Data Science Portfolio

Winston Vu

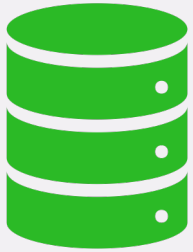
Introduction

- The M.S in Applied Data Science provides students with mathematical knowledge and programming experience with the use of SQL, R, and Python.
- The 7 core concepts of the program:
  - Developing an overview of major practices in data science
  - Collect and organize data
  - Identify patterns visually
  - Statistically, mining, and developing alternative strategies
  - Developing a plan of action
  - Demonstrating communication
  - Learn ethical decisions of data science practice

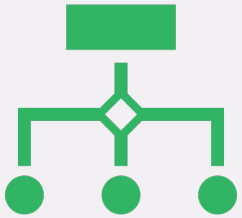
Classes

- There are 4 classes involved that demonstrate the 7 core concepts:
  - IST 659 – Data Administration Concepts and Database Management
  - IST 652 – Scripting for Data Analysis
  - IST 736 - Text Mining
  - IST 718 – Big Data Analytics

IST 659 – Database Administration and Database Management



The class taught the basics of creating a database in SQL, performing an analysis in SQL, and managing a database in SQL.

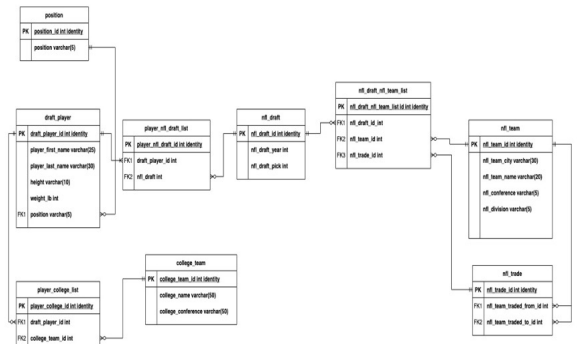


A conceptual model and logical model was taught in order to learn how to organize the databases and link them together.

The project for the class revolved around the top 10 draft picks in the NFL from the years 2010 to 2020.

IST 659 – Project Details

Logical Model



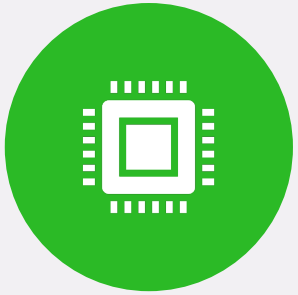
```
CREATE TABLE Draft_Player(
    draft_player_id int identity,
    player_first_name varchar(25) not NULL,
    player_last_name varchar(30) not NULL,
    height varchar(10) not NULL,
    weight_lb int not NULL,
    football_position_id int NOT NULL

    CONSTRAINT PK_Draft_player PRIMARY KEY (draft_player_id),
    CONSTRAINT FK1_Draft_player FOREIGN KEY (football_position_id) REFERENCES football_Position (football_position_id)
)
```

```
116  -- Insert values into draft_player
117  INSERT INTO Draft_Player (player_first_name, player_last_name,
118      height, weight_lb, football_position_id)
119      VALUES ('Cam', 'Newton', '6 ft 5 in', 248, 1);
120  INSERT INTO Draft_Player (player_first_name, player_last_name,
121      height, weight_lb, football_position_id)
122      VALUES ('Von', 'Miller', '6 ft 3 in', 246, 9);
123  INSERT INTO Draft_Player (player_first_name, player_last_name,
124      height, weight_lb, football_position_id)
125      VALUES ('Marcell', 'Darius', '6 ft 3 in', 319, 8);
126  INSERT INTO Draft_Player (player_first_name, player_last_name,
127      height, weight_lb, football_position_id)
128      VALUES ('A.J.', 'Green', '6 ft 4 in', 211, 3);
129  INSERT INTO Draft_Player (player_first_name, player_last_name,
130      height, weight_lb, football_position_id)
131      VALUES ('Patrick', 'Peterson', '6 ft 0 in', 219, 11);
132  INSERT INTO Draft_Player (player_first_name, player_last_name,
133      height, weight_lb, football_position_id)
134      VALUES ('Julio', 'Jones', '6 ft 3 in', 220, 3);
```

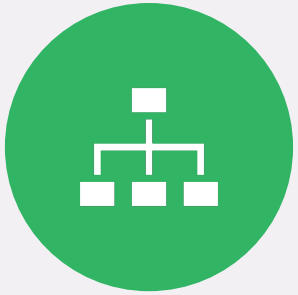
	college_conference	# of Draft Pick
2	SEC	38
3	PAC 12	16
4	ACC	14
5	Big 10	12
6	Big 12	9
7	Independent	4
8	MAC	3
9	AAC	2
10	Missouri Conference	1
11	Mountain West	1

- The project involved creating a problem and relating it to business functions.
  - This project revolved around the top 10 picks in the NFL draft from 2010-2020.
- A logical model was created to organize the database and find connections with the data.
- Questions were created in order to find analytical insight on the data.

IST 659 - Reflection



CLASS TAUGHT ABOUT BUSINESS AND REAL-LIFE SCENARIOS IN WHICH DATA CAN BE USED BY SQL.



THERE WAS ORGANIZATION TO PREPARE HOW THE DATA SHOULD BE PLACED.

DATA WAS MANUALLY COLLECTED, VISUALIZATIONS WERE USED TO ORGANIZE THE DATA, AND PATTERNS WERE FOUND IN DATA.

IST 652 - Scripting for Data Analysis

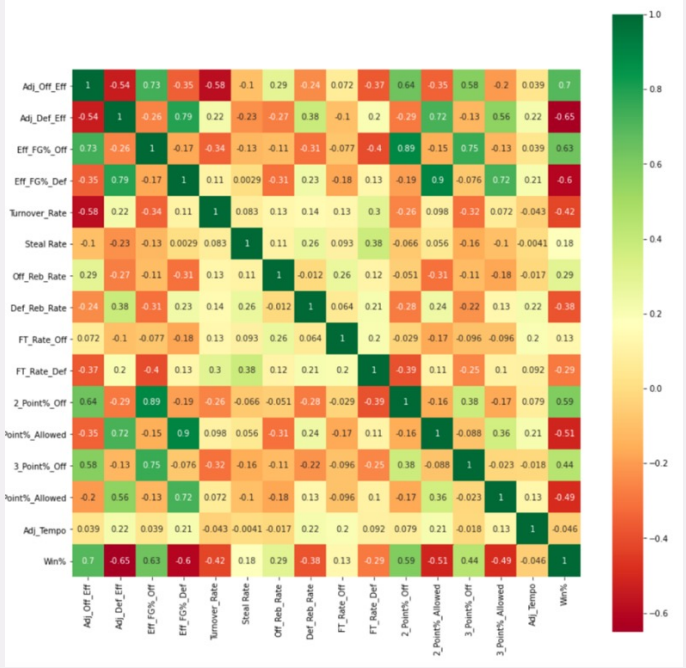
THE CLASS TAUGHT STUDENTS ABOUT HOW TO WRITE SCRIPTS IN PYTHON

SCRIPTS ARE IMPORTANT BECAUSE IT ALLOWS USERS TO CREATE A FUNCTION, SUCH AS DEFINITIONS AND FOR LOOPS, THAT COULD HELP FOR LATER USE IN ANALYSIS.

THE PROJECT FOR THIS CLASS INVOLVED A NCAA BASKETBALL DATASET, WHICH WAS USED TO PROJECT WHAT FACTORS LED TO WINS.

IST 652 – Project Details

```
Win Percentage Correlation:
 Win%                 1.000000
Adj_Off_Eff           0.695083
Eff_FG%_Off           0.634095
2_Point%_Off          0.593644
3_Point%_Off          0.442494
Off_Reb_Rate          0.291860
Steal Rate            0.175753
FT_Rate_Off           0.129505
Adj_Tempo            -0.045864
FT_Rate_Def          -0.285449
Def_Reb_Rate         -0.382775
Turnover_Rate        -0.420977
3_Point%_Allowed     -0.492452
2_Point%_Allowed     -0.511597
Eff_FG%_Def          -0.602441
Adj_Def_Eff          -0.651037
Name: Win%, dtype: float64
```



- The project involves using NCAA basketball data in order to make business related decisions while also accounting for the evolution of basketball.
- Python functions used in the project are Variance Inflation Factor (VIF), linear regression, correlation, and other simple Python functions.
- Visualizations were used to interpret data, such as correlation matrices or simple ordered correlation.

IST 652 - Reflection

THE CLASS WAS THE FIRST INTRODUCTION INTO PYTHON.

CLASS TAUGHT A LOT ABOUT HOW TO USE THE SIMPLE FUNCTIONS IN PYTHON, BUT ALSO, IT ENCOURAGED STUDENTS TO LOOK FURTHER AND USE ITEMS THAT WERE NOT LEARNED.

THOUGH THE CLASS WAS A BASIC INTRODUCTION, IT PROVIDED THE FRAMEWORK FOR FUTURE CLASSES.

IST 736 – Text Mining

The class taught about using machine learning techniques to predict text analysis.

Text Mining is often used in voice recognition, chat bots, and classification.

The project consists of providing lyrics of different genres and predicting what genre the lyrics belonged to.

IST 736 – Project Details

	Lyrics	Song	Artist	Genre
0	i heard life is what pass when your too busi l…	Late July	Zach Bryan	Country
1	if two fillup is all it cost i guess ill make …	Crooked Teeth	Zach Bryan	Country
2	he wa a boy who wa a dreamer and he flew so hi…	Heading South	Zach Bryan	Country
3	im condemn im condemn oh my heart is on the me…	Condemned	Zach Bryan	Country
4	in a life where we work out there a hous upon …	A Life Where We Work Out	Flatland Calvary	Country
..	…	…	…	…
235	oh yeah oh yeah everyth gonna be alright thi m…	Mannish Boy	Muddy Water	Blues
236	lie to me and tell me everyth is alright lie t…	Lie to Me	Jonny Lang	Blues
237	you sing a song while sit at a red light you t…	Red Light	Jonny Lang	Blues
238	come on oh babi dont you wanna go come on oh b…	Sweet Home Chicago	Eric Clapton	Blues
239	uh aw sooki sooki now hey ow uh come on babi h…	Groove Me	King Floyd	Blues

[240 rows x 4 columns]

Unigram Random Forest

	precision	recall	f1-score	support
Blues	0.50	0.75	0.60	12
Country	0.64	0.58	0.61	12
Pop	0.50	0.50	0.50	12
R&B	0.50	0.42	0.45	12
Rap	0.92	1.00	0.96	12
Rock	0.62	0.42	0.50	12
accuracy			0.61	72
macro avg	0.61	0.61	0.60	72
weighted avg	0.61	0.61	0.60	72

Bigram Random Forest

	precision	recall	f1-score	support
Blues	0.62	0.83	0.71	12
Country	0.50	0.42	0.45	12
Pop	0.62	0.42	0.50	12
R&B	0.33	0.33	0.33	12
Rap	1.00	1.00	1.00	12
Rock	0.21	0.25	0.23	12
accuracy			0.54	72
macro avg	0.55	0.54	0.54	72
weighted avg	0.55	0.54	0.54	72

- Project was to create a classifier to match lyrics to its genre
- Vectorization, tokenization, and train/test split were used to pre-process data.
- Many classifiers were used, but the Random Forest provided the best metrics and the highest accuracy at 52.26%
- Accuracies and metrics were not particularly high, but it is more realistic.

IST 736 - Reflection

Class taught about how process text data, use machine learning on text data, and apply it to real life business decisions.

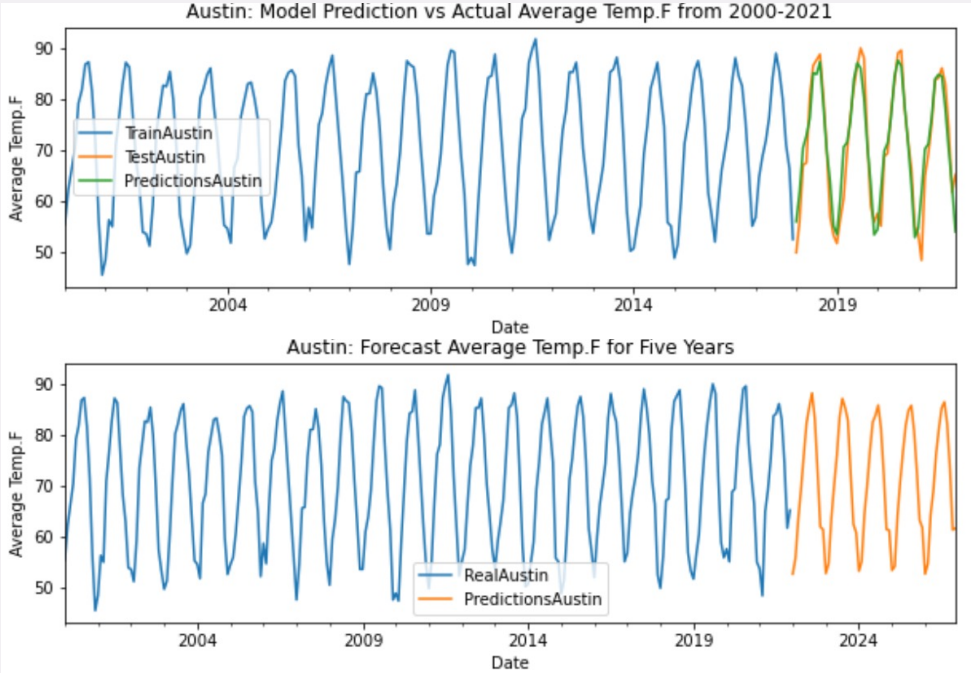
Accuracies were not high, which means there could be much more to be done to improve the metrics.

Applied core concepts from the previous class, Natural Language Processing, in order to complete project.

IST 718 – Big Data Analytics

- The class taught about many ways to approach data, such as approaching a time-series problem or a classification problem.
- Different sets of code and machine learning tactics were learned to approach these types of problems.
- The project was focused on climate change in Texas and how insurance premiums are impacted from climate change.

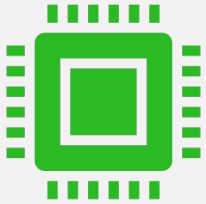
IST 718– Project Details



Fiscal Year	Premium Price
2022-2023	$1,406.10
2023-2024	$1,446.09
2024-2025	$1,487.14
2025-2026	$1,516.30
2026-2027	$1,555.24

- The project focused on a time-series based problem, focusing on predicting future temperatures, with the use of climate variables, as well as insurance premiums.
- Results showed increases in premiums every year due to rising temperatures
- The scikit-learn forecast package was used to train, test, and predict the future temperatures
  - Used because data was multivariate
- The Prophet package was used to predict the insurance premiums price
  - Used because data was univariate

IST 718 - Reflection



THERE WAS MANY CORE CONCEPTS INVOLVED WITH THE PROJECT, SUCH AS COLLECTION OF DATA, COMMUNICATION, AND ALTERNATIVE STRATEGIES

SCIKIT-LEARN FORECAST WAS A RECENTLY RELEASED PACKAGE IN 2022, AND IT WAS TESTED FOR THIS PROJECT.

COLLABORATED WITH OTHER MEMBERS TO ENSURE TIMELINES WERE MET, AND PROJECT WAS ORGANIZED AND COMPLETED.

Reflecting on Core Concepts

- Developing an overview of major practices in data science
  - Each class provided details on the applications to data science and how to use techniques in SQL, Python, and R to accomplish tasks
- Collect and organize data
  - The data could have been obtained through Kaggle, but most data was manually collected to reflect real life business scenarios
  - Pre-processing is an essential part due to its effect on the accuracies and metrics on a model or to let the model run.
- Identify patterns visually, statistically, and data mining
  - Graphs and visuals were used to understand the data and communicate the data to an audience
  - Basic graphs, heatmaps, basic statistics, or simple tables were used to see if there are any patterns or insights within the data
- Developing alternative strategies
  - Learning new functions and algorithms not presented during the class helped grow the knowledge for the data science world.
- Developing a plan of action
  - A plan of action was created in order to prepare for a project and how to approach certain scenarios
  - It is essential to plan ahead before running wild with the problem a person is trying to solve
- Demonstrating communication
  - Communication can involve with people or results
  - Communication could involve working with a team to complete tasks for a project or communicating results within a paper or a presentation
- Learn ethical decisions of data science practice
  - Classes taught about how to be ethical with data by not being bias towards certain results and implementing data integrity and data security to protect the data from outside invaders.

Final Thoughts

The program has taught a lot about not only the coding aspect of data, but also the ethical concerns and business implications as well

Each project allowed for a team to work together to solve the problem or work solo in order to prepare for the real world

Enjoyed learning all aspects of data, the approaches, and will apply all my learnings in my job as a Jr. Data Scientist.