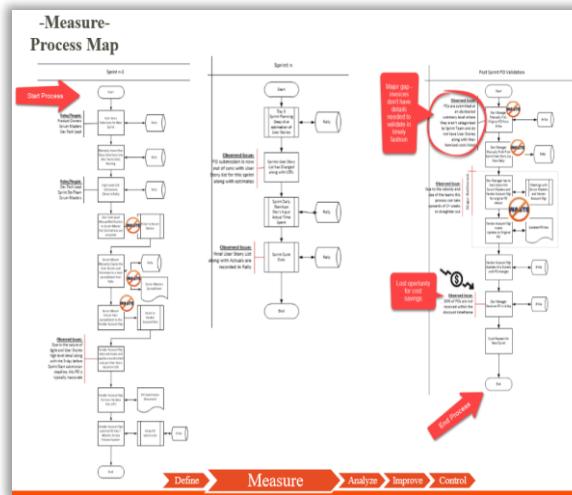
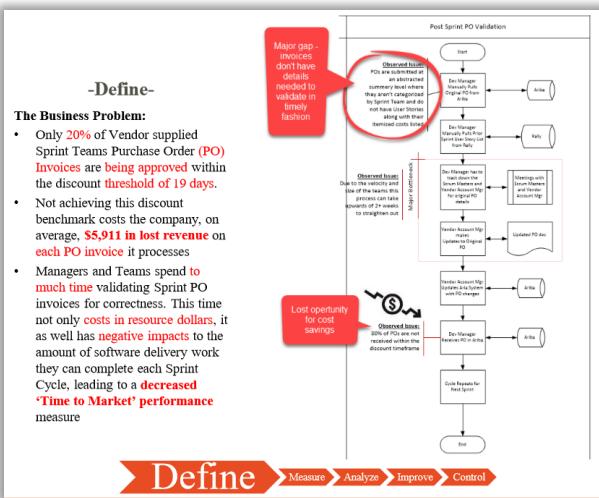
Greater User Stories Count

Define Measure Analyze Improve Control

Demonstrate communication skills regarding data and its analysis

MBC 638 Data Analysis and Decision Making

-Purchase Order Process Improvement-



Demonstrate communication skills regarding data and its analysis

MBC 638 Data Analysis and Decision Making

-Purchase Order Process Improvement-

Project Presentation Deck

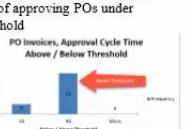
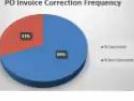
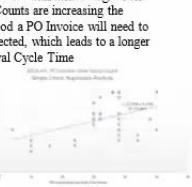
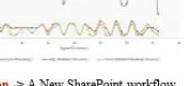
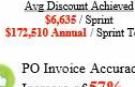
1 of 28 |

Sprint Team, Purchase Order Invoice Approval Cycle Time Reduction

Process owner: Ryan Timbrook



Key Dates --> Team Launch 7/18 **Define** 7/21 **Measure** 7/25 **Analyze** 8/1 **Improve** 8/29 **Control** 9/12

DEFINE	MEASURE	ANALYZE	IMPROVE	CONTROL
<p>Only 20% of Vendor supplied Sprint Teams Purchase Order (PO) Invoices are being approved within the discount threshold of 19 days.</p> <p>Not achieving this discount benchmark costs the company, on average, \$5,911 in lost revenue on each PO invoice it processes</p> <p>Managers and Teams spend to much time validating Sprint PO invoices for correctness. This time not only costs in resource dollars, it as well has negative impacts to the amount of software delivery work they can complete each Sprint Cycle, leading to a decreased 'Time to Market' performance measure</p>  	<p>Managers have a low frequency rate of approving POs under threshold</p>  <p>PO Invoices, Approval Cycle Time Above / Below Threshold</p>  <p>Mgr Validation Time with Goal Threshold</p>  <p>PO Invoices having to be corrected leads to longer Approval Cycle Times</p>	<p>PO Invoices with mid-to-high User Story Counts are increasing the likelihood a PO Invoice will need to be corrected, which leads to a longer Approval Cycle Time</p>  <p>PO Invoices vs Approval Cycle Time</p> <p>The more time a Team and Mgr have to spend on Validating a PO Invoice the more likely the PO Approval Cycle Time will increase</p>  <p>PO Invoice Correction Frequency</p>	<p>New SharePoint Workflow = Efficiency</p>  <p>PO Invoice Approval Cycle Time</p> <p>Avg Discount Achieved \$6,635 / Sprint Team \$172,510 Annual / Sprint Team</p>  <p>PO Invoice Accuracy Increase of 57%</p>	<p>Control charts show a significant improvement in Approval Cycle Time starting at PO37, which was when the new SharePoint Workflow process was implemented</p>  <p>PO37 - New SharePoint Workflow Implemented</p>

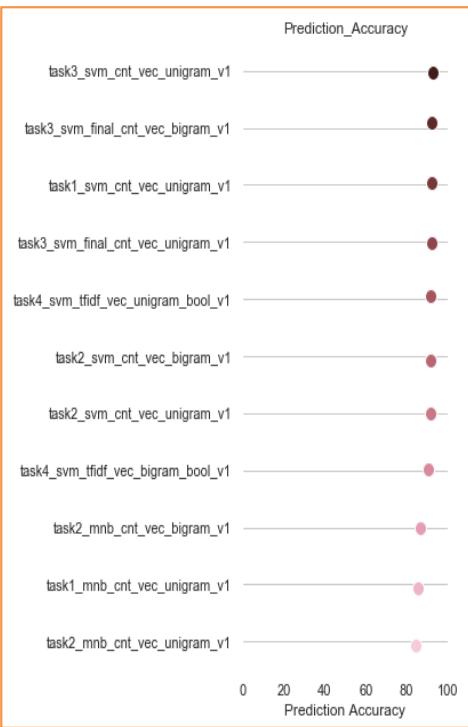
TEAM MEMBERS

Ryan Timbrook

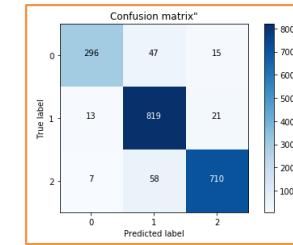
Learning Objective

- Demonstrate communication skills regarding data and its analysis-

➤ Public Sentiment Toward NFL Teams, Coaches, and Players



	precision	recall	f1-score	support
negative	0.94	0.83	0.88	358
neutral	0.89	0.96	0.92	853
positive	0.95	0.92	0.93	775
micro avg	0.92	0.92	0.92	1986
macro avg	0.92	0.90	0.91	1986
weighted avg	0.92	0.92	0.92	1986



Model_Type	Vectorizer	N_Gram	Experiment	Model_Nam	Cross_Fold	Prediction_Accuracy	Total_Prediction_Point	Test_Recall_Score_Avg	Test_Precision_Score_Avg	Train_Recall_Score_Avg	Train_Precision_Score_Avg	Total_Build_Time	Total_Predict_Time	Confusion_Matric
LinearSVM	count	unigram	task3_svm_cnt_vec_unigram		10	92.42	1953	0.9133	0.9312	0.9877	0.9923	7.9983	0.001	[[283 37 22] [9 788 22] [9 49 735]]
LinearSVM	count	bigram	task3_svm_final_cnt_vec_bigr		10	91.4	1953	0.905	0.9218	0.9932	0.996	7.389	0.0017	[[29 40 22] [9 770 32] [9 23 522]]
LinearSVM	count	unigram	task2_svm_cnt_vec_unigram		10	90.99	3905	0.8963	0.9165	0.9951	0.9967	7.3672	0.0014	[[29 535 56] [578 77 56] [23 5522 60]]
LinearSVM	count	unigram	task1_svm_cnt_vec_unigram		10	90.73	3905	0.9003	0.9189	0.9954	0.9972	5.5715	0.0008	[[27 119 1443] [584 81 52] [26 3558 66]]
LinearSVM	count	bigram	task2_svm_cnt_vec_bigram_		10	90.6	3905	0.8956	0.9141	0.9969	0.9979	9.7347	0.002	[[22 120 1396] [532 67 64] [20 541 99]]
LinearSVM	tfidf	bigram	task4_svm_tfidf_vec_bigr		10	89.73	3905	0.8853	0.9138	0.9697	0.9795	5.126	0.0022	[[21 130 1411] [576 79 56] [30 4488 98]]
LinearSVM	tfidf	unigram	task4_svm_tfidf_vec_unigr		10	89.3	3905	0.8755	0.9043	0.9588	0.9733	5.4828	0.0014	[[13 342 1423] [562 68 87] [54 1410 204]]
Multinomial_Naive_Bayes	count	bigram	task2_mnb_cnt_vec_bigr		10	84.1	3905	0.8443	0.8478	0.9047	0.9099	0.1792	0.0015	[[539 66 94] [59 1410 201] [544 86 80]]
Multinomial_Naive_Bayes	count	unigram	task1_mnb_cnt_vec_unigram		10	84.07	3905	0.822	0.8274	0.8821	0.8866	0.1242	0.0024	[[72 140 1343] [65 1361 233] [53 134 1349]]
Multinomial_Naive_Bayes	count	unigram	task2_mnb_cnt_vec_unigram		10	83.33	3905	0.8226	0.8262	0.886	0.8919	0.0947	0.0011	[[53 134 1349]]

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

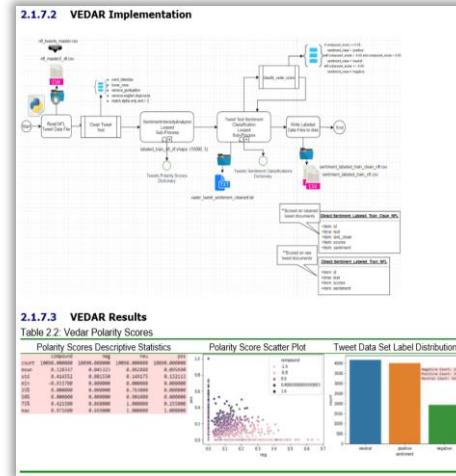
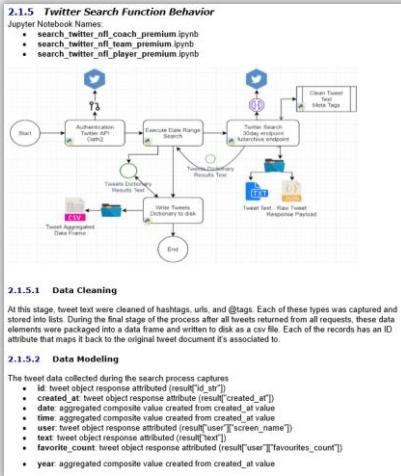
Demonstrate communication skills regarding data and its analysis

IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players-

Types of Communication:

- Final Report (word document) – written as academic paper
- Final Presentation (power point deck) – live presentation summary of results and insights learned
- Autogenerated Model Performance Reports
- Source Code, Jupyter Notebooks Markup Language Comments



6 Conclusion

The most time-intensive, challenging, yet most critical aspect of this research is in mining the data. To be able to accurately apply machine learning algorithms to this type of unstructured, continuous-timeseries data for the purposes of trend analysis, mechanisms have to first be in place that can search social media sources like Twitter for relevant tweets on a daily and or historical bases. Setting up these programs should be well thought out and designed such that the data, in its various forms, is thought of as the final product. How the data is stored, cleaned and propagated through its pipeline will determine the quality of analysis and experimentation that can be done on it.

Technologies such as VADER (Hutto & Gilbert, 2014) for sentiment scoring, the Natural Language Toolkit (NLTK) (Bird, Loper, & Klein, 2009) for natural language processing and SciKit-Learn (Pedregosa, et al., 2011) for general machine learning have helped to commoditize these disciplines. The techniques made available to the average programmer today offer a wide range of options to attempt to extract value from data that is available either freely (albeit sometimes in limited quantities) or at a known subscription fee.

The National Football League (NFL) remains one of the most popular sports in America (Norman, 2018) and Fantasy Football has grown by leaps and bounds. In this age, popularity almost inevitably leads to an explosion of unstructured data across social networks such as Facebook and Twitter. This freely provided user opinion when correctly gathered and curated can provide a wealth of information about how those who follow both the NFL and fantasy sports feel about a particular team, coach, or player at any given week, especially during the active season. This data can be collected, and the sentiment of the fan base can be measured and tracked over time.

It is possible to look at this same user-provided data to determine the subjects about which changes to user sentiment are driven. The addition of topic modeling to this data may be a useful tool for the purposes of marketing. Not only could the NFL determine what the socially vocal fans are discussing, but it can determine the sentiment within those categories. The implications of this may hold well from a fantasy perspective as well, as these topics/sentiments can be used for assessing the cost of adding players to a fantasy team in games that are week to week.

The validity of predicting player performance based on sentiment is, at the moment, inconclusive. While there is undoubtedly immediate benefit from analyzing social media sentiment for the NFL (it can be used to determine marketing strategy), nothing in the presented analysis can conclusively determine that a player's performance can be obtained from the consciousness of the 'Twittersphere.' Future efforts to determine the true validity of using sentiment in the prediction of a players' performance need to include data from each of the fantasy football service providers. Each provider uses different statistical methods and AI to forecast player performance. While this analysis did not show immediate promise based on the limited slice of data used, it is possible that the addition of new data sources and the altering of methods may lead sentiment to be a variable in predictive models.

2.1.5.2 Data Modeling

The following code shows the steps for the sentiment pipeline:

```

# Read in the raw data for the sentiment pipeline
# id, tweet object response attribute (result["id"])
# created_at, tweet object response attribute (result["created_at"])
# date aggregated composite value created from created_at, value
# time aggregated composite value created from created_at, value
# user, user object response attribute (result["user"])
# screen_name, user object response attribute (result["user"]["screen_name"])
# text, tweet object response attribute (result["text"])
# favorite_count, tweet object response attribute (result["user"]["favourites_count"])
# year, aggregated composite value created from created_at, value

```

2.1.5.3 Model Results

Output confusion matrix, precision and recall values for the Sentiment training data.

```

# Output confusion matrix, precision and recall values for the Sentiment training data
# Build a simple NB model and a logistic SVM model
# Print the top 10 related words for the most positive category and the most negative category from the NB and SVMs models respectively
# Visualize the confusion matrix and the classification report for both models and explain why
# Report the confusion matrix, precisions, and recalls to see whether you model performed equally well on all categories, or some categories turn out to be easier or more difficult to build or train
# Run more tests on NB or SVMs
# Report your increased insight along with your report

```

Sentiment Model Results

```

In [24]: # Building a report off of pd.DataFrame()
1  summary_report_MF_model_top1 = model_top1
2  summary_report_MF_model_top1['model_type'] = 'Multinomial Naive Bayes'
3  summary_report_MF_model_top1['category'] = 'positive'
4  summary_report_MF_model_top1['precision'] = 0.85
5  summary_report_MF_model_top1['recall'] = 0.85
6  summary_report_MF_model_top1['f1_score'] = 0.85
7  summary_report_MF_model_top1['accuracy'] = exp_model_names
8  summary_report_MF_model_top1['dataset_version'] = dataset_versions
9  summary_report_MF_model_top1['prediction_accuracy'] = prediction_accuracy
10 summary_report_MF_model_top1['confusion_matrix'] = confusion_matrix
11 summary_report_MF_model_top1['precision_recall'] = precision_recall
12 summary_report_MF_model_top1['f1_scores'] = f1_scores
13 summary_report_MF_model_top1['test_recall_scores'] = test_recall_scores
14 summary_report_MF_model_top1['train_recall_scores'] = train_recall_scores
15 summary_report_MF_model_top1['train_recall_scores'] = train_recall_scores
16 summary_report_MF_model_top1['total_build_time'] = total_build_time
17 summary_report_MF_model_top1['total_train_time'] = total_train_time
18 summary_report_MF_model_top1['confusion_matrix'] = confusion_matrix
19 summary_report_MF_model_top1['accuracy'] = accuracy
20 summary_report_MF_model_top1['precision'] = precision
21 summary_report_MF_model_top1['recall'] = recall

```

Model Type Category Vectorizer N_Gram Experiment_Model_Name Dataset_Version Cross_Fold Prediction_Accuracy

Model Type	Category	Vectorizer	N_Gram	Experiment_Model_Name	Dataset_Version	Cross_Fold	Prediction_Accuracy
1 Multinomial_Naive_Bayes	Naive_Bayes	tfidf	unigram	test_tf_idf_multinomial_nb_g1	v1	10	88.80
1 Multinomial_Naive_Bayes	Naive_Bayes	tfidf	unigram	test_tf_idf_multinomial_nb_g1	v1	10	82.04
2 Multinomial_Naive_Bayes	Naive_Bayes	tfidf	unigram	test_tf_idf_multinomial_nb_g1	v1	10	84.29
3 Multinomial_Naive_Bayes	Naive_Bayes	tfidf	unigram	test_tf_idf_multinomial_nb_g1	v1	10	88.33
4 LinearSVM	LinearSVM	tfidf	unigram	test_tf_idf_linear_svm_g1	v1	10	91.87
5 LinearSVM	LinearSVM	tfidf	unigram	test_tf_idf_linear_svm_g1	v1	10	91.82
6 LinearSVM	LinearSVM	tfidf	unigram	test_tf_idf_linear_svm_g1	v1	10	92.95
7 LinearSVM	LinearSVM	tfidf	unigram	test_tf_idf_linear_svm_g1	v1	10	82.40
8 LinearSVM	LinearSVM	tfidf	unigram	test_tf_idf_linear_svm_g1	v1	10	81.89
9 LinearSVM	LinearSVM	tfidf	unigram	test_tf_idf_linear_svm_g1	v1	10	81.82

Multinomial Naive Bayes Classification (NB)

Python package sklearn.mnsmv2.1.3 sklearn_naive_bayes.MultinomialNB

Naive Bayes classifier for multinomial models

The multinomial Naive Bayes classifier is suitable for classification of discrete features (e.g. word counts for text classification). The multinomial distribution is notably capable of handling sparse input data well. However, in practice, fractional counts work as well as integer counts.

Read more in the [User Guide](#).

```

1  token_pattern = r'\b\w+\b'
2
3  # Initialize a unigram vectorizer
4  vectorizer = TfidfVectorizer(tokenizer=tokenize, stop_words='english', lowercase=True, binary=False, max_features=1000, max_df=1.0, min_df=1.0, ngram_range=(1, 1))
5
6  # Create a 2D matrix of feature vectors
7  X_train_tf_idf_unigram = vectorizer.fit_transform(df_train['text'].values)
8  X_test_tf_idf_unigram = vectorizer.transform(df_test['text'].values)
9
10 # Train and predict
11 nb_clf = MultinomialNB()
12 nb_clf.fit(X_train_tf_idf_unigram, y_train)
13 nb_clf.score(X_test_tf_idf_unigram, y_test)
14
15 # Compute accuracy
16 nb_clf_accuracy = nb_clf.score(X_test_tf_idf_unigram, y_test)
17
18 # Export results
19 nb_clf_report = classification_report(y_test, nb_clf.predict(X_test_tf_idf_unigram))
20
21 # Importance of naive bayes model
22 nb_cm = confusion_matrix(y_test, nb_clf.predict(X_test_tf_idf_unigram))
23
24 # Write to file
25 nb_cm.to_csv('nb_cm.csv')

```

2.1.7.4 Model Results

Output confusion matrix, precision and recall values for the Sentiment training data.

```

# Output confusion matrix, precision and recall values for the Sentiment training data
# Build a simple NB model and a logistic SVM model
# Print the top 10 related words for the most positive category and the most negative category from the NB and SVMs models respectively
# Visualize the confusion matrix and the classification report for both models and explain why
# Report the confusion matrix, precisions, and recalls to see whether you model performed equally well on all categories, or some categories turn out to be easier or more difficult to build or train
# Run more tests on NB or SVMs
# Report your increased insight along with your report

```

Model Type Category Vectorizer N_Gram Experiment_Model_Name Dataset_Version Cross_Fold Prediction_Accuracy

Model Type	Category	Vectorizer	N_Gram	Experiment_Model_Name	Dataset_Version	Cross_Fold	Prediction_Accuracy
1 Multinomial_Naive_Bayes	Naive_Bayes	tfidf	unigram	test_tf_idf_multinomial_nb_g1	v1	10	88.80
1 Multinomial_Naive_Bayes	Naive_Bayes	tfidf	unigram	test_tf_idf_multinomial_nb_g1	v1	10	82.04
2 Multinomial_Naive_Bayes	Naive_Bayes	tfidf	unigram	test_tf_idf_multinomial_nb_g1	v1	10	84.29
3 Multinomial_Naive_Bayes	Naive_Bayes	tfidf	unigram	test_tf_idf_multinomial_nb_g1	v1	10	88.33
4 LinearSVM	LinearSVM	tfidf	unigram	test_tf_idf_linear_svm_g1	v1	10	91.87
5 LinearSVM	LinearSVM	tfidf	unigram	test_tf_idf_linear_svm_g1	v1	10	91.82
6 LinearSVM	LinearSVM	tfidf	unigram	test_tf_idf_linear_svm_g1	v1	10	92.95
7 LinearSVM	LinearSVM	tfidf	unigram	test_tf_idf_linear_svm_g1	v1	10	82.40
8 LinearSVM	LinearSVM	tfidf	unigram	test_tf_idf_linear_svm_g1	v1	10	81.89
9 LinearSVM	LinearSVM	tfidf	unigram	test_tf_idf_linear_svm_g1	v1	10	81.82

Demonstrate communication skills regarding data and its analysis

IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players-

Project Presentation Deck

The screenshot shows a presentation slide with the following content:

Public Sentiment Toward the NFL
Can it predict weekly Fantasy Football outcomes?

• Ryan Timbrook
• Diego Vales
• David Madsen

SCHOOL OF INFORMATION STUDIES
SYRACUSE UNIVERSITY

ischool.syr.edu

The slide is numbered 18 in the top left corner. A vertical navigation bar on the left lists slide numbers from 1 to 18, with slide 1 highlighted.

Demonstrate communication skills regarding data and its analysis

IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players-

Project Final Report

The screenshot shows a Microsoft Word document window. The title bar indicates the file is located at `file:///G:/My%20Drive/Timbrook_Portfolio_Milestone/ist736_text_mining/final_project/05-Report/Final_Project_Timbrook_Ryan.docx`. The document content is as follows:

2019-1002 IST 736
Text Mining

Final Project

**Public Sentiment Toward NFL Team, Coach, Player
Can it predict weekly Fantasy Football outcomes?**

Ryan Timbrook
David Madsen
Diego Vales

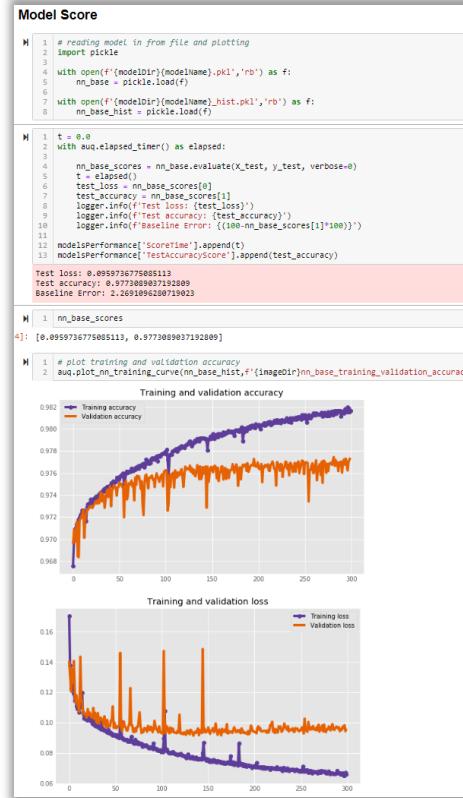
Course: IST 736 Text Mining
Term: Fall, 2019

Learning Objective

- Demonstrate communication skills regarding data and its analysis-

➤ IEEE-CIS Fraud Detection

The poster features the iSchool at Syracuse University logo (a yellow square with a black atom symbol and the number 42) on the left. In the center, the title "IEEE-CIS FRAUD DETECTION" is displayed in white capital letters. Below the title, the names of the team members are listed: Ryan Timbrook, Amanda Carvalho, Luigi Penalosa, and Charles Cheung. At the bottom right, it says "School of Information Studies SYRACUSE UNIVERSITY" and includes the website "ischool.syr.edu".



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Demonstrate communication skills regarding data and its analysis

IST 718 Big Data Analytics

-IEEE-CIS Fraud Detection-

Types of Communication:

- Final Presentation (power point deck) – live presentation summary of results and insights learned
- Autogenerated Model Performance Reports
- Source Code, Jupyter Notebooks Markup Language Comments

Business Question and Problem to Solve

Business Question

- Improve the efficacy of fraudulent transaction alerts, helping hundreds of thousands of businesses reduce their fraud losses and increase their revenue while securing consumer's peace of mind and wallets!

Problem to Solve

- Identify real-time fraudulent e-commerce transactions, using advanced Machine Learning algorithms, by automating alerts that block highly suspicious activities.

About the Data:

- The core data set for this project is provided by VESTA, the worlds leading payment service company, and is a kaggle competition being facilitated by the ISCHOOL.
- The data is broken into two files **identity** and **transaction**, which are joined by **TransactionID**. Not all transactions have corresponding identity information.

DATA SCIENCE AT THE ISCHOOL AT SYRACUSE UNIVERSITY

Get the  Data Science at the ISchool at Syracuse University

Data Details

Transaction Table

Variables in this table are items or other gift-giving goods and services, like you booked a ticket for others, etc.

- TransactionID: Unique from a given reference timeline (not an actual timestamp)
- ProductCD: Product name paid amount in USD
- ProductCD: product code, the product amount for each transaction.
- amt: total amount of payment card information, such as card type, card category, issue bank, country, etc.
- addr1: address
- addr2: address
- P_ and R_ emaildomain: purchaser and recipient email domain
- C1-C4: counting, such as how many addresses are found to be associated with the payment card, etc.
- trans_type: payment method, such as direct debit payment transaction, etc.
- trans_category: merchant category, such as food and drink, etc.
- V1-V49: Vesta engineered rich features, including ranking, and other entity relations.

339 Features - No data definitions provided due to ILP

Categorical Features - Transaction:

- ProductCD
- card1 card2
- addr1(addr2)
- P_emaildomain
- R_emaildomain
- M1-M9

Categorical Features - Identity:

- DeviceType
- DeviceInfo
- id_12_id_38

Training dataset:

- Rows: 590,548
- Columns: 434
- Missing Values: 414
- Total NaN count: 115,523,073

Testing dataset:

- Rows: 500,632
- Columns: 433
- Missing Values: 414
- Total NaN count: 90,186,908

For categorical features we will apply OneHot transformation, but only for most common values for each feature to reduce sparsity. Also there is an embedding approach for categorical features transformation. It was implemented in this kernel <https://www.kaggle.com/mirychevskeras-nn-starter-v-time-series-sgd>

With embedding approach I didn't get any significant improvement comparing to this.

```
In [0]: M 1 # categorical_features = []
2 continuous_features = []
3 for c in train.columns:
4     if str(train[c].dtype) == 'category':
5         categorical_features.append(c)
6     else:
7         if not (c == 'isFraud' or c == 'TransactionID'):
8             continuous_features.append(c)

In [0]: M 1 # hold out test data
2 X = train.drop(columns=['TransactionID', 'isFraud'])
3 y = train['isFraud']
4
5 # split the data 1 of 2
6 X_train, y_train = train_test_split(X, y, test_size=0.2, random_state=13)
7
8 logger.info('X shape: %s', X.shape)
9 logger.info('y shape: %s', y.shape)
10 logger.info('X_train shape: %s', X_train.shape)
11 logger.info('y_train shape: %s', y_train.shape)
12
13 X shape: (472432, 28)
y shape: (472432,)
X_train shape: (118108, 28)
y_train shape: (118108,)
```

Build a model consisting of multiple dense layers...

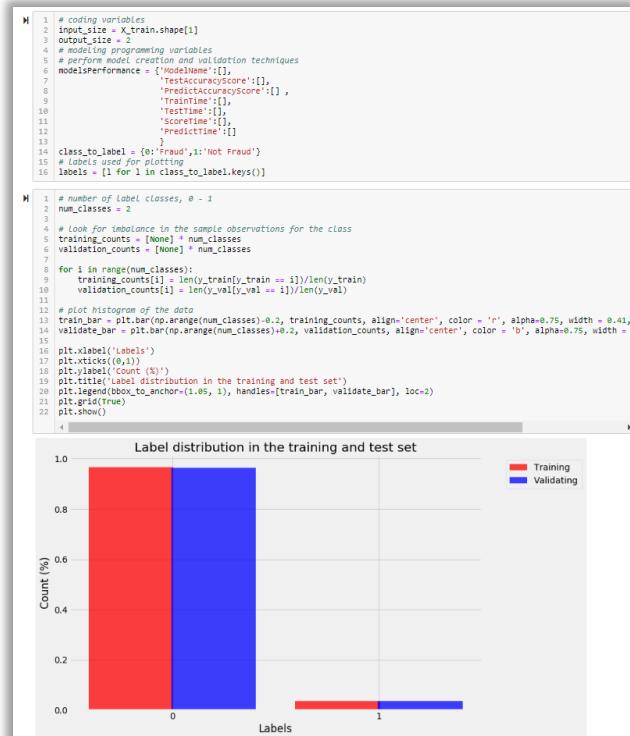
We'll use the Keras sequential API.

Then we will add the hidden layers to the model using `model.add()`. For each `Dense()` layer, you can specify its name, the number of units, its activation function, etc.

Input layer - Keras takes a simple approach and defines it together with the first hidden layer via the parameter `input_dim` or `input_shape`.

```
M 1 # Model - Build the baseline
2 def baseline_model():
3     # Create model
4     model = Sequential()
5
6     model.add(Dense(input_dim, input_dim=input_size, kernel_initializer='normal', activation='relu'))
7     model.add(Dense(output_size, kernel_initializer='normal', activation='softmax'))
8
9     # Compile model loss - categorical_crossentropy - used for identifying multiple images
10    model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
11
12 return model

M 1 m_base = baseline_model()
2 logger.info(m_base.summary())
3 logger.info(m_base.count_params())
None
42398
Model: "sequential_1"
Layer (type)      Output Shape     Param #
=====
dense_1 (Dense)   (None, 98)      42398
=====
dense_2 (Dense)   (None, 2)       412
=====
Total params: 42,810
Trainable params: 42,810
Non-trainable params: 0
```



Demonstrate communication skills regarding data and its analysis

IST 718 Big Data Analytics -IEEE-CIS Fraud Detection-

Project Presentation Deck

The slide is a presentation slide with the following elements:

- Header:** A yellow rectangular icon containing a black atom symbol with the number "42" in the center.
- Title:** **IEEE-CIS FRAUD DETECTION** (in large white capital letters)
- Team Members:** A list of four names: **Ryan Timbrook**, **Amanda Carvalho**, **Luigi Penaloza**, and **Charles Cheung**.
- Syracuse University Logo:** A stylized orange "S" icon.
- Syracuse University Text:** **SCHOOL OF INFORMATION STUDIES** and **SYRACUSE UNIVERSITY** (both in orange).
- Footer:** The URL ischool.syr.edu (in orange).

The left side of the slide shows a vertical navigation bar with numbered items from 1 to 17, each accompanied by a small preview icon.

-Synthesize the ethical dimensions of data science practice- Overview

- A4B KPI BizOps Organization Database
- IEEE-CIS Fraud Detection

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Learning Objective

-Synthesize the ethical dimensions of data science practice-

- A4B KPI BizOps Organization Database

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Synthesize the ethical dimensions of data science practice

IST 659 Database Admin and Management
-A4B KPI BizOps Organization Database-

- IAM Rules – ‘least privilege’
- Hierarchical Relationship Model Design
- SQL Database Role-Based Permissions

Learning Objective

-Synthesize the ethical dimensions of data science practice-

- IEEE-CIS Fraud Detection

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

Synthesize the ethical dimensions of data science practice

IST 718 Big Data Analytics
-IEEE-CIS Fraud Detection-

- Machine Learning Algorithms to identify real-time fraudulent transactions
 - Automatic alerting and blocking of highly suspicious activities
 - Protect the people and the banking systems
-
- Just about **everyone** uses **e-commerce technology** and modern banking systems are at risk of being a victim of fraud.
 - It cost both the individual as well as the merchant who offers refunds for fraudulent transactions; and not all scenarios are covered, leaving many individuals to pay.
 - Company's who have had data security breaches put everyone who uses electronic forms of payment at risk.

A few of the larger breaches such as:

- **eBay in 2014 with 145 million user accounts compromised**
- **Heartland Payment Systems in 2008 had 134 million users credit cards stolen**
- **Target in 2013 had up to 110 million customers credit/debit card**
- **Yahoo in 2014 with it's 1 billion user accounts and passwords compromised**

Thank you!

MS Applied Data Science program at Syracuse University School of Information Studies

Learning Outcome Objectives Covered:

- ❖ Collect and Organize Data
- ❖ Identify Patterns in data via visualization, statistical analysis, and data mining
- ❖ Develop alternative strategies based on the data
- ❖ Develop a plan of action to implement the business decisions derived from analysis
- ❖ Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals
- ❖ Synthesize the ethical dimensions of data science practice



About Me: Microsoft Word Document

RYAN TIMBROOK

Seattle, WA 98107
ryan.timbrook@gmail.com

206.516.9956

<https://github.com/timbro185>

SR. LEADER IN SOFTWARE ENGINEERING

Hands-on Technologist Leading High-velocity Product Development Teams That Exceed Quality Delivery Expectations

A top-performing Leader in Software Engineering with 21+ years of Product Development experience combining expert SDLC management skills with an extensive background in Enterprise Software Development, Speech/Text -- Natural Language Technologies, and Applied Data Science. Directs multiple projects/teams simultaneously; Hands-on technologist with a proven track record of leading multi-disciplinary, high-velocity, product development teams that exceed quality delivery expectations through Agile DevOps implementation methodologies. Relentless focus on solutions and continuous improvement opportunities. Display genuine expertise in business and technology, thus excelling in maintaining alignment of business and technology teams to achieve all business objectives — strong analytical, organizational and problem-solving skills, ability to identify operational deficiencies, and excellent relationship management skills. Recognized as a trusted advisor to executive management teams.

Areas of Expertise

- Applied Data Science
- Natural Language Technologies
- Speech Recognition, Voice Self-Service
- Enterprise Software Development
- Contact Center Technologies
- Agile DevOps, Continuos Product Delivery

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE



Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA