Course Project

Foundations of Data Science

Prof. Rumi Chunara

Project for Office of the Graduate Admissions NYU

By
Kunal Relia

Topics to be Covered

- Problem
- Data being used
- Assumptions
- Model being built
- Tasks Completed & Self-Evaluation

The Problem

- A real world data science project where customer requirements will be satisfied

- Task is to analyze the Applications to find various results like
  - Average GRE scores
  - Number of applications for each semester
  - Average UnderGrad GPA
  - Average TOEFL Score
  - Countries with highest applications

The Problem

- A real world data science project where customer requirements will be satisfied

- Task is to analyze the Applications to find various results like
  - Average GRE scores
  - Number of applications for each semester
  - Average UnderGrad GPA
  - Average TOEFL Score
  - Countries with highest applications

BUT WAIT!!!

WHAT HAS ALL THIS TO DO WITH DATA SCIENCE??

The "REAL" Problem is THE DATA

The "REAL" Problem

- As we all know, NYU Tandon School has undergone various affiliations over the last decade
- This means that with each affiliation, the administrative tasks differed which in turn meant different software being used each time
- This has led to highly non-coherent data of the applications stored in spread sheets

Solution to the "REAL" Problem

- Major task currently being carried out is data cleaning
- Office of Graduate Admissions is helping me understand the standard of the required data
- The light on the background is thrown upon by the customer during the understanding of their requirements

Topics to be Covered

- ~~Problem~~
- Data being used
- Assumptions
- Model being built
- Tasks Completed & Self-Evaluation

THE DATA

Data Table

Attribute Name	Attribute Type	Description
Decision	String	Decision to the application
Acad_plan	String	Department applied to
ACAD_PROG	String	Master or PHD
ADMIT_TERM	String	Term to which applied for
GRE_Verbal	Numeric	GRE verbal score
GRE_Quant	Numeric	GRE quantitative score
TOEFL_Total	Numeric	Total TOEFL Score
Transcript1 GPA	Numeric	UnderGrad GPA
Transcript1 School Name	String	UnderGrad School Name
Stage of Process	String	Current stage of the application
Submitted Date	Date	Date on which application is submitted
Decision Publish Date	Date	Date on which decision on application was posted
Scholarship_Amount	Numeric	Scholarship awarded
First_Condition	String	Conditional admit or not
COUNTRY OF CITIZENSHIP	String	Country of citizenship of the applicant
Deposit Paid	String	Whether admission offer was accepted or not

Problems with the data set given

- On initial scrutiny the data is found to be incomplete and inconsistent.
- It is incomplete in the form that GRE score for some of the applications is missing while it is inconsistent because there is a mix and match of new & old GRE score for newer admit terms.
- The data set is inconsistent in the form that same attributes have different values. For example the attribute "Decision" was initially "Pre_Deposit_Admin_Letter"(during the use of applyyourself) which then changed to "Pre_Deposit_F10" to "FT Admit" (current software).

Original Data Sets

1-GRE Verbal	GRE Quantitati	1-GRE Total
151	159	310
350	730	1080
148	165	313
151	162	313
152	160	312
156	161	317
151	164	315
144		301
145	161	306
149	163	312
155	161	316
152	162	314
155	153	308
147	170	317
133	161	294
148	167	315
155	168	323
155	166	321
156	163	319
160	164	324

Old GRE score

Missing GRE score

Fall 2014 Data Set

Problems with the data set given

- On initial scrutiny the data is found to be incomplete and inconsistent.
- It is incomplete in the form that GRE score for some of the applications is missing while it is inconsistent because there is a mix and match of new & old GRE score for newer admit terms.
- The data set is inconsistent in the form that same attributes have different values. For example the attribute “Decision” was initially “Pre_Deposit_Admin_Letter”(during the use of applyyourself) which then changed to “Pre_Deposit_F10” to “FT Admit” (current software).

Original Data Sets

l 2009	2792369	393426		Pre_Deposit_Admin_Letter	Admit	[redacted]	Fall 2009
2010	3539473	408753		Pre_Deposit_F10	Admit	[redacted]	Fall 2010
2014	6311777	550890	N1	Fall 2014 - FT admit	Admit	[redacted]	Fall 2014

Introduction of "N" number

Topics to be Covered

- ~~Problem~~
- ~~Data being used~~
- Assumptions
- Model being built
- Tasks Completed & Self-Evaluation

Assumptions

- We have seen the data set. Hence, it is clear that no assumption can be made without taking Office of the Graduate Admissions into the picture
- Making wrong assumptions can prove to be disastrous
- In some cases assumptions cannot be made without contacting the customer

Assumptions

- Cases where assumptions can not be made

Assumptions

- Cases where assumptions can not be made
  - Various applications are missing the Admission Decision. Hence a talk with the office helped me understand that those are the incomplete applications that were submitted but missed some key requirement for admission consideration

Assumptions

- Cases where assumptions can not be made
  - Various applications are missing the Admission Decision. Hence a talk with the office helped me understand that those are the incomplete applications that were submitted but missed some key requirement for admission consideration

  - Another key assumption that I was told to make was that if an admitted applicant deposited money, I can assume that they accepted the admission offer and eventually enrolled

Topics to be Covered

- ~~Problem~~
- ~~Data being used~~
- ~~Assumptions~~
- Model being built
- Tasks Completed & Self-Evaluation

Model to be used

- Entire project is being done following the Jeff Hammerbacher's model – Hence the first tasks carried out are data collection and data cleaning
- We have already discussed the non-coherent data source and its incompleteness and inconsistency
- We now have a dataset that is ready for analysis

Model to be used

- The main outcome of the various tasks will be "Decision" – we wish to know what decision is made based on a set of predictors
- Consider a scenario where "*GRE_Quant*" and "*ACAD_PROG*" are the predictors and "*Decision*" is the outcome
- Here, we will categorize the GRE score into various ranges and we will try to determine the outcome for each range

Model to be used

- We will use **Linear Regression** to find out various other pairs of predictors and outcome

- For example,
  - Predictors – GRE_Quant, Transcript1 GPA, ACAD_PROG
  - Outcome – Scholarship_Amount

- Some statistical inferences will also be made
  - Average Life of an application
  - Percentage of applications who accept the admission decision

Model to be used

- Once we have a set of predictors & outcome, we will have a corresponding data set that is large enough to carry out a supervised learning

- KNN algorithm will help find some interesting patterns and determine an outcome for a set of predictors
  - Say if the predictors are ACAD_PROG & GRE_Quant, we will try to predict (based on historical data) whether an applicant will get a decision or not

Topics to be Covered

- ~~Problem~~
- ~~Data being used~~
- ~~Assumptions~~
- ~~Model being built~~
- Tasks Completed & Evaluation

Task Completed

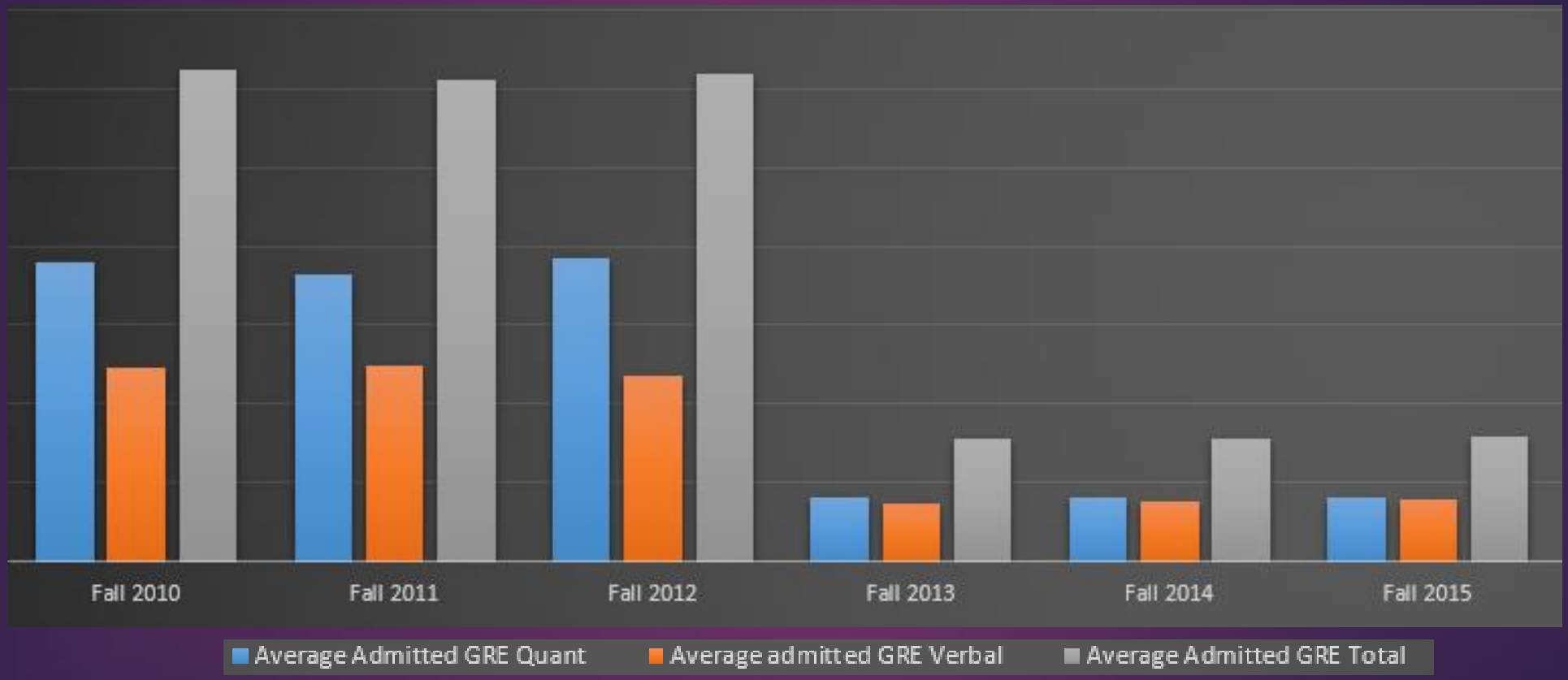
- Data Collection & Understanding the problem from customer
- Data cleaning (all Fall applications completed)
- Some statistical analysis for the customer
  - GRE score for the PhD applications of CS department over the years

Ready for the Numbers?

We are crunching
a few numbers for you..
Please wait!!

Sorry ☹
Can not share the
statistics with you all but..

Here is a sneak peek



Here is a sneak peek

Evaluation

- Evaluation for accuracy now becomes one of the most important thing
- We can use F-measure to evaluate the given data set
- Outliers can be removed by using a simple graph visualizations – like the old GRE score amongst the newer ones

Thank You!

- I thank Prof Rumi Chunara for giving all of us an opportunity to work on such projects and for her comments on the project proposal
- I also thank Qi for his inputs on the project proposal

Thank You!

- I also thank Office of Graduate Admissions for entrusting me with this confidential information
- I thank Dr. Raymond Lutzky and his entire team for their support
- I thank Fatema Dalal for allowing me to work on this project

Thank You All!

All questions are most welcome