An investigation of business analytics and machine learning modles in the healthcare industry.

by

Mohamed Megahed

December 2020

A Capstone
Research Project
submitted to the College of Business
Administration
Kent State University in partial fulfillment of the
requirements for the
degree of Master of Science

in

Business Analytics

Acknowledgements

I am grateful to many people who have supported me during my study and my life at Kent State University, and whom I can never thank enough: First and foremost, to god who guided me through my Master journey, to my parents, **Mr. Arafa Awad** and **Mrs. Zienb Lotfi**, to whom I have dedicated this Research Project. Without them my accomplishment would not have been achieved, especially during this challenging year 2020.

Most sincerely thanks to my advisor and mentor, **Prof. Murali Shanker**, for his invaluable advice, help, and encouragement throughout this endeavor,.

Special thanks to all my friends who helped make my time at KSU memorable.

Abstract

Healthcare analytics is developing and is anticipated to change the nature of health care institutions quickly. Boosting fostering and implementation of digital health records in medical facilities, centers and other medical care setups will benefit in healthcare analytics market development. This project investigates the underlying technologies, features that are necessary predictors for how much an individual will be charged, and the methodologies used in developing the models, the major objective is to present ways of the data storytelling. Likewise, we will service a model that might give us an estimate as to what will be the charges. Nevertheless, we need to deeply investigate what factors affected the fee of a certain person. to do that, we should seek patterns in our data evaluation as well as gain comprehensive understanding of the relationships among variables and features and what the data is trying to tell us. finally, we will certainly go detailed to recognize the story behind the patients in the dataset just via this manner we can have a far better understanding of what data will certainly help our model to have a better accuracy of real patient charge.

TABLE OF CONTENTS

INTRODUCTION

Data analytics have become the primary driver for business operations and strategies. the future of digital business provides companies with almost unlimited possibilities to develop organization worth. The potential for data-driven organization methods and information products is higher than before to accelerate digital business operations.
This shift to data-driven business requires business as well as data leaders, such as (CDO) chief data officers to boost data and create new methods to solve business problems. It additionally greatly affects the work of data and analytics organizations and the venture competencies that must be constructed. *100 Data and Analytics Predictions Through 2024 (gartner.com)*

Healthcare providers, health insurance plans and healthcare clearinghouses gather patient medical data stemmed from their regular operations every day. If data mining methods are applied upon these data sets, these data can significantly profit the health care organizations. Nonetheless, individual identifiable person information needs to be shielded based on Medical insurance Mobility and Accountability Act (HIPAA), and the quality of client information additionally requires to be guaranteed for data mining jobs to achieve accurate results(Petajan, 2000). This project provides a patient data modeling which predicts patient charges with high accuracy and patient records that appropriates for data mining objectives.

Project Case Study (Allaint)

- The case study of this research is Allaint Treatment Center. Which is **Cleveland's opioid treatment headquarters** Located in Shaker Heights, Allaint has 4 branches across the state of OHIO. The drug recovery treatment centers focus on a wide variety of chemical dependency treatment options in an outpatient setting and Services are in an outpatient setting and include a variety of therapy methods such as:

  – Individual therapy
  – Group therapy
  – Recreational therapy

  With these considerable dedicated features, Alliant could be seen as an ideal model for research purpose.

Problem Description

Alliant has a serious problem regarding the billing and revenue system. Despite the company uses a comprehensive software "AXXESS" for billing, it lacks an obvious and organized system for estimating and calculating the patient charges also there is a lack of matching the claims with the remittance advices received from payers, which causes a misleading estimates of the patients charges and payments. Alliant uses the EFTs for payments which enables the company to record and collect data electronically for all claims and payments, unfortunately there is no internal unified database to collect and store all the data processed for the billing operations, Which causes unclear vision and misunderstanding of the RA amounts posted to Alliant.

Main Questions:

Is it possible to match claims with RAs and determine outstanding claims?

Can we build a ML model to predict the charges and payments for each patient?

Which predictive model is the best for Alliant?

What will be the financial position for the next 6 months?

Objective

The main objective of this project is to apply a business analytics and data science algorithms to the healthcare system at Alliant to provide business solutions and answer the questions pointed above, present methods of the data storytelling, deeply investigate what factors affected the fee of a certain person. to do that, we should seek patterns in our data evaluation as well as gain comprehensive understanding of the relationships among variables and features and what the data representing. Nevertheless, we need to deeply investigate what factors affected the fee of a certain person. to do that, we should seek patterns in our data evaluation as well as gain comprehensive understanding of the relationships among variables and features and what the data is representing, providing a new analytical method to understand, facilitate the billing system and provide accurate data of the financial position.

Methodology

The Project Approach is to collect the data needed for the 2020 year, through the Waystar, Axxess, and Payers portals (Insurance Companies). Applying exploratory data analysis and understanding the data types.

Excel tools such as (Power Pivot and Queries and connections) were used to match claims with the RAs received from Payers.

Using R tools to analyze, visualize the relations between variables, building ML Models to predict the charges for patients, and investigate different models to help to identify the high accuracy.

Data Collection:

Two ways used to collect data. First, patients records (Demographic data) for from the Operations and patient's relationship management, Second, claims data submitted to payers through the Waystar software since January 1, 2020. "Waystar software has been used by the company for more than 3 years" and the RAs. Received from the payers through the provider portals. The data from the portals received in the form of 835 coded files.

Data Analysis

Exploratory data analysis

Through the exploratory analysis, I was able to answer the following questions

1. Was the data read incorrectly?
2. Is the number of rows and columns accurate, and what you expected?
3. Are the columns correctly classified?
4. Are there specific types of claims?
5. Are there different patient types?

The data was not ready to be matched and merged well as the RAs. Received from Payers are in the form of 835 files we added a new software to the company information systems that helped us to convert the 835 files easily to an excel sheet. the patient's claim was pulled out from WayStar.

The data pulled from WayStar was clean and no missing data, but with the RAs. The data has 4770 observations and 23 columns, that's close to what we expect.

The columns were not classified appropriately, and there are many columns with no data so we had to eliminate the unnecessary columns like in the following table:

Table (1)

Rendering Provider ID	Check/EFT Date	Payer Message	Contract Code	Supplemental Info.	MOA Remark Codes	MIA Remark Codes	Crossover Carrier	Crossover Carrier ID	DRG Code	DRG Weight	App. Sender Code	Payee ID	Filing Ir
1962852533	12/10/2019										INSTAMED	1760978035	MC - M
1962852533	12/13/2019			$102.31 - Coverage Amount							INSTAMED	1164918413	MC - M
1962852533	12/13/2019			$102.31 - Coverage Amount							INSTAMED	1164918413	MC - M
1962852533	12/13/2019			$102.31 - Coverage Amount							INSTAMED	1164918413	MC - M
1962852533	12/13/2019			$102.31 - Coverage Amount							INSTAMED	1164918413	MC - M
1962852533	12/13/2019			$102.31 - Coverage Amount							INSTAMED	1164918413	MC - M
1962852533	12/13/2019			$102.31 - Coverage Amount							INSTAMED	1164918413	MC - M
1962852533	12/13/2019			$102.31 - Coverage Amount							INSTAMED	1164918413	MC - M
1962852533	12/13/2019			$102.31 - Coverage Amount							INSTAMED	1164918413	MC - M
1962852533	12/13/2019			$102.31 - Coverage Amount							INSTAMED	1164918413	MC - M

There were no different types of claims, but the claims are different in the charge amounts and insurance company. the patients are different in age, gender, smoker or not, number of children and therie address.

Data merging and matching

Using Excell power pivot and queries & Connections, we were able to match and merge the RAs. Files with the claims that have been processed and deleting the duplicated claims. After the matching process, we filtered the claims to paid and outstanding claims for each payer by weekly. The following is a figure showing 2 weeks in November from the matching sheet report:

Table (2): Matching Sheet Report

Alliant Treatment Center

Weeks of 2020

	11/2 - 11/6								11/9 - 11/13					
	Billed		Payments				Received	Outstanding	Billed		Payments			
			Old		Current						Old		Current	
Payer	# of Claims	Total Charges	# Claims	Amount	# Claims	Amount			# of Claims	Total Charges	# Claims	Amount	# Claims	Amount
Buckeye	87	$10,052.15	28	$2,456.08					36	$4,828.59	91	$10,090.78		
Caresource	96	$9,064.10	98	$10,010.12				Oct	33	$3,083.03	44	$3,708.69	2	$252.11
Medicaid	46	$4,636.93	18	$1,061.18			63	10	15	$1,331.05	43	$4,355.35		
Molina	67	$8,282.48	27	$3,295.17			68	4	30	$2,476.09	29	$3,255.90		
Paramount	57	$6,738.56	20	$1,942.74			52	6	56	$5,600.51	78	$7,398.30		
United Healthcare	71	$6,141.99	19	$2,743.30				2	64	$6,502.56	124	$12,264.88		
Total	424	$44,916.21	210	$21,508.59	0	$0.00	210	22	234	$23,821.83	409	$41,073.90	2	$252.11
Balance					(214)	($23,407.62)	(214)	22						$17,504.18

Statistical Analysis

Through the primary analysis by R we were able to obtain the following:

1. 43 Render Providers are responsible for the treatment services and meetings, for the last 11 months the number of claims generated for each Render Provider and the total charge amounts were significantly different among them see table 3.

Table (3): Total number of claims and chargers per Render Provider

Rendering.Provider <fctr>	Claims <int>	percent <dbl>	RP$TotalCharges <dbl>
ADAIR, DOUGLAS	206	4.31865828	25577.63
BRITZMAN, CYNTHIA	25	0.52410901	2393.67
CROSS, ANTON	304	6.37316562	54517.61
DAVIS-FLETCHER, EDITH	163	3.41719078	102445.59
DAVIS, CHAVELA	96	2.01257862	8879.07
DAVISFLETCHER, EDITH	55	1.15303983	3689.87
DAWSON, LYNELL	26	0.54507338	2195.69
GILSON-MOORE, NICO	122	2.55765199	14361.31
GILSONMOORE, NICO	12	0.25157233	1416.88
HARPER, TERRA	145	3.03983229	10779.84

1-10 of 43 rows Previous 1 2 3 4

- Francine Paknik has the highest number of claims 502 with a 10.5% of the total claims processed, while Edith Davisfletcher has the lowest number of claims 55 with a 1.1% of the total claims processed starting January 1, 2020 till December 15, 2020.
- EDITH DAVIS-FLETCHER has only 163 claims but has the highest total charges of $102,445.59 for the services he provided.

2. Alliants has been dealing with 11 payers since January 2020, each payer was charged with a certain number of claims, the biggest partner for Alliant was Caresource of Ohio, which has 1334 claims, while, Medical Mutual of Ohio has the lowest number of claims, only 7 claims since January 2020, see figure (1).

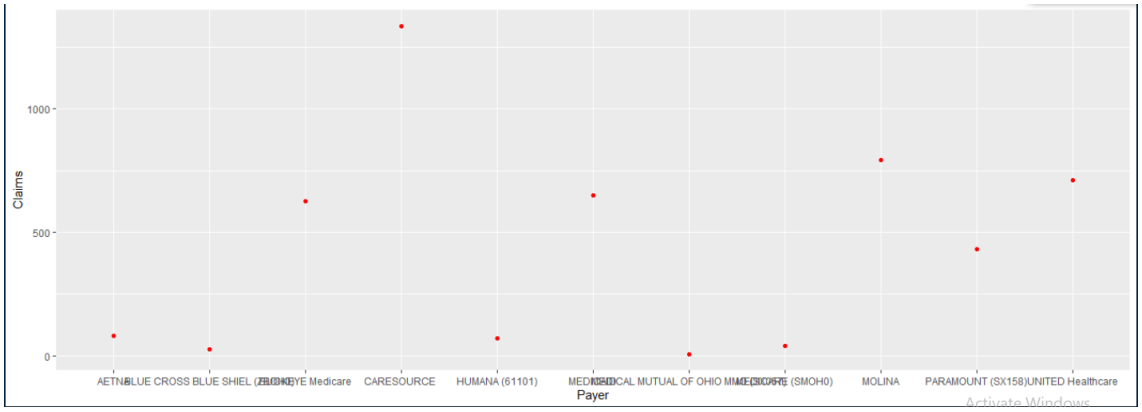


Figure (1): Total number of claims per payer

3. Out of the 4770 claims, 4671 claims were received by Payer and paid, while 99 are still counted as outstanding claims and rejected claims with billing errors, see table (4).

Table (4): Paid and Outstanding Claims

Status	Claims
Received by Payer	4671
Rejected - Name Matching Required	14
Rejected by Payer	49
Rejected by Waystar	5
Sent to Payer	31

4. The remits analysis shows that Alliant received the highest amount of payments from Medicaid of Ohio with $144,107.70 and 25.73% of the total payments, while CareSource, which has the highest total number of claims comes in the 2nd place with $ 130,475 and 23.29% of the total payments. On the other hand, Medical mutual has 0.13% of the total payments $740, which was expected as it has the lowest number of claims, see table (5).

Table (5): Total Payments per Payer

Payer.Name	Totalpayments	percent
MEDICAID	144107.70	25.7342067
CARESOURCE	130475.00	23.2997308
MOLINA	80285.11	14.3370106
BUCKEYE Medicare	69601.20	12.4291184
UNITED Healthcare	68633.21	12.2562584
PARAMOUNT (SX158)	47720.40	8.5217281
AETNA	6816.69	1.2172987
HUMANA (61101)	6222.07	1.1111137
MEDICARE (SMOH0)	3953.27	0.7059600
BLUE CROSS BLUE SHIEL (ZBOH0)	1430.36	0.2554283
MEDICAL MUTUAL OF OHIO MMO (SX057)	740.00	0.1321464

Correlation between Transaction date and Claims Status

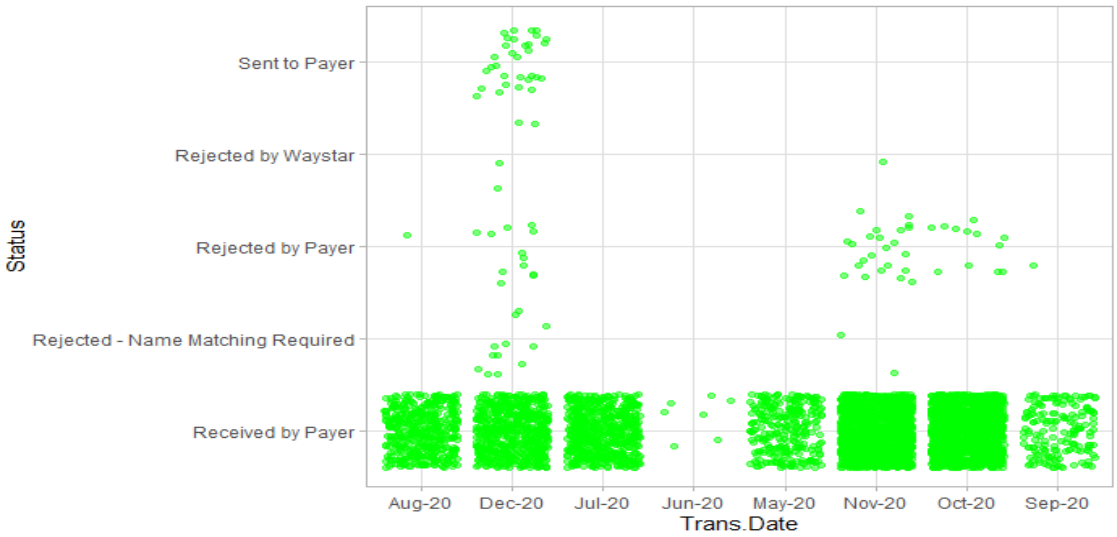


Figure (2)

Correlation between Payers and claims status

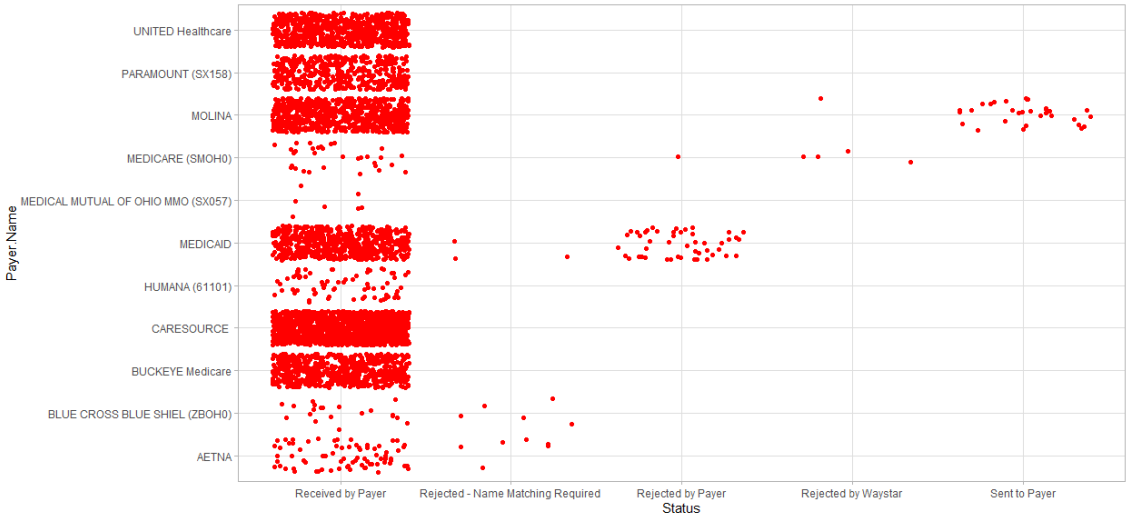


Figure (3)

Correlation between Charges, Sex and Region

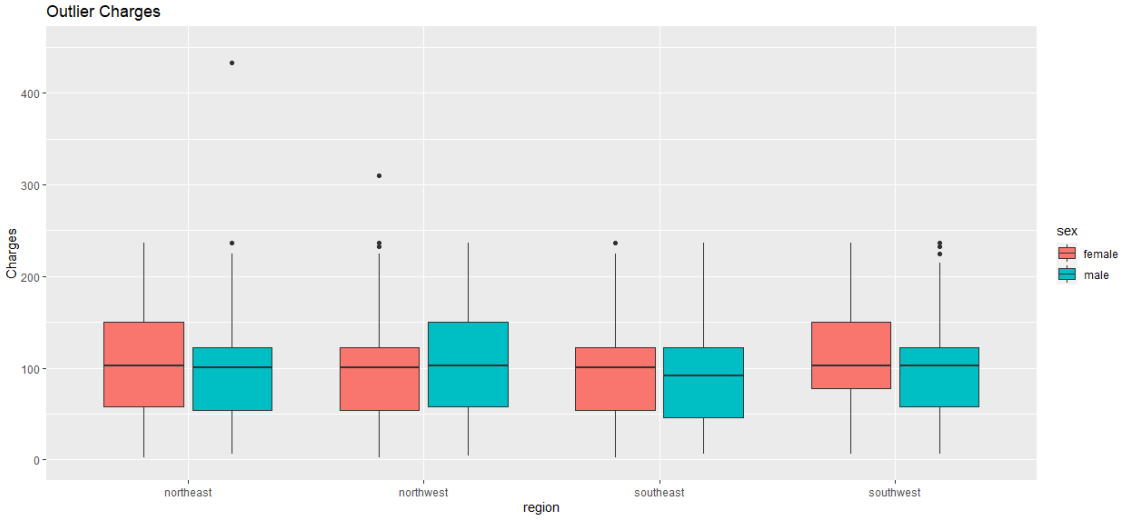


Figure (4)

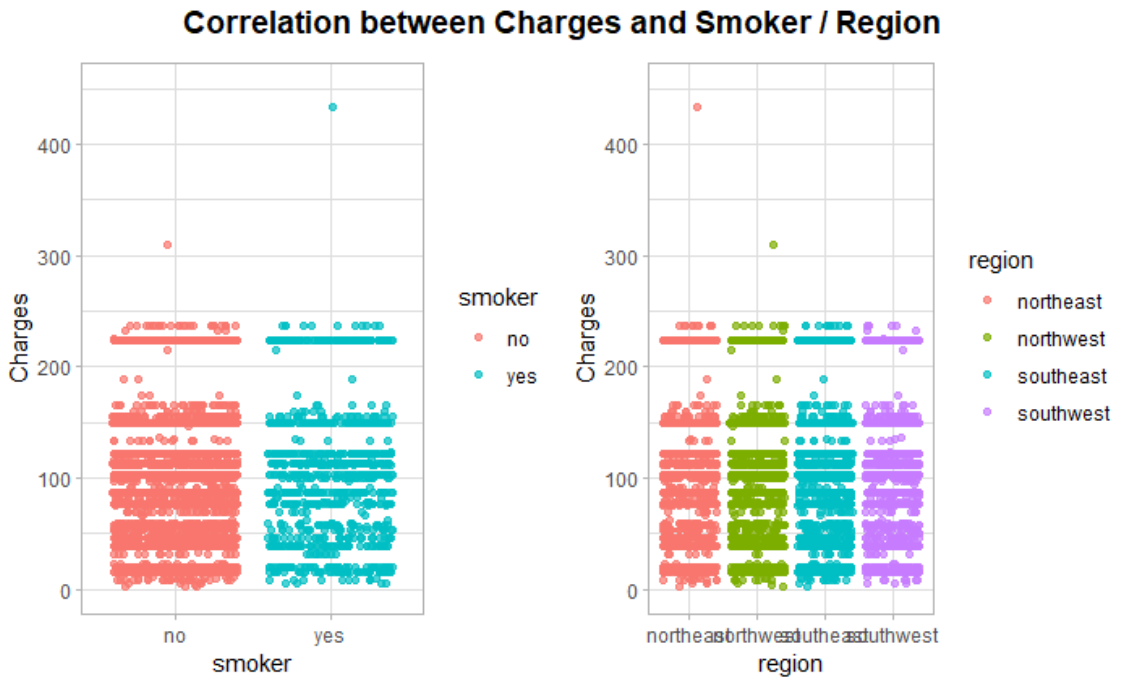


Figure (5)

Linear Regression Analysis

We used age, sex, bmi, children, smoker and region as the predictors and Charge as the target variable**.**

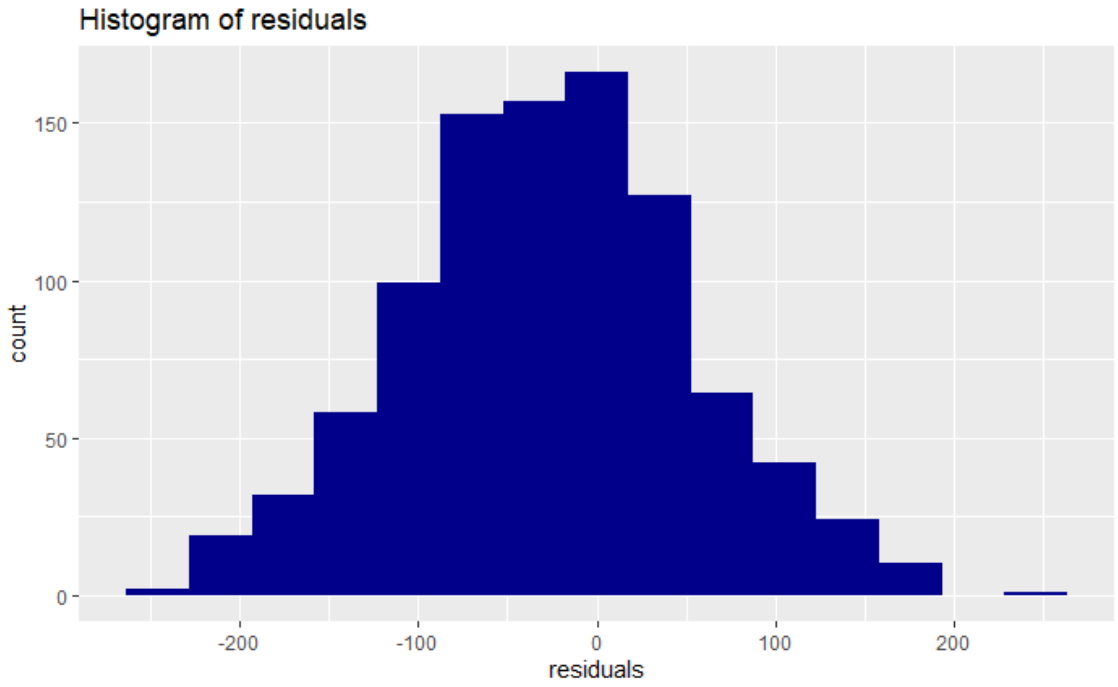
```
Call:
lm(formula = Formula, data = Data_train)

Residuals:
   Min     1Q Median     3Q    Max
  -268    -79    -24     30  90569

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      345.0383   141.2549   2.443   0.0146 *
age               -0.4973     1.7056  -0.292   0.7706
sexmale          -48.3121    47.8117  -1.010   0.3123
bmi               -3.9322     4.0924  -0.961   0.3367
children          16.5376    19.7644   0.837   0.4028
smokeryes        -24.3350    59.8374  -0.407   0.6843
regionnorthwest -101.8284    68.5505  -1.485   0.1375
regionsoutheast  -83.4383    68.1025  -1.225   0.2206
regionsouthwest  -89.8037    68.2401  -1.316   0.1883
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 1470 on 3807 degrees of freedom
Multiple R-squared:  0.001609,  Adjusted R-squared:  -0.00048
F-statistic: 0.7669 on 8 and 3807 DF,  p-value: 0.6321
```

From the above ummary of the linear model we can notice that some of the variable are not significant (*sex*), while *smoking* seems to have a huge influence on *charges*. So, let's Train a model without non-significant variables and check if performance can be improved.

```
[1] "R-squared for first model:0.0016"
[1] "R-squared for new model: 0.0013"
[1] "RMSE for first model: NaN"
[1] "RMSE for new model: NaN"
```

After comparing the 2 models, the performance is quite similar between the 2 models so, this mean that the "Sex" variable is not significant in our data analysis.

- Test the model on the testing data



the errors in the model are close to zero so model predicts quite well.

Figure (6)

Machine Learning Models

Building Random Forest, XGBoost, GBM and SVM Models

xgboost model is the best model.

TrainRMSE <dbl>	TrainRsquared <dbl>	TrainMAE <dbl>	method <chr>

TrainRMSE <dbl>	TrainRsquared <dbl>	TrainMAE <dbl>	method <chr>
534.9401	0.001578744	81.78441	xgbTree

Gradient Boosting: Model Details

TrainRMSE <dbl>	TrainRsquared <dbl>	TrainMAE <dbl>	method <chr>
539.2506	0.002083892	83.59675	gbm

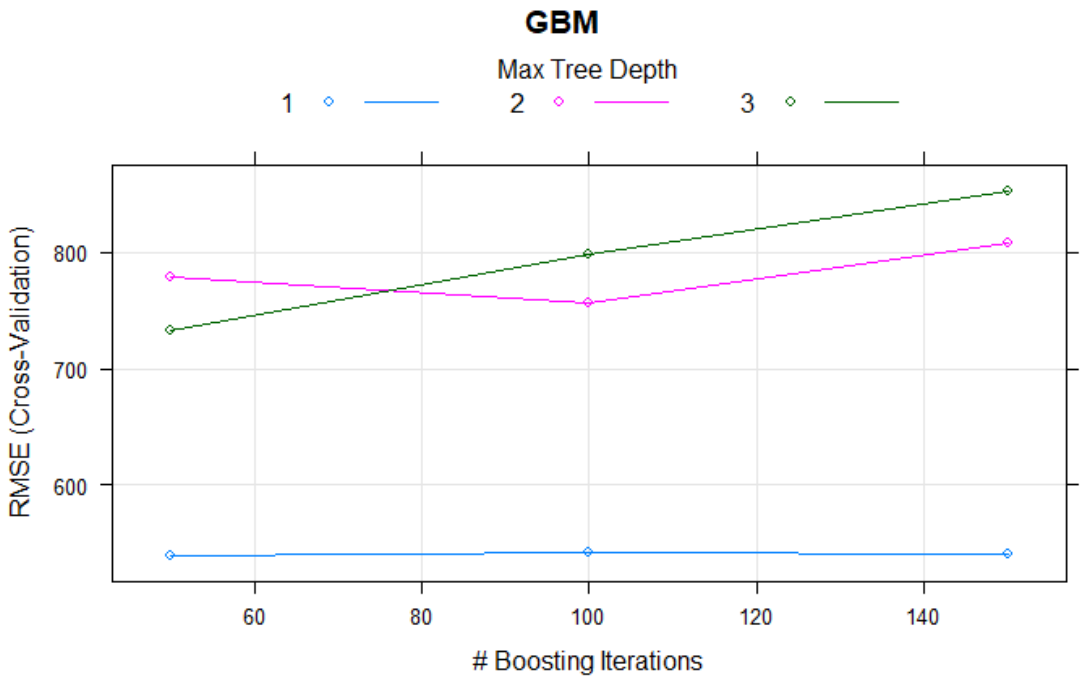


Figure (7)

Support Vector Machine Model Details

TrainRMSE <dbl>	TrainRsquared <dbl>	TrainMAE <dbl>	method <chr>
522.1599	0.001910396	74.65594	svmRadial

RandomForest Model Details and Feature Importance

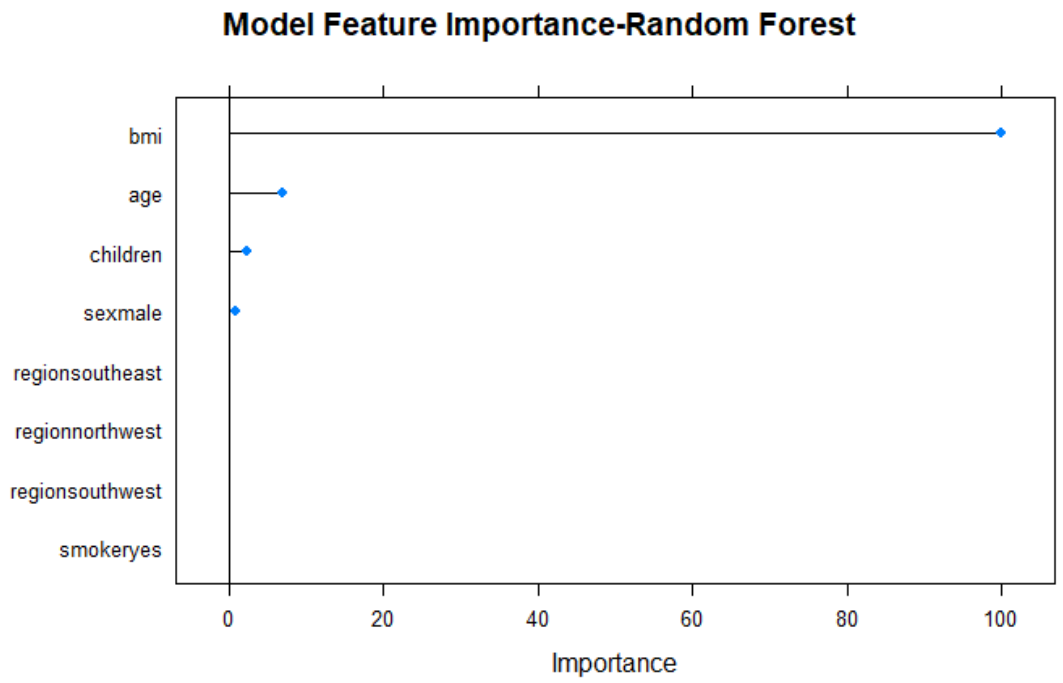


Figure (8)

We can see that the most important feature is whether a person smokes. As expected, the BMI and Age are also high up in the feature importance hierachy possibly due to the high correlation with disease states. Think lung cancer, diabetes, Alzheimer's and hypertension.

We see that the XGB algorithm has an R squared of 0.86 and so does the GBM. The SVM does not do so bad at 0.82. I'll choose the GBM .

*From the analysis we can see that our model does a pretty good job on our "unseen" data. *

The model has a high accuracy for predicting the charges and payments for the next 6 months. And it’s financial position will be even better than now, as the business solutions that we started to provide the decision maker with are effective and the main problems related to the billing system are fixed very well.

Conclusion limitations and future research:

The healthcare industry is thought to be a highly ineffective sector, where one-third of its expenses are squandered and also do not contribute to high quality outcomes. the healthcare system remains to apply commercial and systems design devices to achieve a reliable organized system, data analytics have the potential to boost care, conserve lives and reduced costs by identifying associations as well as recognizing patterns and trends within the data.

Despite the disruptions to conventional practices, all actors in health care should be excited about the possibilities that new data tools will bring. But obtaining this enormous potential is not around the corner and will require overcoming challenges by all of the relevant components of the health care system.

There is no doubt about the benefits of integrating data mining service into existing health care information systems. Certain factors, however, present problems. First, individual identifiable health information cannot be protected if data mining analysis is directly applied upon the patient data collected. Any party who is performing data mining task over these data sets will be able to identify individual patient information. Secondly, quality of the patient data directly influences the outcome of the data mining analysis. we recommend for future researchs to; mprove our models via tuning and Build a model on the most important features.

References

Andreu-Perez, J., Poon, C. C. Y., Merrifield, R. D., Wong, S. T. C., & Yang, G. Z. (2015). Big Data for Health. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1193–1208. https://doi.org/10.1109/JBHI.2015.2450362

Andreu-Perez, J., Poon, C. C. Y., Merrifield, R. D., Wong, S. T. C., & Yang, G. Z. (2015). Big Data for Health. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1193–1208. https://doi.org/10.1109/JBHI.2015.2450362

Aryal, A., Liao, Y., Nattuthurai, P., & Li, B. (2018). The emerging big data analytics and IoT in supply chain management: a systematic review. *Supply Chain Management*, *25*(2), 141–156. https://doi.org/10.1108/SCM-03-2018-0149

Aryal, A., Liao, Y., Nattuthurai, P., & Li, B. (2018). The emerging big data analytics and IoT in supply chain management: a systematic review. *Supply Chain Management*, *25*(2), 141–156. https://doi.org/10.1108/SCM-03-2018-0149

Bateman, H. V. (1998). Information To Users Umi. *Dissertation*, 274.

Behling, O., & Dillard, J. F. (1984). A Problem in Data Analysis: Implications for Organizational Behavior Research. *Academy of Management Review*, *9*(1), 37–46. https://doi.org/10.5465/amr.1984.4277902

Behling, O., & Dillard, J. F. (1984). A Problem in Data Analysis: Implications for Organizational Behavior Research. *Academy of Management Review*, *9*(1), 37–46. https://doi.org/10.5465/amr.1984.4277902

Choi, M. (2014). Book Review: Data Smart: Using Data Science to Transform Information into Insight. *Healthcare Informatics Research*, *20*(3), 243. https://doi.org/10.4258/hir.2014.20.3.243

Choi, M. (2014). Book Review: Data Smart: Using Data Science to Transform Information into Insight. *Healthcare Informatics Research*, *20*(3), 243. https://doi.org/10.4258/hir.2014.20.3.243

Ho, Y.-T. (2020). Building Information Modeling Applications in Construction Management by Ya-Ting Ho Conferral on August 31 , 2020 A thesis submitted to the Faculty of the Graduate School of The University at Buffalo , The State University of New York in partial fulfillme.

Jacobs, J., Roper, L. H., & Van Ruymbeke, B. (2014). From the Editors. *Journal of Early American History*, *3*(2–3), 127–129. https://doi.org/10.1163/18770703-00303003

Jacobs, J., Roper, L. H., & Van Ruymbeke, B. (2014). From the Editors. *Journal of Early American History*, *3*(2–3), 127–129. https://doi.org/10.1163/18770703-00303003

Lim, C., Kim, M. J., Kim, K. H., Kim, K. J., & Maglio, P. P. (2018). Using data to advance service: managerial issues and theoretical implications from action research. *Journal of Service Theory and Practice*, *28*(1), 99–128. https://doi.org/10.1108/JSTP-08-2016-0141

Lim, C., Kim, M. J., Kim, K. H., Kim, K. J., & Maglio, P. P. (2018). Using data to advance service: managerial issues and theoretical implications from action research. *Journal of Service Theory and Practice*, *28*(1), 99–128. https://doi.org/10.1108/JSTP-08-2016-0141

Petajan, E. (2000). Approaches to. *Middle East*, *00*(c), 575–578.

Salazar-Reyna, R., Gonzalez-Aleu, F., Granda-Gutierrez, E. M. A., Diaz-Ramirez, J., Garza-Reyes, J. A., & Kumar, A. (2020). A systematic literature review of data science, data analytics and machine learning applied to healthcare engineering systems. *Management Decision*. https://doi.org/10.1108/MD-01-2020-0035

Salazar-Reyna, R., Gonzalez-Aleu, F., Granda-Gutierrez, E. M. A., Diaz-Ramirez, J., Garza-Reyes, J. A., & Kumar, A. (2020). A systematic literature review of data science, data analytics and machine learning applied to healthcare engineering systems. *Management Decision*. https://doi.org/10.1108/MD-01-2020-0035

Shirley, R. C. (1973). Analysis of Employee and Physician Attitudes Toward Hospital Merger. *Academy of Management Journal*, *16*(3), 465–480. https://doi.org/10.2307/255007

Shirley, R. C. (1973). Analysis of Employee and Physician Attitudes Toward Hospital Merger. *Academy of Management Journal*, *16*(3), 465–480. https://doi.org/10.2307/255007

Submitted, A. T., Partial, I. N., Of, F., Requirements, T. H. E., The, F. O. R., & Of, D. (2011). *a Thesis Submitted in Partial Fulfillment of*.