[Add literature review here]

[Data description; the variable definition; we use clustering method to reduce the variable with many levels; paste the clustering graph]

**[Back ground, introduction]**

30-day readmission means the patient was re-admitted to the hospital within 30 days after the initial discharge. We propose here applying the recent advancements in machine learning to predict the 30-day readmissions. We aim to provide a predictive model that can be useful to related entities including hospitals, health insurance actuaries and Medicare to help to reduce the cost, improving the clinical outcome of our healthcare system.

About 20% of Medicare patients are readmitted within 30 days [1], 34% readmitted within 90 days [2]. More than $41billion [3] spent on diabetes patients only, by hospitals in the United States in 2011, on their 30-day readmission, not even mention patients with other diseases. $17billion could be saved per year for avoidable readmissions of Medicare patients alone [4], note that there are other 44% of the readmissions patients are not Medicare patients [5].

One readmission can double the cost of patient care. It costs Medicare $15,000 for a patient who admitted once, but $33,000 for a patient who readmitted once, according to the statistics in 2012 [1]. It is predicted that Medicare will be insolvent in 2026 [6] if we don’t make an effort.

The prediction of readmission is of significance in multiple aspects. From the patients’ view, this can improve the quality of recovery. From the view of the hospitals and Medicare system, 30- day readmission means low health outcome, ineffective treatment, increasing the burden of the patients and the caregivers, causing waste for valuable and expensive healthcare resources. The research we propose here can help in reducing the huge spending of the system. From the view of hospitals, this can help hospitals improving the effectiveness of their healthcare service and also reducing the penalties they may get from the Centers for Medicare & Medicaid Services (CMS) because of the high rate in readmission. 78% of the hospitals in the US received the Hospital Readmission Reduction Penalty because of exceeding the readmission rate limit set by CMS [7] in the financial year 2015. Also, the development of this predictive model can help further analyze the factors that cause readmission, which can help to increase the quality of healthcare the patients receive in the future. Once the high-risk readmission discharges are identified, more resources can be spent on them to reduce their readmission rate, for example, contacting them within 24~48 hours, depending on the patient’s risk of readmission to make sure they are following the discharge plan. Discharging the high-risk readmission patients to the most appropriate transitional care setting can also be a choice to reduce their readmission after the model predicted that the patient was at high risk of readmission prior to discharge.

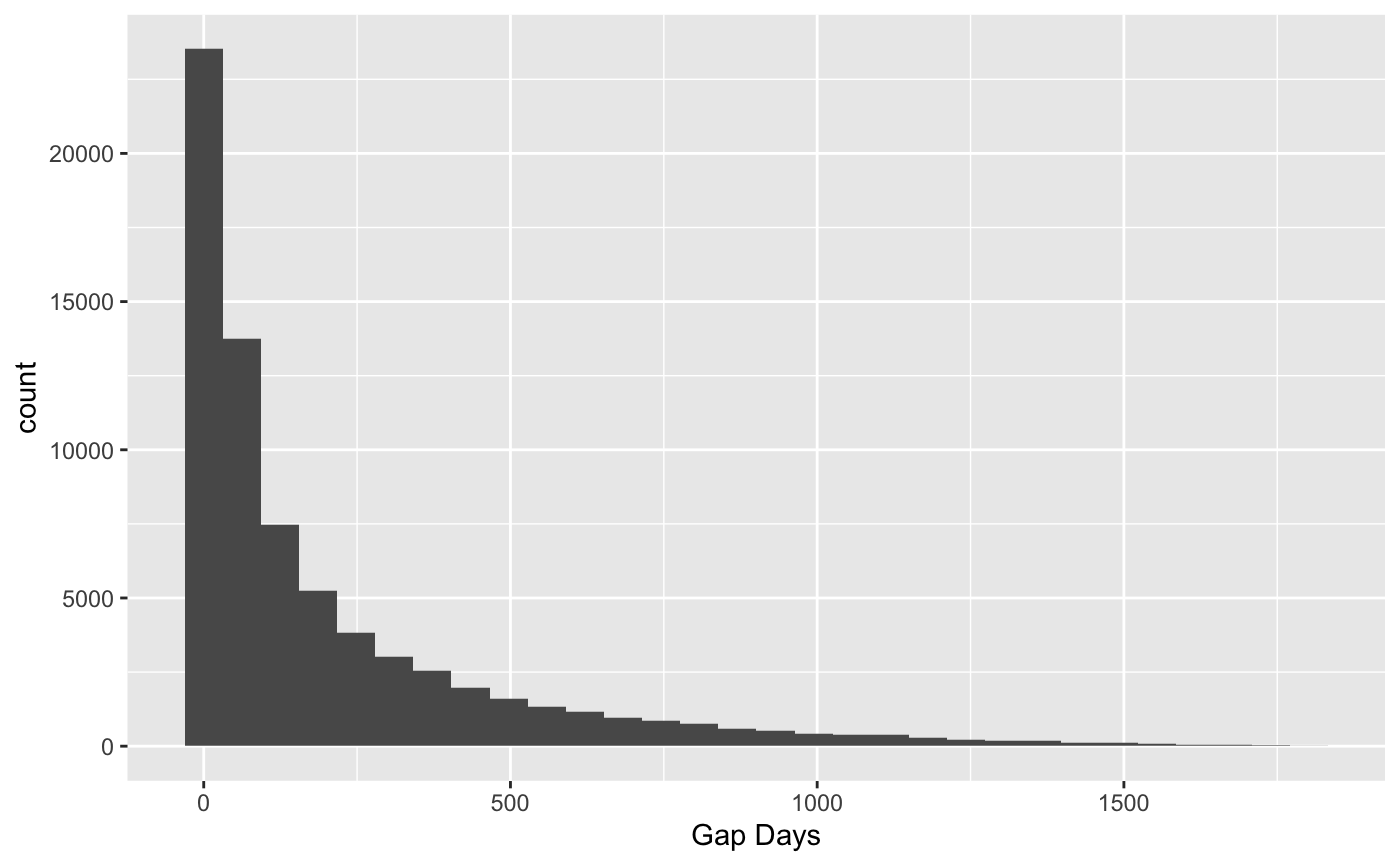
**[Literature Review]**

CMS published 3 models in 2008 to predict readmission for three types of diseases. The performances of their models are relative poor, which is c-statistics (a measurement of model accuracy) 0.63 for acute myocardial infarction (AMI), 0.61 for congestive heart failure (CHF) and 0.63 for pneumonia [8, 9, 10]. Note that only when c-statistics is greater than 0.8, the model can be considered as a strong model. In 2011, a study [11] reviewed 30 models in readmission prediction, it turns out most of them didn’t have high enough prediction accuracy, with c- statistics ranging from 0.55-0.83. In 2015, a study [12] reviewed the machine learning models including Random Forest, LogitBoost etc. on readmission projection, their Area under the curve (AUC, a measurement of model classification accuracy) ranging from 0.75~0.8, which are fair but not strong accuracy. In 2018, Health Catalyst published a report says they developed a model with AUC=0.784 [13] that met the goal of their hospital client, which outperformed the LACE index method that is currently being used in many hospitals.

[Task 1]

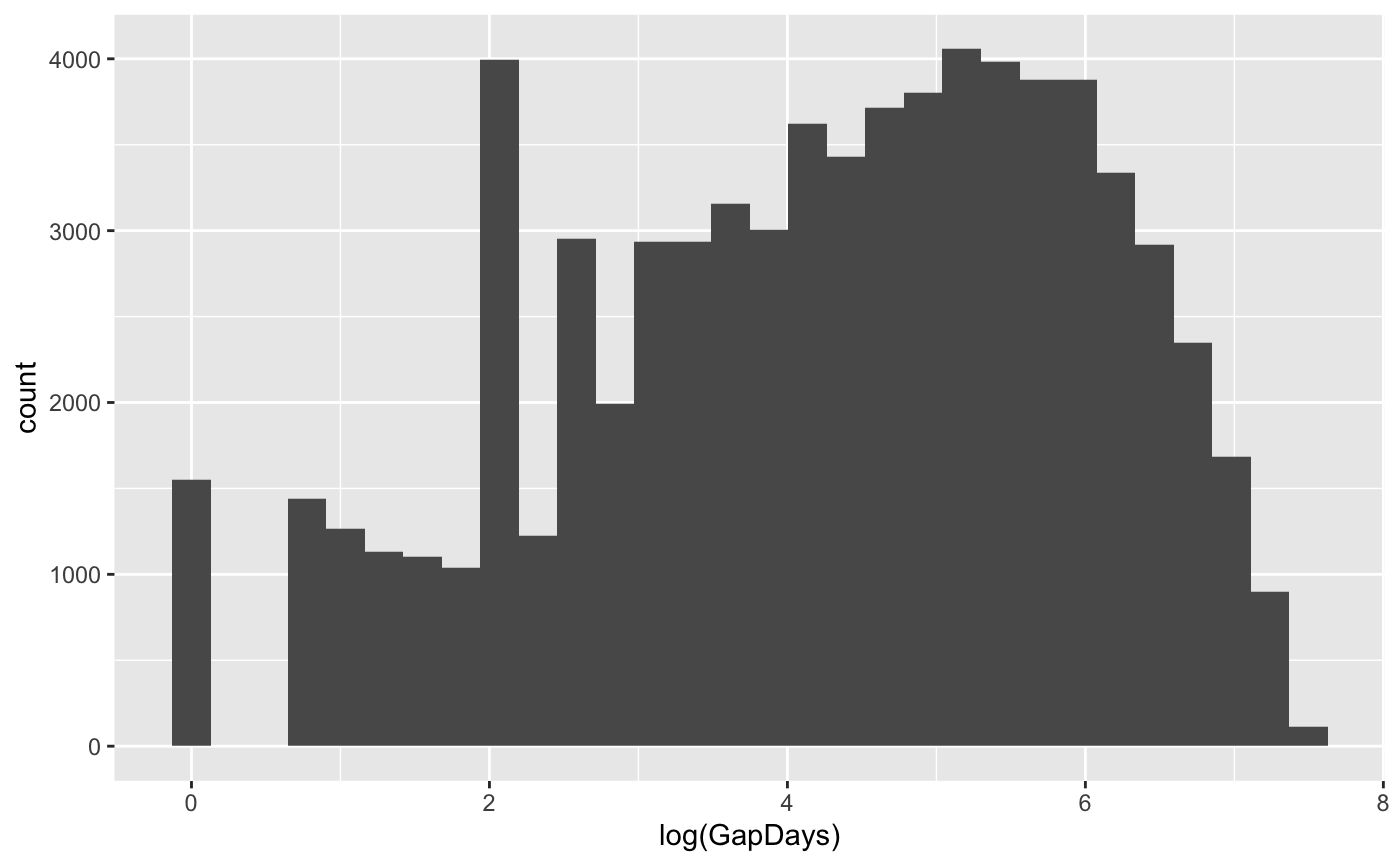
We define the target variable Gap Days (GD) as the days between this discharge to the next hospital visit (readmission) for the same patient. The smaller Gap Days means faster readmission, which indicates bad healthcare outcome. [we try to predict …]

Figure 1 is the histogram of the Gap Days, which is right skew distributed.



**Figure 1.**

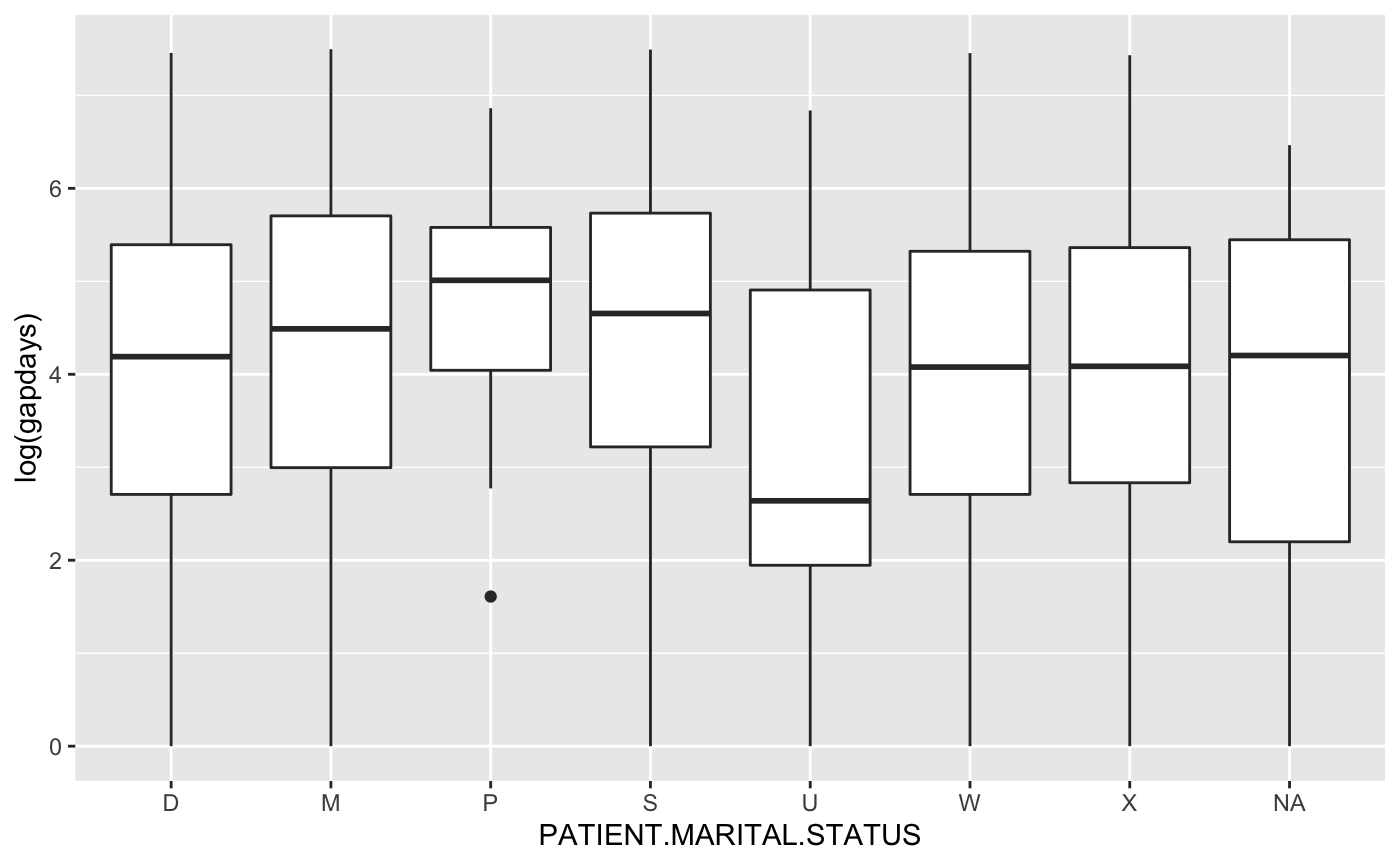
As a common practice to deal with right skewed distribution, we take the natural log of the Gap Days, so that its distributed will be transformed to a normal-like distribution (Figure 2), which makes a later analysis more convenient. The reason it’s GapDays+1 is …



**Figure 2.**

To explore the relationship of each predictor to the target variable, we did a boxplot for each predictor split by it levels. One of the plots is given in Figure 3. Where D means divorced, M for married, P for domestic partner, S for single, U for unknown, W for widow, X for legally separated, NA is missing data. From this figure, there is higher Gap Days for marital status code = P (Domestic partner), lower for code W (Widow). This makes sense because research shows being widowhood undermines people’s health [Williams, K. 2004]. The larger difference among the levels implies the better this predictor is in terms of predicting the target variables. Other variables with different target values within the levels are:

* HOSPITAL.SERVICE.CODE: low target value for NP, IVT, WND, WWC.
* ADMIT.SOURCE: high for ADMIT.SOURCE=9, 1, 6. Low for 2, 4, 5.
* PATIENT.SEX.CODE: high target value for female.
* PATIENT.RACE.CODE: high for target value for PATIENT.RACE.CODE=T, low for D.
* Income Level: low for income less than 10k, or greater than 90k. High for income in 50-60k.



**Figure 3.**

By looking at the mean and median of each levels of the predictors, some aditional differences are observed:

* Age: high for age 0-10, low for 100+. The target value decreases with age increases.
* Patient.Days: low for Patient.Days 20-30.

An example of using Patient.Days is shown below:

| **Patient.Days**  <ord> | **mean**  <dbl> | **median**  <dbl> | **n**  <int> |
| --- | --- | --- | --- |
| 1 | 0-19 | 4.34939 | 4.54329 | 63322 |
| 3 | 30+ | 3.86552 | 3.98744 | 26 |
| 2 | 20-30 | 2.59868 | 2.07944 | 2340 |

[Task 2 Reduce the number of factor levels where appropriate]

To further explore the relationships between the predictors and the target, we computed the statistics of the target variables in each level of the predictors. One example is given in Table 1.

| **PATIENT.MARITAL.STATUS**  <chr> | **mean**  <dbl> | **median**  <dbl> | **n**  <int> |  |
| --- | --- | --- | --- | --- |
| D | 4.051373 | 4.204693 | 8827 |  |
| M | 4.308189 | 4.488636 | 29433 |  |
| P | 4.620087 | 4.986553 | 22 |  |
| S | 4.402142 | 4.653960 | 23839 |  |
| U | 3.307373 | 2.708050 | 129 |  |
| W | 4.035886 | 4.077537 | 7855 |  |
| X | 4.051213 | 4.077537 | 1599 |  |
| *NA* | 3.784187 | 4.216797 | 20 |  |

**Table 1.**

Based on the similarity the target behaves in two levels, we can combine the levels, to reduce the levels within a predictor so that the predictive power of the model can be increased. The marital status D, U, W, X, NA are combined to one level renamed as Unmarried, status P (Domestic partner), S (Single) are combined to another level called Single, M stays as the other level called Married. The reason D, W, X can be combined together is the target has similar mean and median on these three levels. U and NA has very few data, so we decide to group them to Unmarried also, because of their low mean and median of target value. It’s clearer after levels are reduced to 3: ranging from low, middle to high GD. [Can add some explanation here about why P and S has good GapDays, maybe because they are younger people who are not married yet]

[Should change here about what combinations made]

The following other combinations are made:

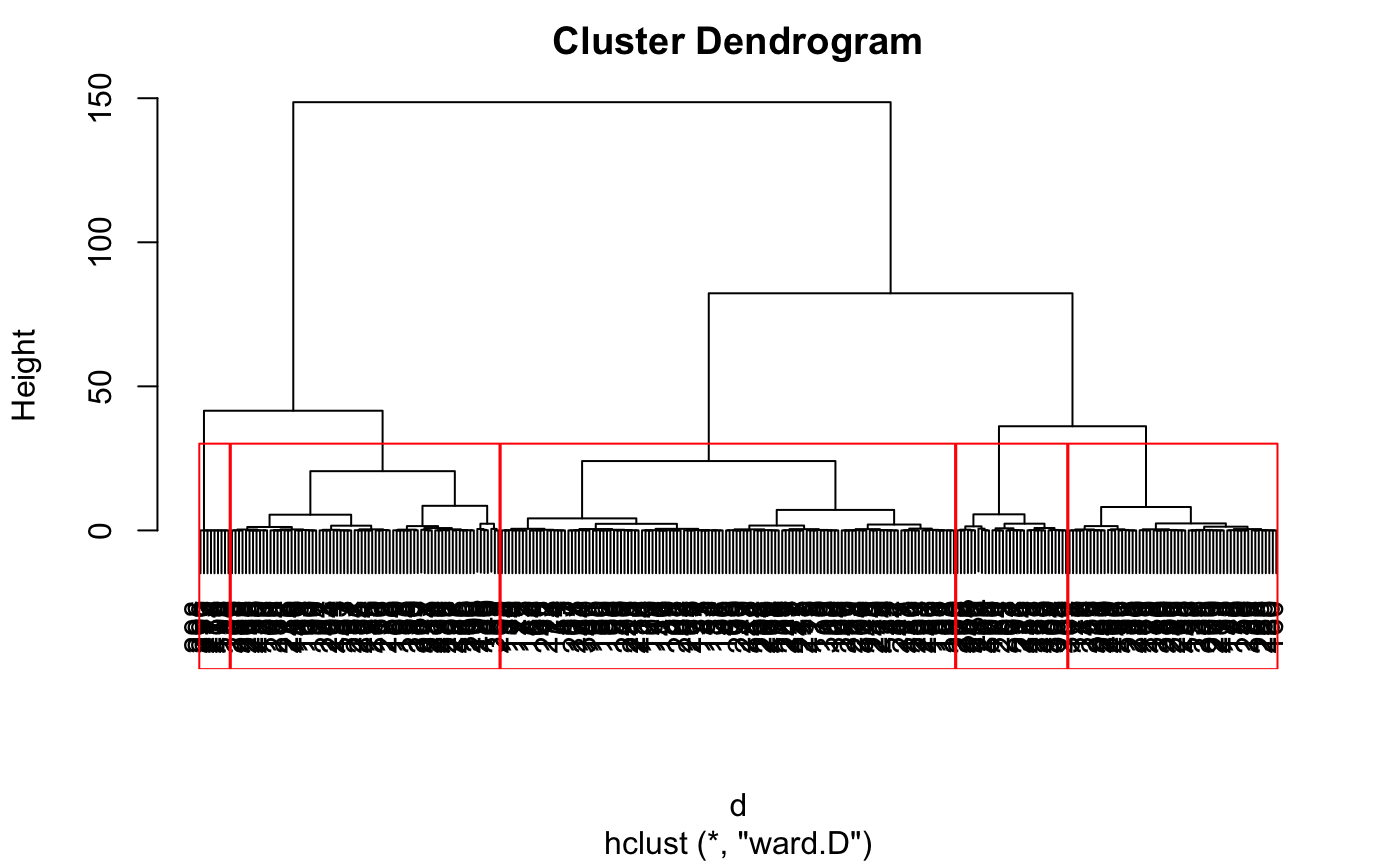
* Patient Days less than or equal to 19 are combined as one level called 0-19. Patient Days from 20 to 30 are grouped to the level called 20-30. The others are grouped to 30+.
* Patient\_Race\_Code: W and X are combined to a new group W-X, because they similar relationship to target. B stays as one level B. All other codes are grouped as Other, because the data counts are small.
* Hospital\_Service\_Code: WND, WWC, PAT, SO, NP, IVT are combined as Level1; DX, EOB, MED, OBS, OPB, OPV as Level2; ER as Level3; CTH, NB, OPS as Level4. The mean and median of target increases from Level1 to Level4.
* Discharge\_Status: Code 2, 5, 43, 62, 63, 65, 70, 72 are combined as D\_level1; Code 1, 7, 30 are combined as D\_level2; others are combined to D\_level3.

| **HOSPITAL.SERVICE.CODE**  <chr> | **mean**  <dbl> | **median**  <dbl> | **n**  <int> |  |
| --- | --- | --- | --- | --- |
| CTH | 4.5409533 | 4.700480 | 481 |  |
| DX | 3.9041673 | 3.850148 | 1368 |  |
| EOB | 3.8209554 | 3.258097 | 5 |  |
| ER | 4.4087035 | 4.634729 | 43964 |  |
| IVT | 2.8353134 | 2.708050 | 422 |  |
| MED | 4.1696288 | 4.219508 | 6617 |  |
| NB | 5.8154321 | 6.059112 | 184 |  |
| NP | 2.3680183 | 1.354025 | 8 |  |
| OBS | 4.1247930 | 4.262680 | 2291 |  |
| OPB | 3.9781478 | 3.806662 | 429 |  |
| OPS | 4.6064572 | 4.844187 | 10704 |  |
| OPV | 3.9640359 | 3.828641 | 381 |  |
| PAT | 0.5364793 | 0.000000 | 3 |  |
| SO | 3.6505878 | 3.496508 | 21 |  |
| WND | 2.6781038 | 2.079442 | 4808 |  |
| WWC | 3.0922665 | 2.708050 | 38 |  |
|  |  |  |  |  |
|  |  |  |  |  |

| **DISCHARGE.STATUS**  <dbl> | **mean**  <dbl> | **median**  <dbl> | **n**  <int> |  | |
| --- | --- | --- | --- | --- | --- |
| 1 | 4.2820393 | 4.4773368 | 66479 |  | |
| 2 | 4.6453560 | 4.7706846 | 1618 |  | |
| 3 | 3.4969854 | 3.4011974 | 1081 |  | |
| 4 | 3.7309671 | 3.8918203 | 52 |  | |
| 5 | 5.3160783 | 5.5243963 | 4 |  | |
| 6 | 3.8399538 | 3.8712010 | 1345 |  | |
| 7 | 4.2257911 | 4.4188406 | 521 |  | |
| 9 | 0.8379309 | 0.6931472 | 5 |  | |
| 21 | 3.0031845 | 3.0445224 | 11 |  | |
| 30 | 4.0928598 | 4.0253517 | 3 |  | |
| 43 | 4.9400071 | 4.9400071 | 2 |  | |
| 50 | 3.8651315 | 3.8903604 | 6 |  | |
| 51 | 2.6944429 | 2.7080502 | 3 |  | |
| 62 | 4.6817401 | 4.7184989 | 101 |  | |
| 63 | 4.7829277 | 4.5926138 | 16 |  | |
| 64 | 3.2614329 | 3.1567740 | 48 |  | |
| 65 | 4.5530608 | 4.4829866 | 416 |  | |
| 70 | 4.5528307 | 4.7184989 | 7 |  | |
| 72 | 6.5581978 | 6.5581978 | 1 |  | |
| 81 | 2.9444390 | 2.9444390 | 1 |  | |
| 82 | 3.6694441 | 3.6694441 | | 2 |  |
| 83 | 3.8333451 | 3.8333451 | | 2 |  |
|  |  |  | |  |  |
|  |  |  | |  |  |
|  |  |  | |  |  |

The predictors like “Doctor.Number” has 307 levels, which means the are 307 different doctors in this dataset. It is not feasible to combine the groups manually. Therefore we apply the clustering method to reduce its levels. The idea is the levels with similar mean of target will be merged. The clustering algorithm used is Ward’s method:

Where is the data, which are the means of target in each 307 levels of the predictor “Doctor.Number”. is the total distance of the all the points in data to the center of the merged cluster of and . The distance between two clusters are measured by . [should be more precise about the description here]

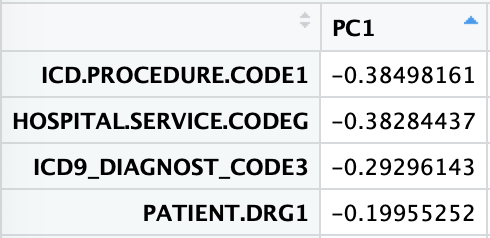
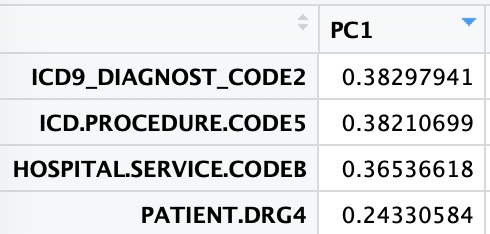


**Figure 4.**

[Can add a table of Map values here]

[Can run PCA based on several variable that could have correlations: for instance, Surgeon and Doctor.Number could be related, PATIENT.DRG and HOSPITAL.SERVICE.CODE and ICD9.PROCEDURE.CODE and ICD9\_DIAGNOST\_CODE could be related]

Table 2 shows the largest loadings in PC1. We can use these loadings to generate a new feature to replace the 3 predictors the levels are from “HOSPITAL.SERVICE.CODE”, “PATIENT.DRG”, “ICD9\_DIAGNOST\_CODE”. This new feature is the linear combination of the levels from table 2, and the coefficients are the loadings.

**Table 2.**

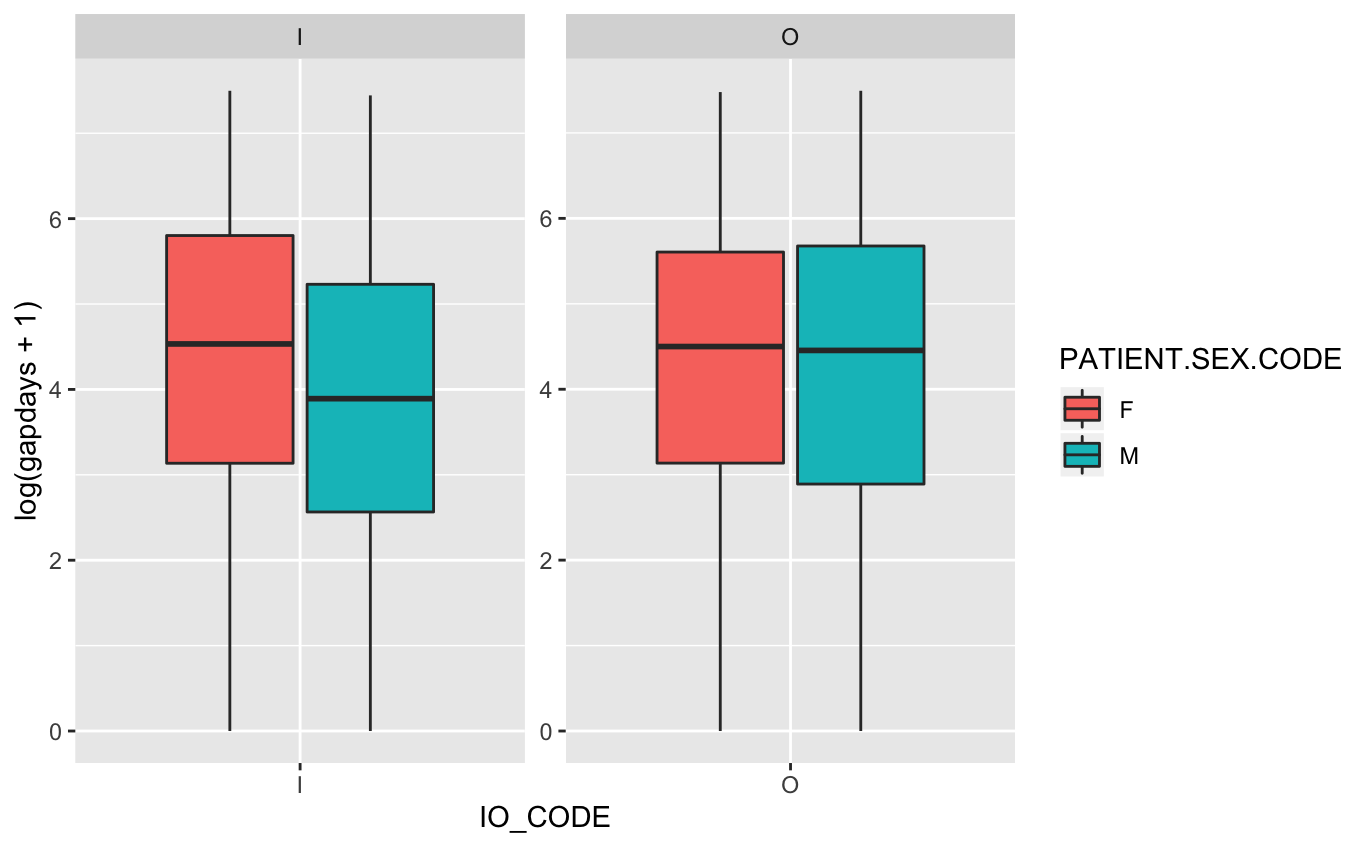
**Interaction**

When we say two predictors have interaction of each other, that means the change of value in one predictor alters the how the other predictor affect the target. We investigated the interactive of IO\_CODE vs PATIENT.SEX CODE, and IO\_CODE vs Age. IO\_CODE stands for inpatient/outpatient code, where code I is inpatient, O is outpatient. From **Figure 5**, we can see the interaction exists between IO\_CODE and PATIENT.SEX CODE. For IO\_CODE=I, the target values are significantly higher in female than in male, while for IO\_CODE=2, the target values are almost the same in both genders. The reason could be the inpatient service has better outcome for female than male, while outpatient service has same outcome on both genders. Or it could be female inpatients involves birth deliveries, who stayed in hospital not because they are sick but to give birth etc. **Figure 6** shows the interaction between IO\_CODE and Age. We can see the GapDays decreases with age increases for outpatients, while GapDays stays the same or even slights increases for inpatients age greater than 40 years old. This indicates investing more resources on inpatients age from 40 to 90 could be effective reducing the readmission rate (increasing the GD), because their GD doesn’t behave in the pattern it supposed to be.

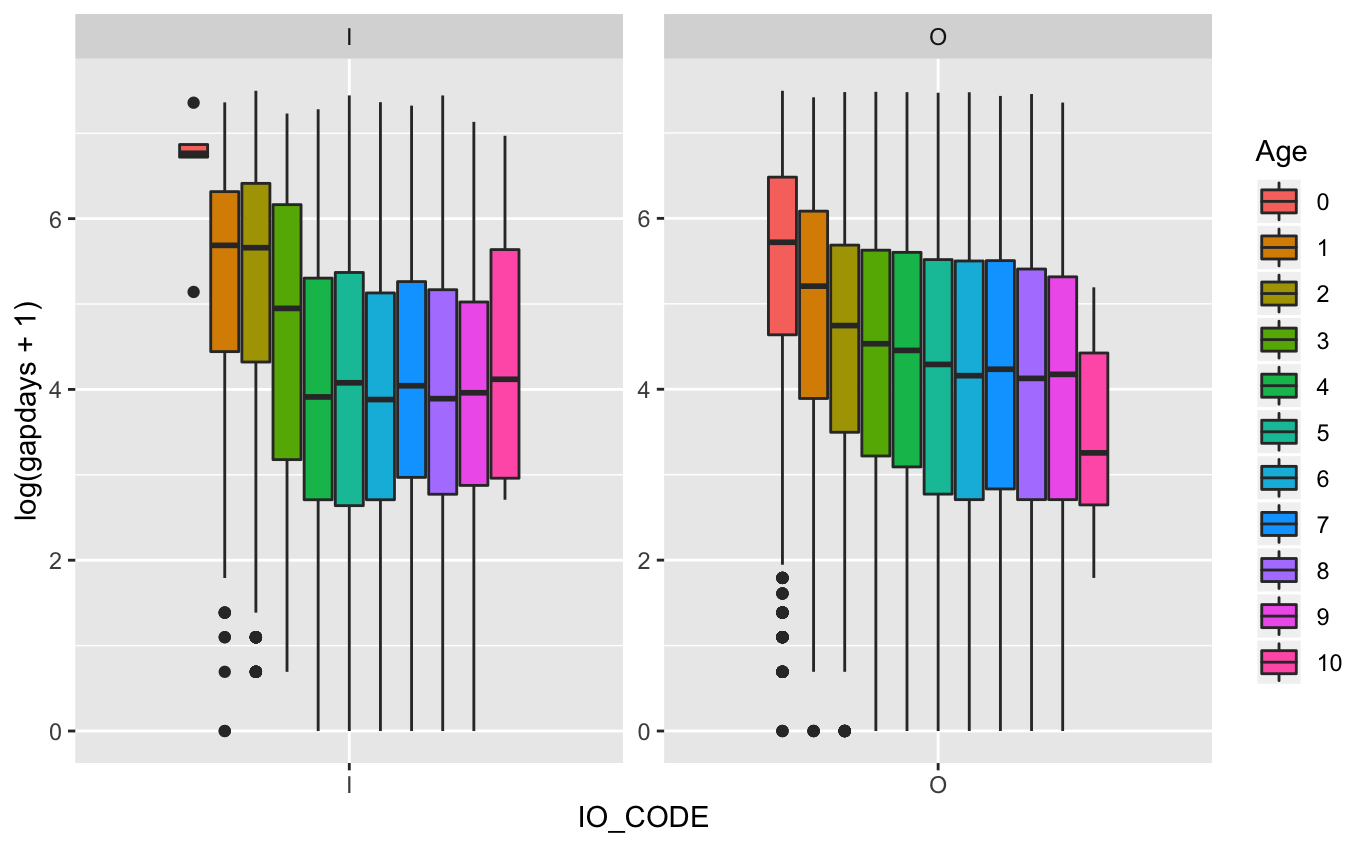
[State here which interaction you may use for later work, or both. “Both interactions will be explored in the later model to see how it can help improving the model.”]

[Can consider add the deleted time columns back, because there could be useful trend effect]

[Can add a formula here to explain the interaction]



**Figure 5**



**Figure 6 [The Age legend should be modified]**

**[Task 5]GLM**

We split data into 75% training data and 25% testing data. The average of training and testing data 207.233 and 207.438. This shows they are produce to have very similar distribution by the building stratification function in R.

Before we run GLM, we first run the linear regression model to get the baseline model. Its AIC [ref] is 692151 and RMSE is 275.102.

We evaluate two combinations of distribution and link function of GLM: Gamma and Inverse Gaussian distributions, both with log link function. Because the target variable is right skewed, we choose the Gamma and Inverse Gaussian distributions, which are both right skewed. The reason we choose log link is that it ensures the non-negative output of the prediction, which is the domain of the GapDays. The GLM models also involves the interactions selected above: the interaction between IO\_CODE and PATIENT.SEX.CODE, IO\_CODE and Age.

The AIC for GLM Gamma distribution with log link is 607627 and RMSE is 271.102. The GLM Inverse Gaussian distribution doesn’t converge when running it in R.

The Gaussian distribution with log link can also produce non-negative output. Running it we get AIC=692012, RMSE=274.642.

Given two models have same number of predictors, the model with lower AIC indicates the larger value of loglikelihood and thus better fit the training data. GLM with Gamma distribution and log link has lower AIC than OLS model and Gaussian GLM, we select it as the model to be used for further analysis.

[Task 6 - Select features using AIC or BIC]

It is likely to have overfitting when too many predictors are involved in the model. AIC and BIC can effectively reduce the overfitting. AIC is defined as , where is the loglikehood of the model on the training data, is the number of parameters (or features) in model. Using AIC as the metric to evaluate overfitting, the model with smaller AIC is better. acts as the penalty term, to prevent the model adding too many parameters. BIC is defined as , where is the number of rows in training data. From the formulas of AIC and BIC, we know BIC is stricter than AIC, thus BIC is more likely to build a model with less features. Since our goal is to select the most important features to be contained in the final model, we prefer BIC than AIC as the criteria.

In forward selection, the algorithm starts from an empty model, then adding feature one by one until the BIC no longer decreases. It is opposite for backward selection, where it starts from the full model with all the features included, then deleting feature one after another until the BIC doesn’t decrease. It is more likely the forward selection produces less features than backward selection. We use forward selection to do feature selection, since we’d like a model with less parameters.

Running the forward selection with BIC, the following features are selected:

* DOCTOR.NUMBER
* ServCode\_DRG\_ICD
* Age
* Surgeon
* PATIENT.MARITAL.STATUS
* Nur.Stat
* DISCHARGE.STATUS
* income\_level
* Patient.Days
* PATIENT.RACE.CODE
* IO\_CODE

When running the GLM model with the above selected features, we get the following output about the p-value of each predictor. Notice that majority features have small p-values except several features (or levels). This indicates the most features selected are significant. There are 6 levels that are not significant. For example, DOCTOR.NUMBER=5 has a p-value=0.13795. To deal with this issue, we combine these levels to their base levels.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.99381 0.16330 36.70 < 2e-16 \*\*\*

DOCTOR.NUMBER2 -0.60542 0.06983 -8.67 < 2e-16 \*\*\*

DOCTOR.NUMBER3 -0.10120 0.04012 -2.52 0.01166 \*

DOCTOR.NUMBER4 0.31089 0.03528 8.81 < 2e-16 \*\*\*

DOCTOR.NUMBER5 0.52643 0.35486 1.48 0.13795

ServCode\_DRG\_ICD -0.20195 0.00603 -33.48 < 2e-16 \*\*\*

Age10-20 -0.26285 0.05948 -4.42 9.9e-06 \*\*\*

Age100+ -1.86377 0.39675 -4.70 2.6e-06 \*\*\*

Age20-30 -0.61615 0.05559 -11.08 < 2e-16 \*\*\*

Age30-40 -0.66055 0.05605 -11.78 < 2e-16 \*\*\*

Age40-50 -0.63653 0.05632 -11.30 < 2e-16 \*\*\*

Age50-60 -0.63357 0.05637 -11.24 < 2e-16 \*\*\*

Age60-70 -0.63019 0.05726 -11.01 < 2e-16 \*\*\*

Age70-80 -0.60506 0.05855 -10.33 < 2e-16 \*\*\*

Age80-90 -0.62985 0.06173 -10.20 < 2e-16 \*\*\*

Age90-100 -0.60780 0.08093 -7.51 6.0e-14 \*\*\*

Surgeon2 -0.27524 0.02279 -12.08 < 2e-16 \*\*\*

Surgeon3 -0.18455 0.06706 -2.75 0.00592 \*\*

Surgeon4 -0.36174 0.04333 -8.35 < 2e-16 \*\*\*

Surgeon5 0.53341 0.34909 1.53 0.12652

PATIENT.MARITAL.STATUSM 0.16705 0.01716 9.73 < 2e-16 \*\*\*

PATIENT.MARITAL.STATUSPS 0.05911 0.02037 2.90 0.00371 \*\*

Nur.StatMS\_CCU\_ER\_PC 0.02249 0.03593 0.63 0.53137

Nur.StatWS 0.22770 0.06175 3.69 0.00023 \*\*\*

DISCHARGE.STATUSR 0.39629 0.04480 8.85 < 2e-16 \*\*\*

DISCHARGE.STATUSY 0.43827 0.05653 7.75 9.2e-15 \*\*\*

income\_levellow -0.11351 0.02611 -4.35 1.4e-05 \*\*\*

income\_levelmeduim -0.09598 0.01506 -6.37 1.8e-10 \*\*\*

Patient.Days20-30 -0.28387 0.04367 -6.50 8.1e-11 \*\*\*

Patient.Days30+ -0.48626 0.32127 -1.51 0.13015

PATIENT.RACE.CODEBO -0.18993 0.14455 -1.31 0.18887

PATIENT.RACE.CODEDH -0.99651 0.20834 -4.78 1.7e-06 \*\*\*

PATIENT.RACE.CODEWX -0.11298 0.14369 -0.79 0.43171

IO\_CODEO -0.18799 0.04251 -4.42 9.8e-06 \*\*\*

---

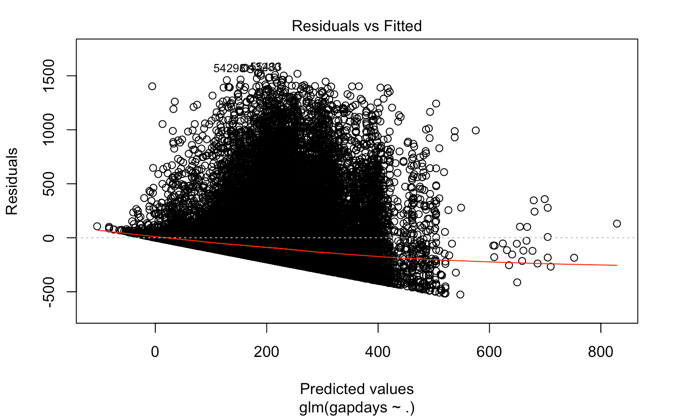
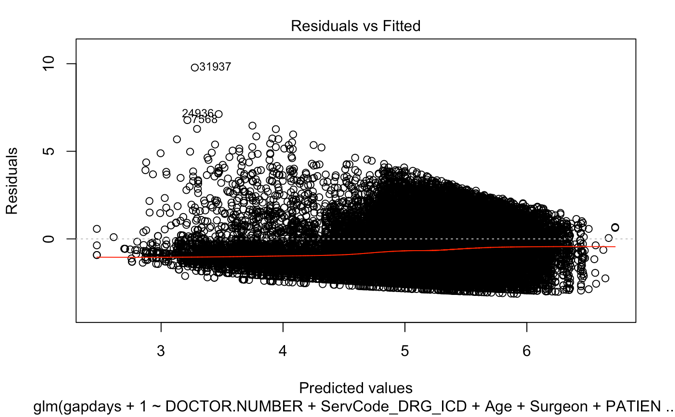
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Table 4**

[Task 7 Validate the model]

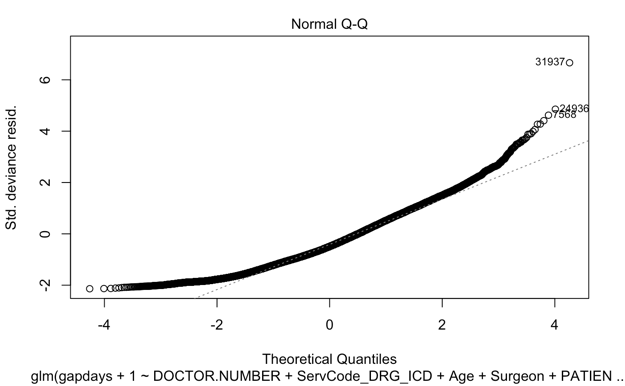
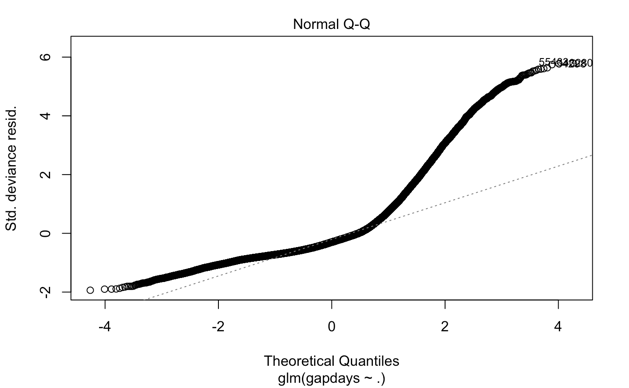
Running the GLM model constructed above, and the ordinary least squares regression (OLS) model in the beginning of Task 5, we get AIC is 607727 for the GLM model and 692151 for OLS model. GLM has significant lower AIC despite having fewer features, suggesting the GLM is a better model than OLS.

We obtained the diagnose plots of two models. From the plot of Residual vs Fitted in Figure 7, the points for GLM are distributed more symmetrically against its mean and centered near zero than the points of OLS in the right model. Therefore the assumption of constant variance and zero mean of residual is more valid for GLM than OLS.



**Figure 7**: Residual vs Fitted. The left figure is for GLM with features selected from above session, right figure for OLS with all predictors.

The Q-Q plot of GLM in Figure 8 shows the normal assumption for residuals holds for most values, except for values near extreme. It suggest a right-skewed distribution like gamma or lognormal would do better for residuals distribution. The Q-Q plot in the right side of Figure 8 shows the normal assumption of residual is not proper for OLS model.

**Figure 8:**

[Task 8]

We run the GLM model built in Session 6 on all data, the coefficients of the regression model are listed in Table 5.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.84694 0.06366 91.84 < 2e-16 \*\*\*

DOCTOR.NUMBER2 -0.58933 0.06102 -9.66 < 2e-16 \*\*\*

DOCTOR.NUMBER3 -0.11050 0.03487 -3.17 0.00153 \*\*

DOCTOR.NUMBER4 0.28857 0.03085 9.36 < 2e-16 \*\*\*

ServCode\_DRG\_ICD -0.20100 0.00528 -38.10 < 2e-16 \*\*\*

Age10-20 -0.30656 0.05169 -5.93 3.0e-09 \*\*\*

Age100+ -1.16735 0.35101 -3.33 0.00088 \*\*\*

Age20-30 -0.63787 0.04820 -13.23 < 2e-16 \*\*\*

Age30-40 -0.69793 0.04861 -14.36 < 2e-16 \*\*\*

Age40-50 -0.67707 0.04885 -13.86 < 2e-16 \*\*\*

Age50-60 -0.67009 0.04887 -13.71 < 2e-16 \*\*\*

Age60-70 -0.66848 0.04966 -13.46 < 2e-16 \*\*\*

Age70-80 -0.63585 0.05076 -12.53 < 2e-16 \*\*\*

Age80-90 -0.63634 0.05360 -11.87 < 2e-16 \*\*\*

Age90-100 -0.66369 0.07005 -9.47 < 2e-16 \*\*\*

Surgeon2 -0.28085 0.01977 -14.21 < 2e-16 \*\*\*

Surgeon3 -0.22311 0.05860 -3.81 0.00014 \*\*\*

Surgeon4 -0.36310 0.03772 -9.63 < 2e-16 \*\*\*

PATIENT.MARITAL.STATUSM 0.18697 0.01491 12.54 < 2e-16 \*\*\*

PATIENT.MARITAL.STATUSPS 0.06817 0.01772 3.85 0.00012 \*\*\*

Nur.StatWS 0.20997 0.04929 4.26 2.1e-05 \*\*\*

DISCHARGE.STATUSR 0.41057 0.03870 10.61 < 2e-16 \*\*\*

DISCHARGE.STATUSY 0.45851 0.04926 9.31 < 2e-16 \*\*\*

income\_levellow -0.08182 0.02262 -3.62 0.00030 \*\*\*

income\_levelmeduim -0.09786 0.01307 -7.49 7.2e-14 \*\*\*

Patient.Days20-30 -0.22549 0.03781 -5.96 2.5e-09 \*\*\*

PATIENT.RACE.CODEDH -0.57857 0.13194 -4.39 1.2e-05 \*\*\*

PATIENT.RACE.CODEWX 0.07868 0.01669 4.71 2.4e-06 \*\*\*

IO\_CODEO -0.21764 0.02759 -7.89 3.1e-15 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Table 5:**

Since the log link was took on the target variable, to explain the effect of the coefficients, we should take exponent of the coefficients and minus 1. The following table is the interpretation of some coefficients:

|  |  |  |
| --- | --- | --- |
| Feature | Coefficients | Interpretation |
| DOCTOR.NUMBER=2 | -0.45 | 45% decrease in GapDays compared to the group DOCTOR.NUMBER=1. This means the group 2 of doctors are less effective in terms of improving the readmission days (GapDays) than group 1. |
| DOCTOR.NUMBER=3 | -0.10 | 10% decrease in GapDays compared to the group DOCTOR.NUMBER=1. |
| DOCTOR.NUMBER=4 | 0.33 | 33% decrease in GapDays compared to the group DOCTOR.NUMBER=1. Therefore group 4 of doctors are more effective improving the GapDays tha group 1. |
| ServCode\_DRG\_ICD | -0.18 | 18% decrease in GapDays for every 1.0 increasing in the feature ServCode\_DRG\_ICD, which is the new artificial feature made using the PCA from the predictors PATIENT.DRG, HOSPITAL.SERVICE.CODE, ICD.PROCEDURE.CODE, ICD9\_DIAGNOST\_CODE |
| Age=10-20 | -0.26 | 26% decrease compared to Age 0-10.This make sense because the younger people are less likely to be readmitted. |
| Age=20-30 | -0.47 | 47% decrease compared to Age 0-10. |
| … |  |  |
| Age=100+ | -0.69 | 69% decrease compared to Age 0-10. |
| Surgeon=2 | -0.24 | 24% decrease of GapDays compared with Surgeon=1. |
| … |  |  |
| PATIENT.MARITAL.STATUS=M | 0.21 | 21% increase of GapDays for patients in marriage compared with base level MARITAL.STATUS=DUWXNA, which is the group for divorced, widowed or unknown status. People in marriage usually can be taken care better thus have better health condition. |
| PATIENT.MARITAL.STATUS=PS | 0.07 | 7% increase of GapDays for patients with domestic partner (P) or single (S) compared with base level MARITAL.STATUS=DUWXNA. |
| income\_level=low | -0.08 | 8% decrease of GapDays for low income patients compared to high income patients. This make sense because the low income patients may not afford enough healthcare to maintain good health condition. |
| income\_level=medium | -0.09 | 9% decrease of GapDays compared to high income patients. The medium income patients have even worse readmission days than low income patients maybe because they have longer working hours, higher mental pressure. |
| Patient.Days=20-30 | -0.20 | 20% decrease compared to its base level, the visits whose Patient.Days smaller than 20 or greater than 30. This suggests the visits whose days of hospital stay is between 20-30 days are most likely to be readmitted with short readmission days. |
| PATIENT.RACE.CODE=WX | 0.08 | 8% increase of GapDays for PATIENT.RACE.CODE is W or X compared to its base level PATIENT.RACE.CODE=AIMNT. |

[Task 9 - Investigate ridge and LASSO regressions]

The other feature selection methods are Ridge regression and LASSO regression. Ridge regression uses L2 penalty. Its penalty term is the sum of the sqaures of all the features selected. LASSO regression uses L1 penalty term, which is the sum of the absolute values of all the selected features. Both penalty terms act as a filter of features. The ridge regression reduces the coeffients of less important features approch to but not equal to 0. While LASSO makes them equal to 0 to elimiate the less important features.

The glmnet function in R requires the binarization of each level of the categorical variables, while this is not required in the stepwise method. In the stepwise method, each level is treated as an object to retain or combined to its base level based on their p-values. In LOSSO, each level is retained or removed based on whether their regression coeffients are 0 or not.

Ridge regression is not suitable for our task, which is to identify the key predictors. Becasue Rigde regression does not significantlly reduce the amount of predictors, just reducing the values of the coefficents.

An advantage of LASSO over the stepwise method is the crossed validation (CV) was used in LASSO to reduce the overfitting and bias, which is not used in the stepwise method. The stepwise methods tend to produce a more complicated model than LASSO.

An RMSE of 274.9847 is obtained when running LASSO. It retained more predictors than the stepwise methods. The RMSE for the GLM model developed by Task 8 is 275.6046.

The LASSO produced a more complicated model but doesn't signicantly reduced the RSME, thus we still recommend the setpwise methods in this study.

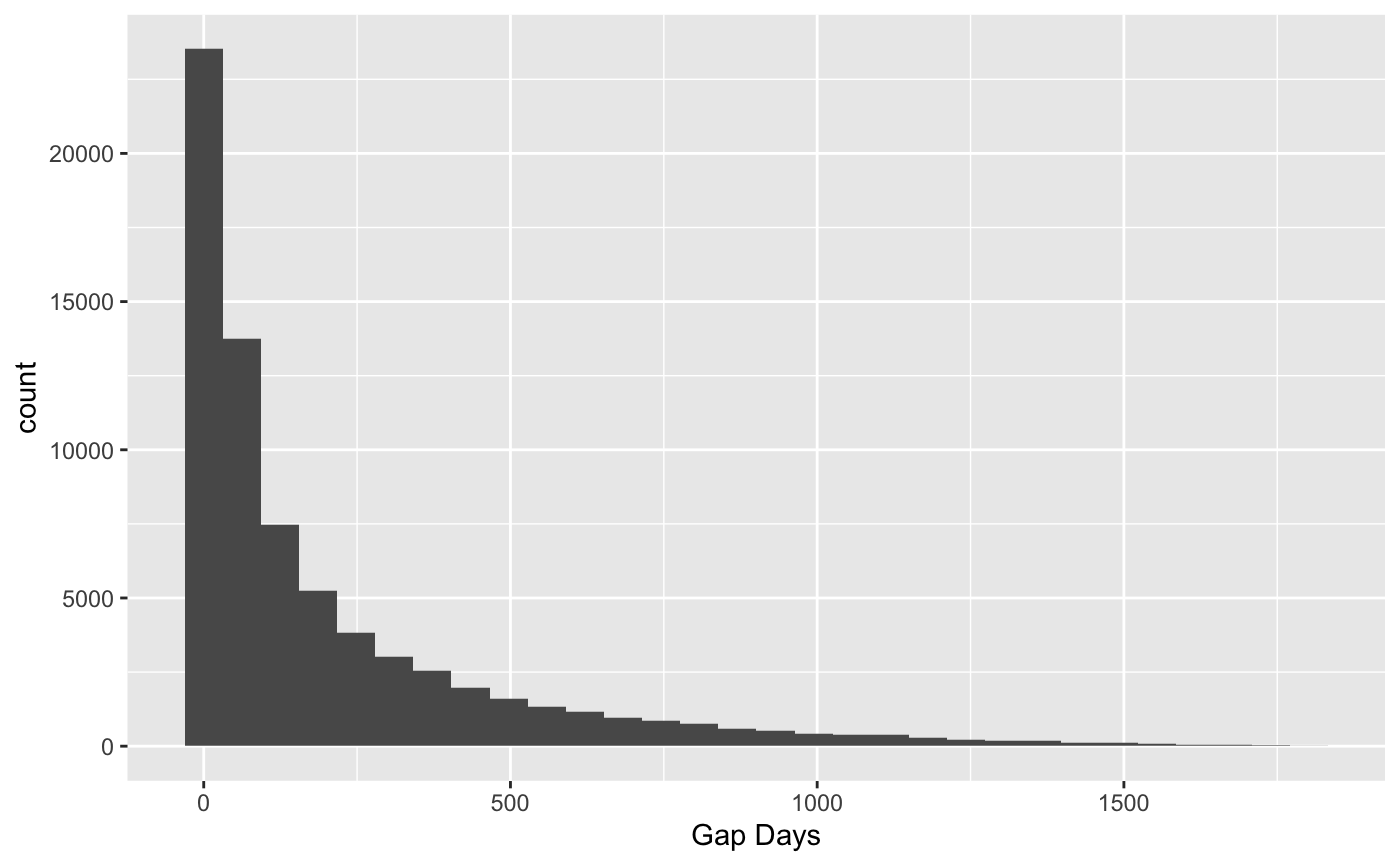
[Task 11]

To: Hospital Client

From: Actuarial Consultant

Re: Explore the factors related to the hospital readmission days.

This report provides an investigation into the factors that contribute more or less to the GapDay, which is defined as the days until next readmission after the most recent discharge for a patient visit. We have used a dataset from a hospital in North Carolina 2011-2016. The data included the variables you believe relate to the GapDays, such as patient gender, DRG, etc. As requested, we have applied analytics techniques to determine which of those variables are related to the GapDays, and the magnitude of the effect of those that do.

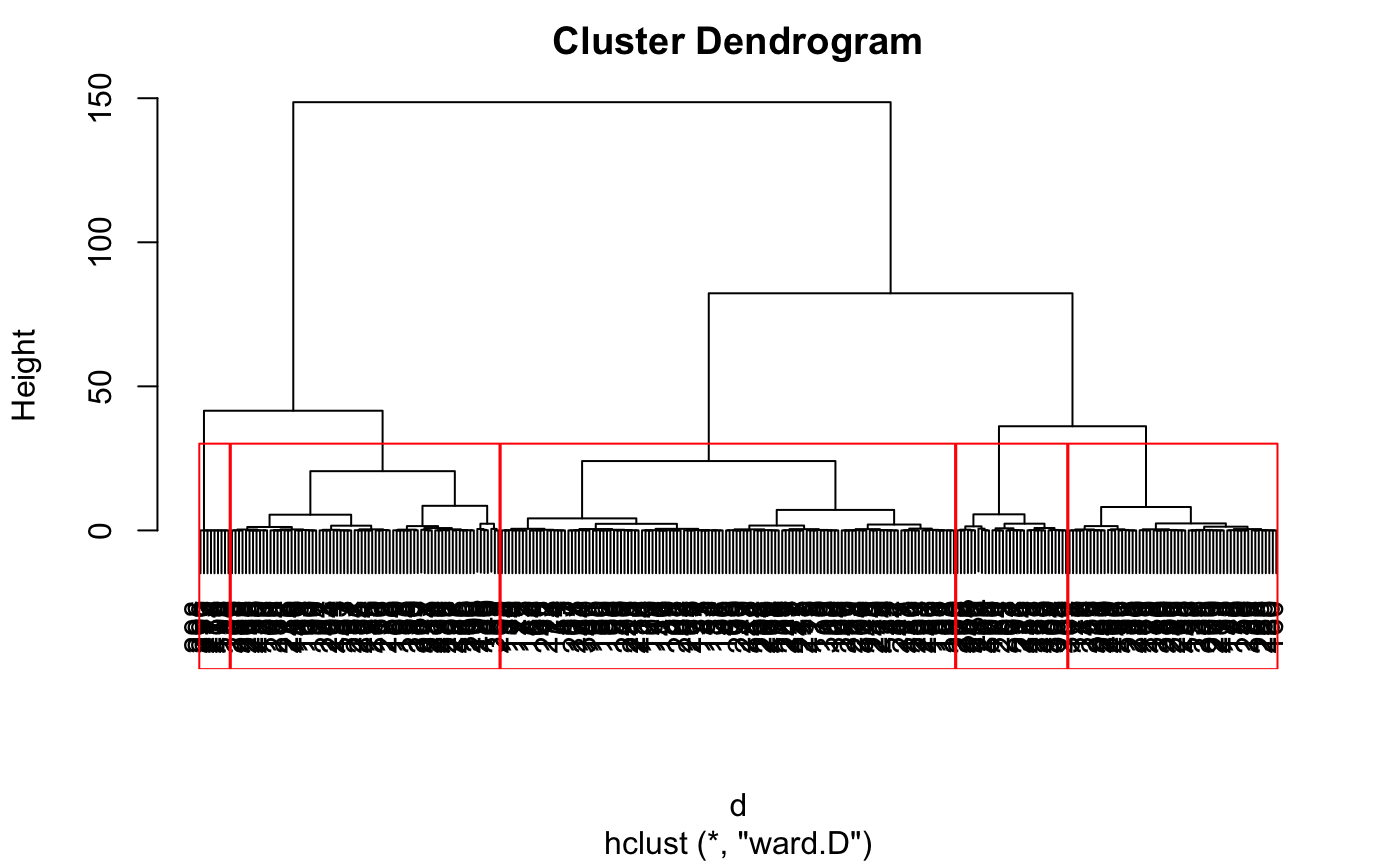
The data you supplied contains 71724 records of patient visits in a network of all the hospitals in North Carolina from 2011 through Early 2016. The amount of missing entries is not relatively small compared with the total amount of data. Therefore we think the data is reliable for the intended purpose. Because the data is from NC state only, the findings in this report may not be generalized well to the readmission GapDays nationwide. The distribution of GapDays is presented below:

It is a right-skewed distribution, with mostly low values and few high values. This characteristic will be considered in the later modeling process.

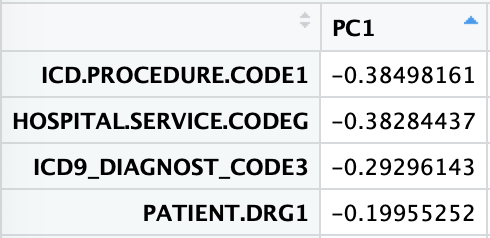
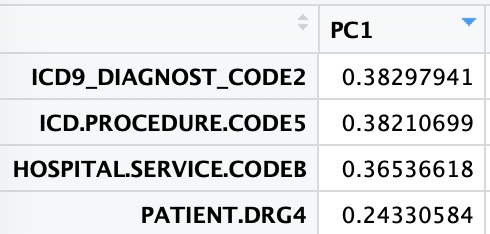
Some of the variables we think does not have predictive power, so we deleted them. These variables include PATIENT/ENCOUNTER.NUMBER, PATIENT.HISTORY.NUMBER etc. In order to use the finding of this report about the relationship between each variable and GapDays to help reduce the readmission GapDays, we need to know which variables are within the control of the hospitals. Variables like Age, PATIENT.DRG, PATIENT.SEX.CODE, PATIENT.RACE.CODE, ICD-9.DIAGNOST.CODE are not within the control of hospitals. The hospitals can only accept the patients as they are. The other variables are within the control of hospitals that can be allocated to help reduce the readmissions. These variables include Nur.Stat, DOCTOR.NUMBER, Surgeon, I/O.CODE (inpatient or outpatient), HOSPITAL.SERVICE.CODE, ADMIT.SOURCE, ICD.PROCEDURE.CODE.

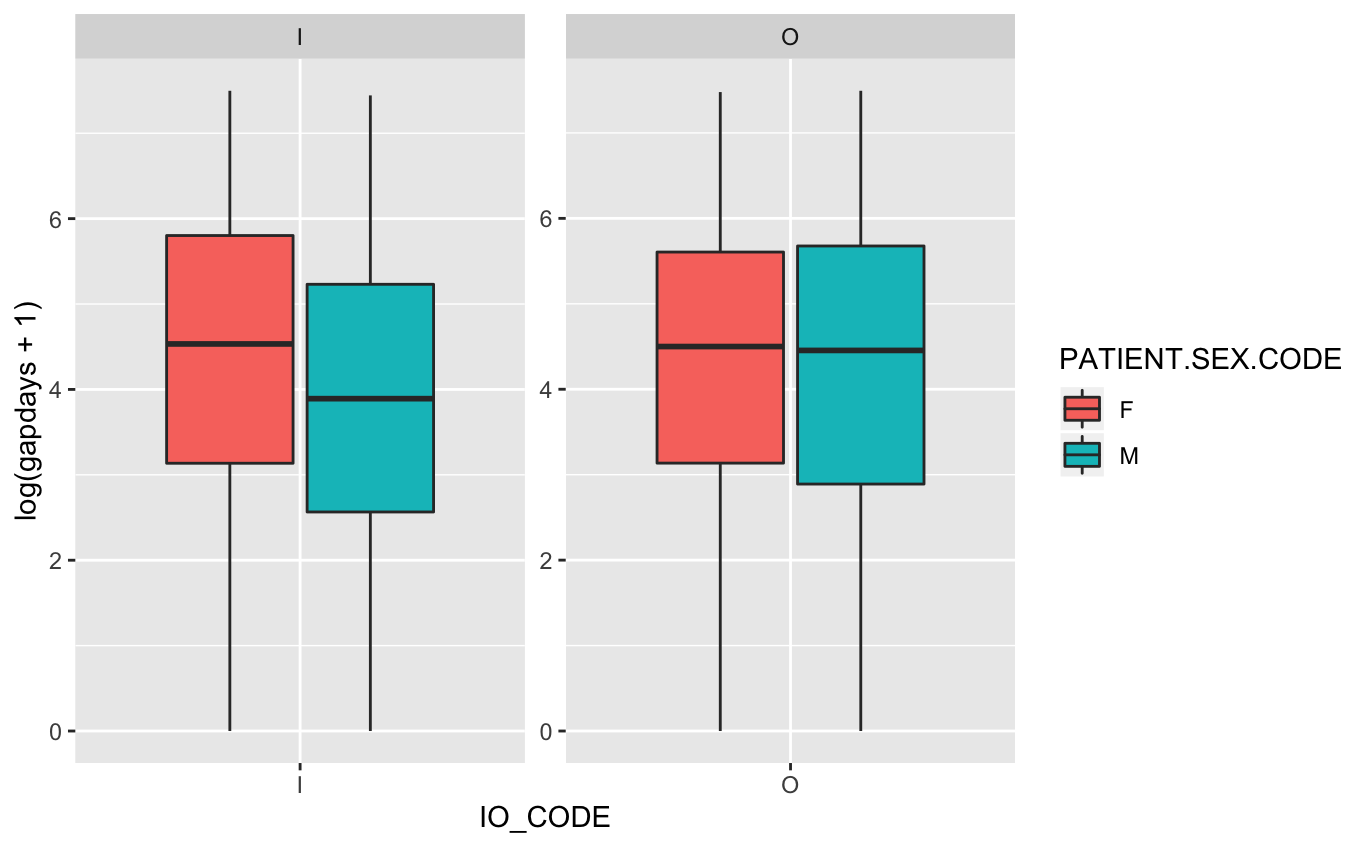
To tell which variables could be good candidates for the features in the final model, we did an exploration analysis of the target value in each level of each categorical predictors using both boxplot and summary statistics about mean and median. We got some interesting findings. For instance, the GapDays is low for income less than 10k, or greater than 90k, high for income in 50-60k. There is an obvious trend the GapDays decrease with the Age increases. The following variables could be predictive based on the magnitude of the variation of the target variable within the levels of predictors: Patient.Days, Hospital.Service.Code, Patient.Marital.Status, Age, etc.

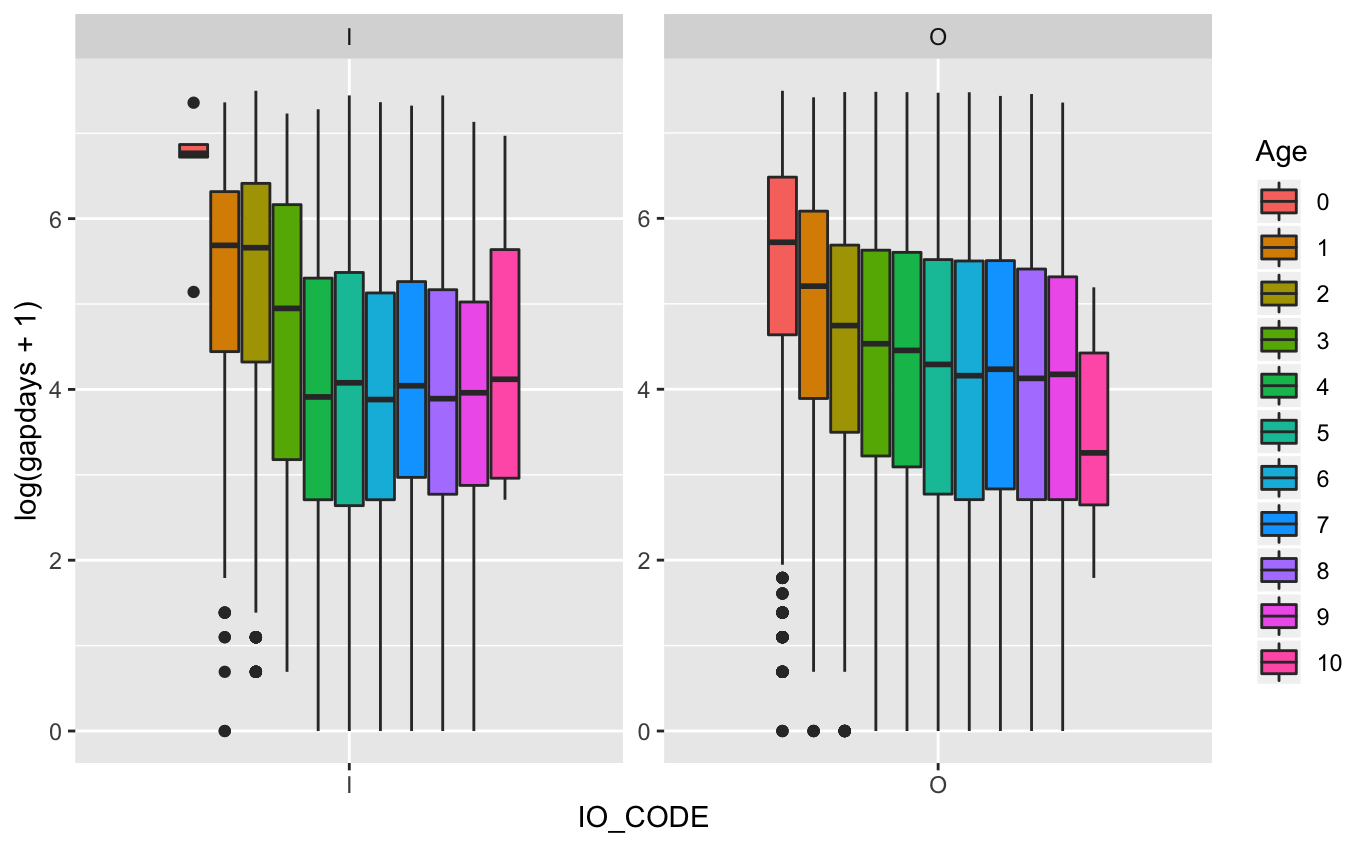
We combined some of the levels within the predictors if they have a similar mean and median of the target variable. For instance, the mean values of GapDays for Patient.Marital.Status is P (Domestic partner), or S (Single) are relatively close, we group these two levels of Patient.Marital.Status into one level called PS. It also makes sense to combine them because they both represent an unmarried status. Some of the factor predictors have too many levels, an example if DOCTOR.NUMBER. Each of the code represents a doctor. There are hundreds of different doctors in the data. It’s not possible to combine them manually. So we applied the clustering methods to group them. Based on the mean values of the target within each level, a cluster tree is built using Ward’s distance method. We group them into 5 clusters. The levels in the same cluster are combined. Reducing the levels of the predictors can help us avoid overfitting thus improving the performance of the model on unknown testing data.



Before we start the formal analysis, another alternation is made to the variables to help select out the most important components of the variables. We applied the Principle Component Analysis (PCA) to the data and found the following levels of predictors have the largest absolute value in the direction of the first principle component, which is the most important component. We then made a new feature which is the linear combination of the levels listed in the following table. The 4 predictors where these levels came from are then deleted. To this point, there are 13 variables (or features) remained in the data.

The interactions between the feature are then investigated. When we say two predictors have interaction with each other, that means the change of value in one predictor alters how the other predictor affects the target. We found the interactions exist between IO\_CODE vs PATIENT.SEX CODE, IO\_CODE vs Age. IO\_CODE. As it shows in the following figures, the target values are significantly higher in female than in male for IO\_CODE=I, while for IO\_CODE=O, the target values are almost the same in both genders. This means the value of IO\_CODE alters how the PATIENT.SEX CODE impacts the target variable. The reason could be the inpatient service has a better outcome for females than males, while outpatient service has the same outcome on both genders. In the second figure, we can see the GapDays decreases with age increases for outpatients, while GapDays stays the same or even slights increases for inpatients age greater than 40 years old.



With the data ready for analysis, we choose the Generalized Linear Model.

With the above data preprocessing done, we use the generalized linear model (GLM) to do the analysis with all the features and interactions. We considered different types of GLM and selected the one that makes the most sense. Compare with the ordinary least squares (OLS) model, the generalized linear model improved the goodness-of-fit of the model, which is measured by the AIC.

Then we considered the selection of the most important features using or BIC, a similar but more strict measurement of the goodness-of-fit as AIC. When it comes to select the futures, there are forward selection and backward selection to implement this process. It is more likely the forward selection produces fewer features than backward selection. Running the forward selection with BIC, the following features are selected:

• DOCTOR.NUMBER

• ServCode\_DRG\_ICD

• Age

• Surgeon

• PATIENT.MARITAL.STATUS

• Nur.Stat

• DISCHARGE.STATUS

• income\_level

• Patient.Days

• PATIENT.RACE.CODE

• IO\_CODE

We combined some of the levels of the features whose coefficients are not very significant to their base levels.

Various methods are used to validate the model constructed using the features selected above compared with the OLS. It turns out the underlying assumptions of linear regression are more valid for the GLM we constructed than the OLS model with all features. For instance, it shows the assumption of constant variance and zero mean of residual works better for GLM than OLS. The Q-Q plot shows the normal assumption for residuals holds better for GLM than OLS.

We then run the GLM model with the above features selected and levels combined. A list of regression coefficients is generated. Here is an interpretation of some coefficients:

|  |  |  |
| --- | --- | --- |
| Feature | Coefficients | Interpretation |
| DOCTOR.NUMBER=2 | -0.45 | 45% decrease in GapDays compared to the group DOCTOR.NUMBER=1. This means the group 2 of doctors are less effective in terms of improving the readmission days (GapDays) than group 1. |
| DOCTOR.NUMBER=3 | -0.10 | 10% decrease in GapDays compared to the group DOCTOR.NUMBER=1. |
| DOCTOR.NUMBER=4 | 0.33 | 33% decrease in GapDays compared to the group DOCTOR.NUMBER=1. Therefore group 4 of doctors are more effective improving the GapDays tha group 1. |
| ServCode\_DRG\_ICD | -0.18 | 18% decrease in GapDays for every 1.0 increasing in the feature ServCode\_DRG\_ICD, which is the new artificial feature made using the PCA from the predictors PATIENT.DRG, HOSPITAL.SERVICE.CODE, ICD.PROCEDURE.CODE, ICD9\_DIAGNOST\_CODE |
| Age=10-20 | -0.26 | 26% decrease compared to Age 0-10.This make sense because the younger people are less likely to be readmitted. |
| Age=20-30 | -0.47 | 47% decrease compared to Age 0-10. |
| … |  |  |
| Age=100+ | -0.69 | 69% decrease compared to Age 0-10. |
| Surgeon=2 | -0.24 | 24% decrease of GapDays compared with Surgeon=1. |
| … |  |  |
| PATIENT.MARITAL.STATUS=M | 0.21 | 21% increase of GapDays for patients in marriage compared with base level MARITAL.STATUS=DUWXNA, which is the group for divorced, widowed or unknown status. People in marriage usually can be taken care better thus have better health condition. |
| PATIENT.MARITAL.STATUS=PS | 0.07 | 7% increase of GapDays for patients with domestic partner (P) or single (S) compared with base level MARITAL.STATUS=DUWXNA. |
| income\_level=low | -0.08 | 8% decrease of GapDays for low income patients compared to high income patients. This make sense because the low income patients may not afford enough healthcare to maintain good health condition. |
| income\_level=medium | -0.09 | 9% decrease of GapDays compared to high income patients. The medium income patients have even worse readmission days than low income patients maybe because they have longer working hours, higher mental pressure. |
| Patient.Days=20-30 | -0.20 | 20% decrease compared to its base level, the visits whose Patient.Days smaller than 20 or greater than 30. This suggests the visits whose days of hospital stay is between 20-30 days are most likely to be readmitted with short readmission days. |
| PATIENT.RACE.CODE=WX | 0.08 | 8% increase of GapDays for PATIENT.RACE.CODE is W or X compared to its base level PATIENT.RACE.CODE=AIMNT. |

This study investigated the relationships of the predictors to the readmission GapDays. We found Age is the most important factor impacting the Gapays. Other predictors like DOCTOR.NUMBER, Surgeon, PATIENT.MARITAL.STATUS and Patient.Days also has a significant effect on GapDays. Our hospital client can use this study to predict which patient visit can lead to short readmission. So that the hospital can better allocate the resources on these patients to reduce and avoid the rate of short readmission. Especially the 30-days readmission rate is a metric both the CMS (Centers for Medicare & Medicaid Services) and the hospitals pay lots of attention. Notice the interactions identified between IO\_CODE and PATIENT.SEX.CODE, IO\_CODE, and Age are not included in the final GLM model. If there is more data, it would be interesting to further investigate this and more interactions. We would also like to implement a loss-gain analysis about each wrong or correct prediction if the information of loss-gain associate to each type of prediction is available.

**References:**

1. MBA Medical, “Cost of Hospital Readmissions”, 6 Nov. 2015,  http://mbamedical.com/cost-hospital-readmissions/
2. Jencks, Stephen F., Mark V. Williams, and Eric A. Coleman. "Rehospitalizations among patients in the Medicare fee-for-service program." New England Journal of Medicine 360.14 (2009): 1418-1428.
3. Hines, Anika L., et al. "Conditions with the largest number of adult hospital readmissions by payer, 2011: statistical brief# 172." (2006).
4. Center for Health Information and Analysis. “Performance of the Massachusetts Health Care System Series: a focus on provider quality.” January 2015. Publication Number 15- 06-CHIA-01
5. Judy Mathias, “Study identifies high-risk groups for readmissions”, 11 July 2017,  https://www.ormanager.com/briefs/study-identifies-high-risk-groups-for-readmissions/
6. Associated Press, “Medicare will become insolvent in 2026, U.S. government says”, 05 Jun. 2018, http://www.latimes.com/nation/nationnow/la-na-pol-medicare-finances- 20180605-story.html]
7. Boccuti, Cristina, and Giselle Casillas. "Aiming for fewer hospital U-turns: the Medicare hospital readmission reduction program." Policy Brief (2015).
8. Krumholz HM, Normand S-LT, Keenan PS, et al. Hospital 30-day acutemyocardial infarction readmissionmeasure: methodology, June 9, 2008.
9. Krumholz H, Normand S-L, Keenan P, et al. Hospital 30-day heart failure readmission measure: methodology, April 23, 2008.
10. Krumholz HM, Normand S-LT, Keenan PS, et al. Hospital 30-day pneumonia readmission measure: methodology, June 9, 2008.
11. Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: a systematic review. Jama, 306(15), 1688-1698.
12. Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.
13. Health Catalyst report, “Accuracy of Readmission Risk Assessment Improved by Machine Learning”, 2018

* Williams, K. (2004). The transition to widowhood and the social regulation of health: Consequences for health and health risk behavior. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 59(6), S343-S349.