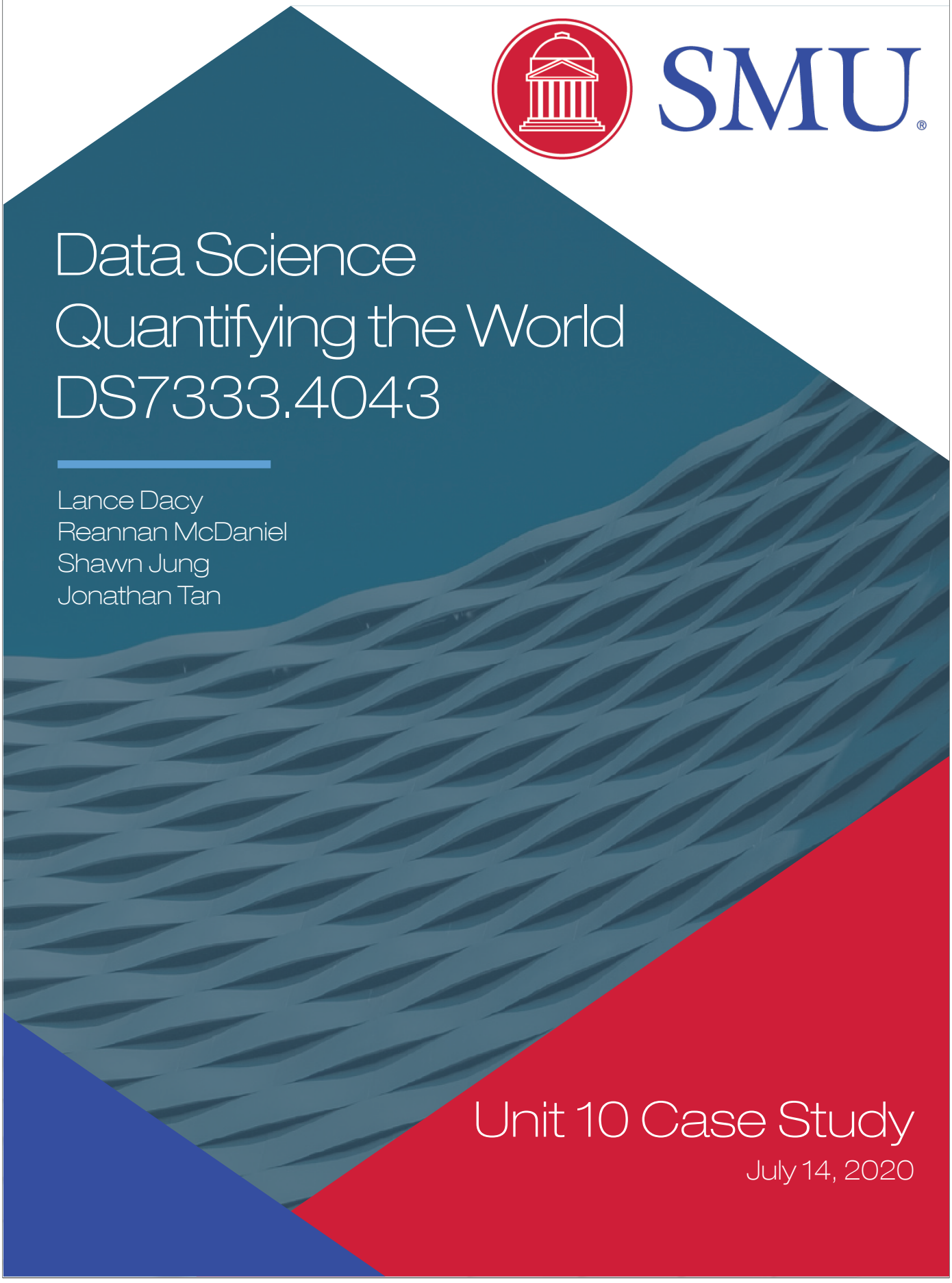


Data Science Quantifying the World DS7333.4043

Lance Dacy
Reannan McDaniel
Shawn Jung
Jonathan Tan

Unit 10 Case Study

July 14, 2020

Table of Contents

Introduction

Technology improvements for storing and retrieving data has progressed rapidly in the last 5 years. In fact, computing power to process complex operations on the data is becoming increasingly more available for the layman to contribute to studies and analysis of data. Still, the data must be cleaned in a manner that allows for proper interpretation of statistical analysis or prediction models. Ever increasing size of data contributes to the issue of data that might go missing or is in a format that is unintelligible for the data scientist.

Myriad methods exist to conduct imputation exercises on the data. Imputation is the process of replacing missing data with substituted values. When substituting for a data point, it is known as "unit imputation"; when substituting for a component of a data point, it is known as "item imputation". It is best for a data scientist to consider various approaches to determine what would be best for the end model. Simply removing data because it doesn't exist isn't always the best measure. Imputing data incorrectly can also skew the results. Careful attention must be taken to ensure data is measured against the approaches to determine which one best fits the need.

There have been many theories embraced by scientists to account for missing data but the majority of them introduce large amounts of bias. A few of the well known attempts to deal with missing data include:

- hot deck and cold deck imputation
- listwise and pairwise deletion
- mean imputation
- regression imputation
- last observation carried forward
- stochastic imputation
- multiple imputation.

Once all missing values have been imputed, the data set can then be analyzed using standard techniques for complete data. Such an exercise will be conducted on a data set that is meant to determine home values in the state of California. Given that California is a large homestead state as well as a multitude of climate and housing development opportunities, this data set will give a real-time view into imputation techniques and the insight into the performance of the solutions that a data scientist might explore.

Method

The dataset chosen for this exercise includes 8 attributes and over 20 thousand records. For additional analysis the target variable is median house value. For the purpose of this analysis the data is converted into a data frame in Python Pandas.

An action view is then generated to view the first 5 records within the data for each variable as well as each variable's descriptive statistics, like mean, standard deviation and percentile of where the data falls.

Figure 1: Action View of First 5 Records

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25

Figure 2: Descriptive Statistics

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude
count	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000
mean	3.870671	28.639486	5.429000	1.096675	1425.476744	3.070655	35.631861	-119.569704
std	1.899822	12.585558	2.474173	0.473911	1132.462122	10.386050	2.135952	2.003532
min	0.499900	1.000000	0.846154	0.333333	3.000000	0.692308	32.540000	-124.350000
25%	2.563400	18.000000	4.440716	1.006079	787.000000	2.429741	33.930000	-121.800000
50%	3.534800	29.000000	5.229129	1.048780	1166.000000	2.818116	34.260000	-118.490000
75%	4.743250	37.000000	6.052381	1.099526	1725.000000	3.282261	37.710000	-118.010000
max	15.000100	52.000000	141.909091	34.066667	35682.000000	1243.333333	41.950000	-114.310000

Using a linear regression model, training and test datasets are created using a 70/30 split. The linear regression model accuracy for predicting median house value is 60.94%. As a general rule, a good accuracy metric is 90% or greater, however, this exercise's purpose is to measure the effects of 'missing value handling', against the baseline model. The focus will be on the delta or diff of the performance through step 1 to 4 of the exercise in details that follow.

Data

Data Source

Over the years the housing market has seen drastic changes in price as well as purchasing power for buyers. California is a desirable market due it the vast climate and scenery options that include mountain living, thousands of miles of coastal property, forest, and deserts. These vast market selections come at a price however; making California housing one of the most lucrative and expensive in the United States.

The sklearn.dataset.fetch_california_housing dataset was derived from the 1990 U.S. census, using one row per census block group. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people). The data was most famously used in Pace, R. Kelley and Ronald Barry, Sparse Spatial Autoregressions, Statistics and Probability Letters, 33 (1997) 291-297/.

Exploratory Data Analysis

Missing data is a common problem in real-world settings and for this reason has attracted significant attention in statistical literature. In this exercise, a flexible framework based on formal optimization to impute missing data with mixed continuous and categorical variables will be explored. The target variable is the median house value for California districts.

Data Set Characteristics

- Number of Instances 20,640
- Number of Attributes 8 numeric, predictive attributes and the target

Attribute Information

- MedInc median income in block
- HouseAge median house age in block
- AveRooms average number of rooms
- AveBedrms average number of bedrooms
- Population block population
- AveOccup average house occupancy
- Latitude house block latitude
- Longitude house block longitude

The abs(max) coef-value is 0.7791446958109843 and the variable associated with this coef-value is AveBedrms.

Question 1

What is the loss and what are the goodness of fit parameters? This will be the baseline for comparison. The measures are

- MAE (mean absolute error): measure of errors of observed vs predictions
- MSE (mean squared error (MSE): measures the average of the squared errors
- RMSE (root mean square error): measures the difference of the population
- R2 (sample measure): measures the proportional variance of the regressor from the defendant variables

The baseline measures for the first step of this exercise are:

- MAE: 0.530
- MSE: 0.537
- RMSE: 0.733
- R2: 0.597

Question 2

In each case [1%, 5%, 10%, 20%, 33%, 50%] perform a fit with the imputed data and compare the loss and goodness of fit to the baseline.

Using random sampling on the full dataset a function is generated for comparison purposes. New measurements are added and are calculated for imputation. The new measurements are "mae_diff", "mse_diff", "rmse_diff" and "R2_diff". Each of these a calculation is the difference between the model score and the baseline score.

Figure 3: Imputed Data Fit with 1%, 5%, 10%, 20%, 33%, and 50%

	data	imputation	mae	mse	rmse	R2	mae_diff	mse_diff	rmse_diff	R2_diff
0	original	none	0.529571	0.536969	0.732781	0.597049	NaN	NaN	NaN	NaN
1	Nullify 1.0%	median	0.530093	0.538481	0.733813	0.595914	0.000522	0.001512	0.001031	-0.001135
2	Nullify 5.0%	median	0.531264	0.540778	0.735376	0.594191	0.001693	0.003809	0.002594	-0.002858
3	Nullify 10.0%	median	0.534101	0.556453	0.745958	0.582428	0.004530	0.019485	0.013176	-0.014622
4	Nullify 20.0%	median	0.540786	0.649251	0.805761	0.512791	0.011215	0.112282	0.072979	-0.084258
5	Nullify 33.0%	median	0.547262	0.673146	0.820455	0.494860	0.017691	0.136177	0.087673	-0.102190
6	Nullify 50.0%	median	0.537447	0.562518	0.750012	0.577877	0.007876	0.025549	0.017230	-0.019172

Question 3

In each case [10%, 20%, 30%] perform a fit with the imputed data and compare the loss and goodness of fit to the baseline.

Figure 4: Imputed Data Fit with 10%, 20%, and 30%

	data	imputation	mae	mse	rmse	R2	mae_diff	mse_diff	rmse_diff	R2_diff
0	original	none	0.529571	0.536969	0.732781	0.597049	NaN	NaN	NaN	NaN
1	Nullify 1.0%	median	0.530093	0.538481	0.733813	0.595914	0.000522	0.001512	0.001031	-0.001135
2	Nullify 5.0%	median	0.531264	0.540778	0.735376	0.594191	0.001693	0.003809	0.002594	-0.002858
3	Nullify 10.0%	median	0.534101	0.556453	0.745958	0.582428	0.004530	0.019485	0.013176	-0.014622
4	Nullify 20.0%	median	0.540786	0.649251	0.805761	0.512791	0.011215	0.112282	0.072979	-0.084258
5	Nullify 33.0%	median	0.547262	0.673146	0.820455	0.494860	0.017691	0.136177	0.087673	-0.102190
6	Nullify 50.0%	median	0.537447	0.562518	0.750012	0.577877	0.007876	0.025549	0.017230	-0.019172
0	Random Missing 2Col 10%	median	0.534778	0.528455	0.726949	0.603438	0.005207	-0.008513	-0.005832	0.006389
0	Random Missing 2Col 20%	median	0.536835	0.532347	0.729621	0.600518	0.007264	-0.004622	-0.003161	0.003468
0	Random Missing 2Col 30%	median	0.540682	0.540558	0.735226	0.594356	0.011111	0.003589	0.002445	-0.002693

Question 4

Perform a fit with the imputed data [25%] and compare the loss and goodness of fit to the baseline.

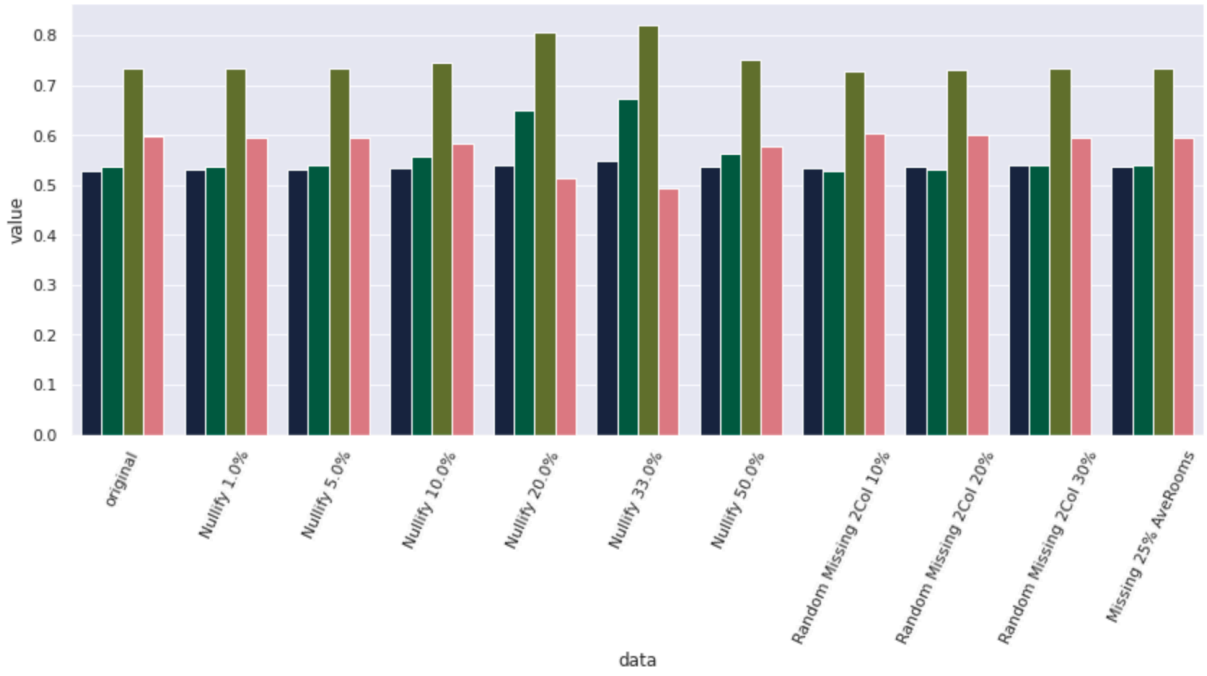
Figure 5: Imputed Data Fit with 25%

	data	imputation	mae	mse	rmse	R2	mae_diff	mse_diff	rmse_diff	R2_diff	model
0	original	none	0.529571	0.536969	0.732781	0.597049	NaN	NaN	NaN	NaN	0
1	Nullify 1.0%	median	0.530093	0.538481	0.733813	0.595914	0.000522	0.001512	0.001031	-0.001135	1
2	Nullify 5.0%	median	0.531264	0.540778	0.735376	0.594191	0.001693	0.003809	0.002594	-0.002858	2
3	Nullify 10.0%	median	0.534101	0.556453	0.745958	0.582428	0.004530	0.019485	0.013176	-0.014622	3
4	Nullify 20.0%	median	0.540786	0.649251	0.805761	0.512791	0.011215	0.112282	0.072979	-0.084258	4
5	Nullify 33.0%	median	0.547262	0.673146	0.820455	0.494860	0.017691	0.136177	0.087673	-0.102190	5
6	Nullify 50.0%	median	0.537447	0.562518	0.750012	0.577877	0.007876	0.025549	0.017230	-0.019172	6
0	Random Missing 2Col 10%	median	0.534778	0.528455	0.726949	0.603438	0.005207	-0.008513	-0.005832	0.006389	7
0	Random Missing 2Col 20%	median	0.536835	0.532347	0.729621	0.600518	0.007264	-0.004622	-0.003161	0.003468	8
0	Random Missing 2Col 30%	median	0.540682	0.540558	0.735226	0.594356	0.011111	0.003589	0.002445	-0.002693	9
0	Missing 25% AveRooms	median	0.536806	0.540218	0.734995	0.594611	0.007235	0.003250	0.002214	-0.002439	10

Conclusion

Describe your imputation approach and summarize your findings. What impact did the missing data have on your baseline model's performance?

Figure 6: Summary of Comparisons



Most of the models are close to the baseline interpretation of the data without any imputation (model 0). The only exceptions are models 4 and 5 which imputed a random 20% and 30% of the data from 'AveRooms'.

Both of these models had significantly higher MSE and RMSE, with proportionately lower R2. The MAE was not much different from the other models, indicating that the 20/30% missing data for these models contained some extreme outliers which are shown in the higher MSE and RMSE.

Because the next model with 50% of the data missing from 'AveRooms' had a lower MSE and RMSE, the random nature of the missing variables perhaps didn't include those same extreme outliers.

A great take-away from this exercise is the power of a simple method when working with missing data. Simply taking the median value arrives at similar results to the baseline encouraging the data scientist to avoid mechanically omitting missing-data records for the sake of saving time. Explore these methods and tools to find the best mechanism.

Appendix

As with any technology; the system required to generate meaningful data, cleaning of the data, as well as the models that were implemented require various software packages. The raw "code" used to explore the methods detailed above can be found in the following libraries:

- GitHub Repository
- Google Colab Worksheet