# Predicting Inpatient Readmissions for Diabetic Patients

Jesse Ashby

# Contents

# Predicting Inpatient Readmissions for Diabetic Patients

## Research question

### Context

Since 2008, the Centers for Medicare & Medicaid (CMS) have emphasized the importance of limiting the time patients with their coverages spend admitted to a hospital as inpatient. One of the primary measures CMS uses to measure this goal is the readmission. A readmission occurs when a patient is discharged from a facility and then admitted again with the same problem within a specified period. The most common period used is 30 days (Centers for Medicare & Medicaid Servers, 2020). CMS motivates healthcare facilities to adopt this goal by penalizing the reimbursement it pays up to 3% (CMS, 2020a).

### Justification

In 2014, Virginia Commonwealth University (VCU) conducted a study of diabetic patients in the Health Facts database from Cerner. Health Facts contains medical record data for patients from hospitals that use the Cerner EMR and participate in the program. VCU obtained deidentified, HIPAA compliant data from Health Facts to conduct their study. The aim of the study was to identify a relationship between HbA1c lab tests conducted during an inpatient stay and the likelihood that patient would be readmitted. VCU did find a significant relationship in the interaction of primary diagnosis and the presence of an A1c test to readmission rates (Strack et al, 2014). However, A1c is considered a measurement of the average blood glucose of a patient for the previous 60-80 days (Nathan et al, 2008) and the study began with a hypothesis that measuring this value could improve patient outcomes and reduce readmissions. No explanation was offered on why this presence of this test would affect those outcomes. The goal of this analysis will be applying multi-variate logistic regression to the VCU dataset to identify

previously overlooked readmission predictors. No bias will be given to any variable. This analysis hypothesizes that other significant relationships exist in the supplied data set. Any additional relationships exposed have the potential for further studies into their nature and clinically actionable potential.

**Hypothesis**

- $H_0$: There are no further significant relationships between a patient's readmission status and the variables collected.
- $H_a$: There is at least one other statistically significant relationship between a patient's readmission status and the variables collected.

## Data collection

**Method**

The authors of the VCU study collected data from Cerner's Health Facts database. Health Facts is a data warehouse owned by Cerner storing patient data collected from participating healthcare organizations. With Cerner's permission, the authors queried Health Facts to find diabetic patients that had an inpatient encounter in a participating hospital. The data was then de-identified to remove protected health information and readmissions were calculated based on a 30-day period between discharge and admission. Only encounters that met the following criteria were included:

1. length of stay between 1 and 14 days (inclusive)
2. One or more laboratory tests during the encounter
3. One or more medications administered
4. One or more diabetic diagnosis (Strack et al, 2014).

After the study had concluded, the dataset was donated to the University of California Irvine's Machine Learning Repository and made available for public download (Dua & Graff, 2019).

## Data extraction and preparation

### Techniques

#### *Data download and SQL Server import*

The dataset VCU compiled was downloaded from the UCI machine learning library and the csv files were extracted. A database was created on a local SQL Server instance to store the data in tables and create backups as the data was cleaned. The IDs_mapping.csv file was loaded into Excel and transformed into a relational format. Both the mapping and detail files were imported into the previously created database using SQL Server's import wizard.

#### *Preliminary cleaning*

Once the VCU data was loaded, it was examined for any values that would hinder logistic regression. The first item found was the presence of redundant levels in three categorical variables: admission_source_id, discharge_disposition_id and admission_type_id. Each of these variables had multiple levels indicating missing values. The redundant levels were collapsed into one level per variable.

The next item was to remove cases that were discharged as expired or to hospice. These patients would not be eligible for a readmission. Cases discharged as intra-facility transfers were also removed. In these cases, the inpatient stay was not ended, another encounter was added for administrative reasons.

As logistic regression requires independent observations, the set should only contain one case per patient_nbr. The VCU study took the case representing the first (index) encounter for each patient. This was not possible in the published dataset as admission and discharge times

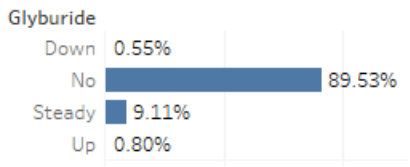
were not included. However, this was approximated using the minimum encounter_id for each patient_nbr. The payer_code and weight variables were removed as the VCU paper recommended. Three variables were removed as they represented the counts of patient care related items, such as num_lab_procedures. The procedures conducted are meaningful, but the number of procedures is not. This would affect the interpretability of the final model.

Two diagnoses related variables were removed: diag_2 and diag_3. These represent two of the secondary diagnoses, but number_diagnoses has a mode of 9 in the dataset. No explanation of which two secondary diagnoses were selected in the VCU study and it could not be determined. A model formed without this understanding may not generalize well when applied to new data with a different selection process for secondary diagnoses.

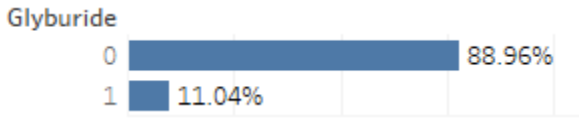
Lastly, seventeen variables were removed because 95% or more cases contained the same value. This type of distribution will risk complete or quasi-complete separation (UCLA, 2020) and not add any value to the model.

Cleaning with univariate analysis

Rare levels in a categorical variable do not add significant value to a logistic regression model (UCLA, 2020). For the remaining variables, all levels that contained less than 5% of the variation were combined into single level that was more than 5%. Tableau histograms were used to check these variances. One example is the variable glyburide. "No" contained 89.53% of the variation, "Steady" had 9.11%, "Up" had 0.80% and "Down" had 0.55%.

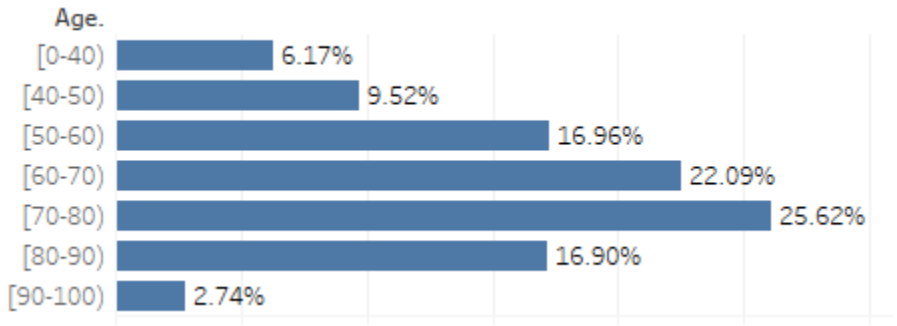


The Up and Down levels were combined with Steady and the new level was relabeled to 1 and the existing “No” label changed to 0.




This was common with most of the medication variables.

Admission_source_id and admission_type_id required a different transformation. The levels in these variables were combined based on the logical flow of a patient through a hospital. Age is an ordinal variable, so only adjacent levels could be collapsed. That left “[90-100)” with its own level and only 2.5% of the variation.




Medical_specialty was collapsed into specialty categories.

Although they started as continuous, number_emergency, number_inpatient and number_outpatient had more than 95% of cases with a value of 0 or 1. This type of skew severely limits the usefulness of a continuous variable. This was addressed by converting the variables into binary variables indicating if a patient had that type of encounter in the 12 months before the admission of the index encounter.

As the subject of this analysis is 30 day readmits, the readmitted variable was also transformed into a binary variable by combining the “No” and “>30” levels into 0 and changing “<30” to 1.

The transformations up to this point had changed the meaning of many of the variables, so the variable names were changed to reflect this.

***Cleaning with bi-variate analysis***

Encounters that were discharged as a transfer to another inpatient facility had a much stronger relationship to readmitted than those that were not discharged as a transfer. The chance of readmission for the data set was around 9%.

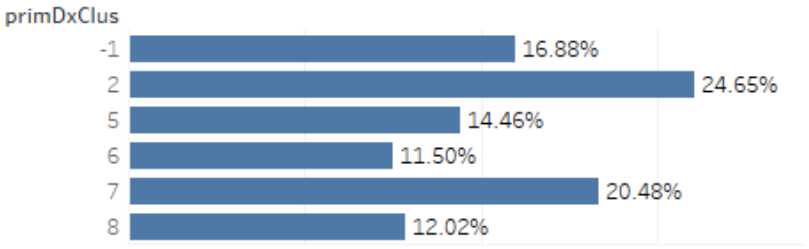

While encounters discharged as a transfer to another inpatient facility had an 18% chance of readmission.


One possible explanation for this is that some transfers were made between two facilities that participate in Health Facts. In this case, the sending facility's encounter would be labeled as readmitted for receiving facilities encounter for the same patient. This type of transfer would not be considered a readmission by funding agencies. The VCU study did not mention how this scenario was handled. Encounters that were discharged to another inpatient facility were removed for this reason. Discharge_disposition_id was then grouped by types of discharges and renamed. Removing these cases did not change the distribution of any other variable enough to warrant removing them.

Diag_1 (now primDx) had over 100 distinct levels, many of which were not represented in both levels of the target variable, readmitted. Grouping these categories hierarchically or by

other medical means would have required specialized medical knowledge that was not readily available. Instead, Greenacre's method for combining categories through clustering was applied. The SAS implementation of this method from their support documentation was used. As Greenacre's method uses chi-squared in its analysis, primDx levels with less than 5 cases in each readmitted level were combined into an "other" category. encId, primDx and readmitted were exported to Excel and then imported into SAS.

Greenacre's method was applied using the SAS developed algorithm (Patetta, 2019) and the results were exported to Excel. The cluster results were imported back into SQL server and a new primDxClus variable was added. A map of diagnoses to cluster was also created for reporting the results of the analysis. There were 4 cluster levels that did not meet the 5% threshold. These were combined into an "other" category labeled -1.




## Analysis

The rest of this analysis was conducted in SAS. The data was imported and split into a training set and a test set for honest assessment. Logistic regression with forward selection of features was used to detect significant interactions. Logistic regression with backward selection of features was used to screen for significant variables. Logistic regression with best subset selection was used to score model candidates and select the champion model.

## Techniques

### Data import and partitioning

The prepared data were exported from SQL Server to Excel and then imported into SAS. The SAS dataset was deterministically sorted by readmitted and encId. Then proc surveyselect was used to split it into a training set (70%) and test set (30%) with a static seed to allow the same split to be repeated.

### Interaction detection and variable screening by forward selection

The search for the model with the best fit to the training dataset began with simultaneously searching for non-significant inputs and interactions in the main effects. Interactions occur when the relationship between a predictor and the target is dependent on the value of a different predictor. When this occurs, coefficient estimates for each of the variables in the interaction must also consider the other variables (Patetta, 2020). The forward selection option of proc logistic was used to test for these interactions. The significance level for model entry was selected based on the Schwarz Bayesian criterion of the model after a candidate variable or interaction was added. Interactions were limited to pairwise as higher order interactions require more compute resources than were available and the odds ratio estimates for higher order interactions are often difficult to interpret or use to take meaningful action. .00409 was used as the p-value for variable entry into the model.

Forward selection revealed 11 effects and interaction terms to be considered. The main effects in these interactions were carried into the analysis to maintain model hierarchy (Patetta, 2020). One disadvantage to forward selection is that as terms are added, they are not reevaluated in later steps. Proc logistic with backward selection was used to address this. Backward selection begins with the full model and removes the variable with the highest p-value over a

selected value. After each removal, variable p-values are reevaluated. Backward selection ends when all remaining variables have p-values under the selected level. The SAS implementation of backward selection includes a fast option that trades accuracy in the regression coefficients for reduced compute complexity. As the goal of this step was only to validate the significance of the effects detected with forward selection, the fast option was used. The same method was used to determine the significance level needed for an effect to stay.

Backward selection found that all effects detected by forward selection maintained their relevance.

Joint Tests

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| los | 1 | 32.9904 | <.0001 |
| ipEnc | 1 | 96.2820 | <.0001 |
| edEnc | 1 | 25.0177 | <.0001 |
| primDxClus | 5 | 87.9292 | <.0001 |
| metformin | 1 | 11.5832 | 0.0007 |
| dmRx | 1 | 39.8870 | <.0001 |
| race | 2 | 9.8312 | 0.0073 |
| age | 6 | 13.9199 | 0.0305 |
| dischargeDisp | 2 | 49.0096 | <.0001 |
| admitSpecialty | 4 | 3.7548 | 0.4402 |
| race*age | 12 | 29.1560 | 0.0037 |
| age*dischargeDisp | 12 | 30.0436 | 0.0028 |
| discharge*admitSpeci | 8 | 27.5076 | 0.0006 |
| ipEnc*age | 6 | 32.7057 | <.0001 |
| los*ipEnc | 1 | 11.1941 | 0.0008 |

Each of the non-significant main effects were included in a significant interaction so were not removed from consideration in the final model.

**Model selection by best subset selection**

The final model was selected by examining all possible candidate models from the main effects and interactions that had not been removed in prior screening. Each model was judged based on a modified c statistic for the test data set and the improvement it offered over the best

model with 1 less variable. Adding variables to a logistic regression model increases the complexity and can decrease interpretability (Patetta, 2020). A model that balances complexity with ability to predict the chance for readmission achieves the goal of this analysis to find potential clinically actionable insights.

Best subset selection was used to examine all possible candidate models. SAS' implementation of best subset selection builds all possible models and returns the top n models for each number of variables based on the chi-squared statistic (SAS, 2019). In previous model building, the class statement from proc logistic was used with categorical variables to apply the coding necessary for logistic regression. This feature is not available with best subset selection and had to be completed manually.

***Dummy coding***

Logistic regression requires binary or continuous input variables (UCLA IDRE, 2020). If categorical inputs are to be used, they need to be coded. The coding method selected for this analysis was reference or dummy coding. In dummy coding, a reference level is selected for each categorical input and new binary variables are created for all other categories. The variable corresponding to the original value is set to 1 for each case and all other dummy variables are set to 0. After the model is completed, the resulting coefficients for the remaining variables are in reference to the reference level (Grotenhuis & Thijs, 2015).

For example, the variable dischargeDisp was coded in this manner. The original variable was categorical with levels "other", "home" and "transfer". "Other" was selected as the reference level and 2 new variables were created: dischHome and dischTrans. In a case where the patient was discharged to home from the index visit, dischHome would be 1 and dischTrans

would be 0. The chart below displays the values of each of the dummy coded variables for each value of dischargeDisp.

| | dummy variable | |
|---|---|---|
| dischargeDisp value | dishcHome | dischTrans |
| Home | 1 | 0 |
| Transfer | 0 | 1 |
| Other | 0 | 0 |

If the final odds ratio for the dischHome variable was 0.5, that would indicate a patient who was discharged to home from the index visit was half as likely to be readmitted.

Dummy coding was appropriate for this analysis as the categorical variables remaining each had levels indicating an absence or unknown value. These levels were assigned as the reference levels. Each combination of non-reference levels in variables from interactions was also dummy coded (Grotenhuis & Thijs, 2015). A new database table was created to store and persist the new dataset. The new table was exported to Excel and then imported into SAS. It was split into training and test sets using the same deterministic sorting, ratio and seed as was previously used.

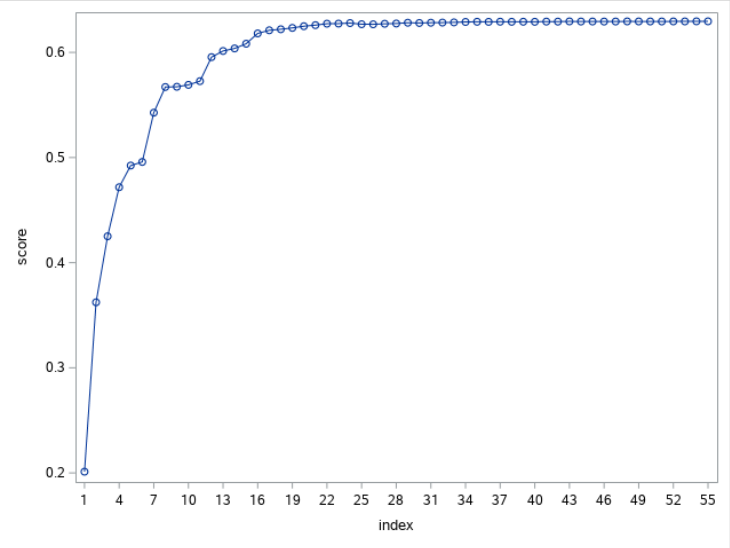
***Model scoring and selection***

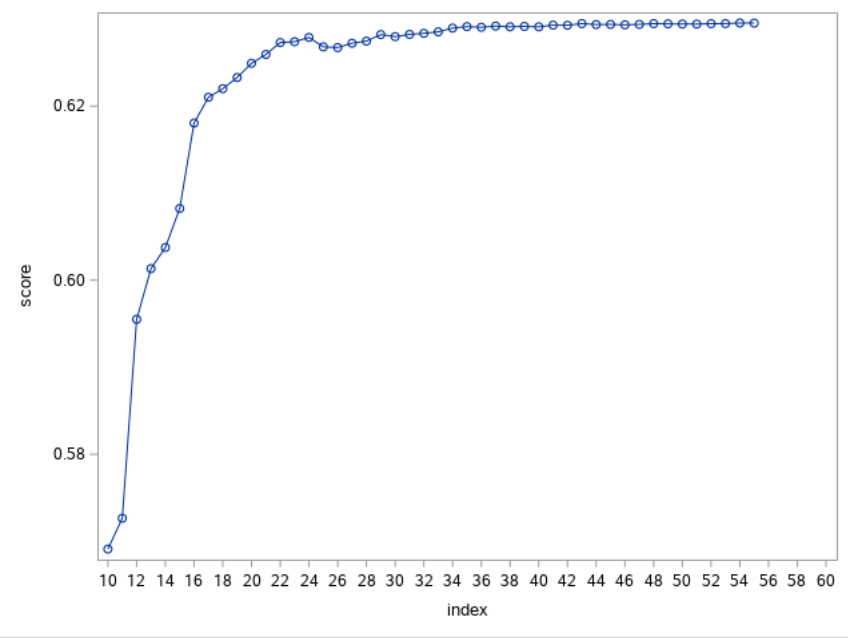
Proc logistic with best subset selection was used to find the top model for each number of variables up to the full model. Each top model was then scored against the data set that had been partitioned for testing.

No model predicted a readmission. However, each case in the scoring sets was assigned a probability of readmission. This was used to build a score for each model. Each case that resulted in a readmission was compared against all cases that did not. The model's score was

then determined by the percentage of cases that were not readmitted and had been assigned a lower probability of readmission than the cases that were readmittted.

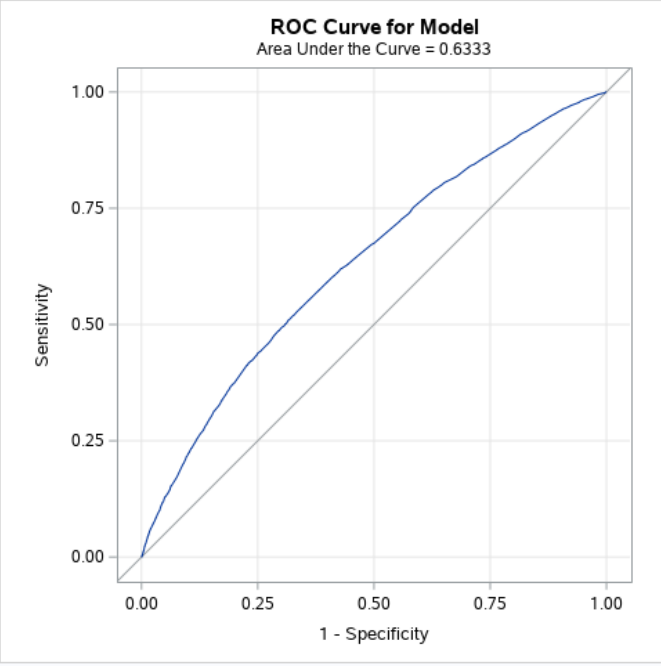
Graphs of these scores were then reviewed to determine the champion model considering both score and complexity.



The 17-variable model was the lowest index to score over 62% and was a good estimate of when returns on accuracy began to diminish as additional variables were added. Odds ratio estimates and the ROC curve for this model are below.

| Odds Ratio Estimates | | | |
|---|---|---|---|
| Effect | Point Estimate | 95% Wald Confidence Limits | |
| ipEnc | 1.806 | 1.658 | 1.967 |
| edEnc | 1.359 | 1.216 | 1.519 |
| metformin | 0.875 | 0.800 | 0.955 |
| dmRx | 1.259 | 1.156 | 1.371 |
| dx2 | 0.840 | 0.770 | 0.917 |
| dx5 | 0.742 | 0.666 | 0.828 |
| dx6 | 1.270 | 1.150 | 1.403 |
| dx8 | 0.572 | 0.500 | 0.654 |
| age8090 | 1.420 | 1.240 | 1.625 |
| raceCage5060 | 1.391 | 1.172 | 1.651 |
| discHomeAge5060 | 0.647 | 0.550 | 0.762 |
| discHomeAge8090 | 0.811 | 0.688 | 0.957 |
| discTransAge6070 | 1.768 | 1.491 | 2.097 |
| discTransAge7080 | 1.808 | 1.590 | 2.055 |
| discHomeAdmitOther | 0.833 | 0.749 | 0.926 |
| discTransAdmitPcp | 1.416 | 1.218 | 1.646 |
| discTransAdmitSurg | 1.795 | 1.297 | 2.485 |



## Data summary and implications

### Results

Unlike the VCU study, this analysis did not find that whether a patient's A1c was tested during the index encounter had any ability to predict that patient's readmission. However, it did discover possible correlations between a patient's inpatient and emergency room encounters in the 12 months prior to the index encounter. A patient with an inpatient encounter in that timeframe is 1.8x as likely to be readmitted while a patient with an ER encounter is 1.4x as likely. It also found several scenarios where a patient who was discharged to home is less likely to be readmitted while some scenarios where a patient is transferred to another, non-inpatient facility are more likely. As expected, elderly patients are also more likely to be readmitted. One interesting result is that a patient who was prescribed metformin either before or during the index encounter was only 0.88x as likely to be readmitted. Another interesting result was that the Caucasians between 50 and 60 years old were 1.4x as likely to be readmitted. This analysis and the VCU study did agree on perhaps the most expected finding of primary diagnoses being related to a readmission. This is seen outside of the diabetic population and CMS has recently adjusted reimbursements and penalties based on the primary diagnosis of the index encounter (Walker, 2016). Using these predictors, this model was able to identify patients with higher chances for readmission correctly in 62% of cases.

#### *Limitations of results*

These results are limited by several factors. The dataset used has a long lineage that originated from many disparate sources. This along with the de-identification processes necessitated by HIPAA make it impossible to validate the accuracy of the set to the data contained in these patient's charts. Several assumptions and inferences were made that could

yield different results than if the true nature of each element were known. The sample was also not collected from a randomized trial. This means that the results are not guaranteed to apply to another sample taken from a larger population. The scope of these results is defined only by the criteria the VCU study used to extract encounters from Cerner's Health Facts database.

**Recommendation for course of action**

Given the results of this analysis, providers wishing to limit readmissions in their diabetic patients may consider prescribing Metformin before or during inpatient stays where diabetes related symptoms are the cause for hospitalization. Many of the strongest relationships to readmissions were also in variables that relate to a patient's admission source and discharge disposition. Especially troubling was that a patient who was admitted for surgery and then transferred to another non-inpatient facility for care was 1.8x more likely to be readmitted. This dataset was limited to the broad scope of surgery. Hospital administrators and surgeons may examine the types of procedures performed and where those patients are discharged to for possible avenues to reduce readmissions. Finally, the diagnosis clustering could be reviewed by providers for possible unexpected associations between diagnoses as they relate to readmissions.

**Future study**

As this data was collected over a 10-year period, ending near the time CMS implemented penalties and incentives based on readmissions, repeating the study with the goal of finding what difference the added focus has made could be warranted. Using the CMS definition for a readmission in future data collection would be advisable. Exploring the relationship between Metformin and readmissions could also yield interesting results. Several other drugs were included in the study, but the data on these were too sparse to be meaningful. Focusing on a

select few of these drugs and their relationship to readmissions could possibly yield additional insight, especially if interactions with a primary diagnosis were discovered.

## References

Centers for Medicare & Medicaid Services (2020), *Outcome Measures* retrieved from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/OutcomeMeasures 2020-09-20

Centers for Medicare & Medicaid Services (2020a), *Hospital Readmission Reduction Program (HRRP)*, retrieved from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction-Program 2020-09-20

Centers for Medicare & Medicaid Services (2019), *Quality Measures Fact Sheet Hospital-Wide All-Cause Unplanned Readmission Measure* retrieved from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/OutcomeMeasures 2020-09-20

Castro, K. (2018). *Advantages of Database Management Systems.* Tutorials Point. Retrieved from: https://www.tutorialspoint.com/Advantages-of-Database-Management-System

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Grotenhuis, M and Thijs, P (2015). *Dummy variables and their interactions in regression analysis: examples from research on body mass index*. Cornell University. arXiv: 1511.05728 [stat.AP]

Hemedinger, C. (2020). *Through the years: SAS Enterprise Guide versions* SAS Blogs. Retrieved from https://blogs.sas.com/content/sasdummy/2020/08/25/eg-versions-through-years/ 2020-10-01.

Jackson, J. E. 1991. A Users Guide to Principal Components. New York: John Wiley & Sons.

Lopez, P (2020). *Transaction Locking and Row Versioning Guide*. SQL Docs. Retrieved from: https://docs.microsoft.com/en-us/sql/relational-databases/sql-server-transaction-locking-and-row-versioning-guide?view=sql-server-ver15

Nathan D., Kuenen J, Borg R, et al (2008): *Translating the A1c assay into estimated average glucose values*. Diabetes Care. 2008 Aug:1473-1478

Narkhede, S. (2018). *Understanding Logistic Regression.* towards data science. Retrieved from: https://towardsdatascience.com/understanding-logistic-regression-9b02c2aec102 2020-09-30

Patetta, M (2019). *Predictive Modeling Using Logistic Regression.* SAS virtual learning environment. Retrieved from: https://vle.sas.com/course/view.php?id=1523 2020-09-30 (subscription required).

PRNewswire (2019). *SAS is No. 1 in Advanced and Predictive Analytics Market Share,* Market Technology Insights. Retrieved from: https://martechseries.com/predictive-ai/sas-no-1-advanced-predictive-analytics-market-share-says-analyst-report-2/ 2020-10-01.

SAS (2019). *The LOGISTIC Procedure.* SAS Support Retrieved from SAS Help Center: Effect-Selection Methods 2020-12-26

SAS (2020). *Cluster Analysis* SAS Support Retrieved from https://support.sas.com/rnd/app/stat/procedures/ClusterAnalysis.html 2020-09-30

Strack, B.; DeShazo, J. P.; Gennings, C.; Olmo, J. L.; Ventura, S.; Cios, K. J. & Clore, J. N. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records *BioMed Research International, Hindawi Publishing Corporation*, 2014, 2014, 781670

Tableau (2020). *Tableau recognized as a Leader in the 2020 Gartner Magic Quadrant for Analytics and Business Intelligence Platforms.* Why Tableau? Retrieved from https://www.tableau.com/reports/gartner 2020-09-21.

Texas Health and Human Services (2017), *Texas DSRIP Measure Bundle Protocol Demonstration Years 7-10* retrieved from https://hhs.texas.gov/sites/default/files/documents/laws-regulations/policies-rules/Waivers/medicaid-1115-waiver/1115-medicaid-waiver-tools-guidelines-regional-healthcare-partnership-participants/dy7-10-final-mbp.pdf 2020-09-20

UCLA (2020). *FAQ WHAT IS COMPLETE OR QUASI-COMPLETE SEPARATION IN LOGISTIC/PROBIT REGRESSION AND HOW DO WE DEAL WITH THEM?.* UCLA Institute for Digital Research & Education. Retrieved from: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-complete-or-quasi-complete-separation-in-logisticprobit-regression-and-how-do-we-deal-with-them/

UCLA IDRE *Coding systems for categorical variables in regression analysis.* UCLA Institute for Digital Research & Education Statistical Counseling. Retrieved from: Coding Systems for Categorical Variables in Regression Analysis (ucla.edu) 2020-12-06.

Walker, B. (2016). *The Top 5 Causes of Hospital Readmissions – and How to Prevent Them.* PatientBond Blog. Retrieved from: https://insights.patientbond.com/blog/the-top-5-causes-of-hospital-readmissions-and-how-to-prevent-them