Prescription Timeliness Prediction Using Machine Learning

Author: Ian de la Cruz
Advisor: Jeffrey Duffany, Ph.D.

Abstract

Adherence is a very important aspect for each patient enrolled in a Prescription Drug Plan, a large gap between claim fills is an indicator that they might be disrupting their therapy. The prediction of when each patient will go and get a refill or set a medical appointment to receive a new prescription would highly reduce such gaps and increase adherence. The purpose of the program is to predict when these occurrences might take place, using demographic data, to better tailor adherence programs to a patient's schedule. The methodology involves implementing Machine Learning utilizing R Services within SQL Server 2017. Using the information available an accurate prediction was not established using only demographic information. Additional historical information on an individual patient basis is necessary to be able to establish a more robust prediction

Introduction

Currently for any Prescription Drug Plan you have a set of Patients who are not consuming their therapies as they should. This affects both the Patient health wise and the rating of the Prescription Drug Plan from a ranking standpoint. A possible solution for this would be to establish a method to identify patient behavior to determine when he or she usually obtains their medication from the pharmacy so that an action can be taken when a patient is identified as not continuing their therapy as required.

The purpose of this project is to predict when these patients require a reminder or an intervention from the Plans utilizing Machine Learning by implementing R in SQL Server and using real patient paid claim data from Prescription Drug Plans. The claim being the pharmacy transaction that contains each unique patient's medication based on processed date, date of service, patient, medication, prescription number, fill number and refill number. A series of processes can be established after these patients have been identified to maintain or increase their medication adherence, these can range from dedicating Customer Call Centers to reach out to patients and determine if they require Assistance Programs or to even sending specialized nurses to educate these Patients in the importance of their therapy adherence.

Methodology

This project was developed utilizing SQL Server 2017 Machine Learning Services (In-Database), configured as established by Microsoft materials [1] and utilizing data from different Lines of Business Types. Data exploration and manipulation was executed to determine which fields were necessary to develop a prediction.

Selected fields for prediction were based on hands on PBM experience. The selected data included information from the following information groups: Patient Information, Drug Plan Information, Prescription Drug Information and Claim Information. Within each group the preliminary variables selected were the following. Patient ID, Date of Birth, Gender, Prescription Drug Plan, Line of Business and Line of Business Type from the Patient Information.

Generic Product Identifier (GPI), Drug Name, Drug Dispensable Name, GPI Group, GPI Class, Medical Indication, OTC Indicator, Maintenance Drug Indicator, Brand/Generic Indicator, Dose Form, Route from the Drug Information. Claim ID, Fill Number, Service Date, Quantity Dispensed, Total Amount of Refills, Days Supply, Days Supply Category, Patient Paid Amount were selected from the Claim Information. Additionally, a series of calculations were added based on the same data to obtain a Next Date variable (1).

Next Day Claim = Service Date + Days Supply (1)

The Actual Next Claim Date was found based on the actual next processed Claim Date of Service for the drug and that patient. The difference between these two variables deemed Days Between (2) was used as the dependent variable that is to be predicted.

Days Between = Actual Next Claim Date – Next Claim Date (2)

Data Selection Method

It was decided that the data to be used would involve information from Service Date between the years 2016 and 2018. The data to train the model would range from 2016-01-01 to 2017-03-31, while the data to test the model would range from 2017-04-01 to 2017-09-30. After obtaining the initial data the first step was to determine which values of the Days Between variable would be considered outliers and mark them as such. A program using R was developed to identify outliers using a Probability Density Function. Figure 1 demonstrates the distribution obtained of the Days Between. As it's presented, most mayor values lies in the zero axis, which displays a favorable distribution of Days Between.

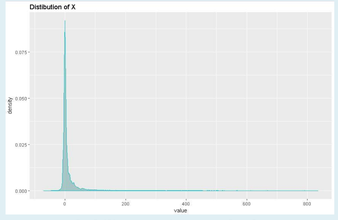

Figure 1. Probability Density Graph of Days Between

A preprocessing was performed for the Days Between fields, from which a probability distribution was calculated using the Probability Density Function of a multivariate normal (3) [3].

$$f(x) = (2\pi)^{-k/2}|\Sigma|^{-1/2}e^{1/2(x-\mu)'\Sigma^{-1}(x-\mu)} \quad (3)$$

The graphical output of this equation was generated using R standalone and a random data sample of a million records, this output is displayed in Figure 2. When comparing this output to the output of using all historical data as displayed in Figure 3 both graphs it can be observed that both generated graphs were very similar. This demonstrates how uniform the data is regarding Days Between and that the outliers reside in similar low probabilities.

Methodology

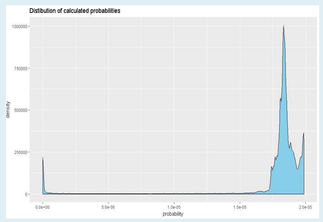

Figure 2 . Distribution of Days Between (Sample Data Set: 100k records)

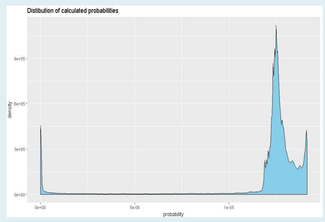

Figure 3. Distribution of Days Between (Full Data Set: 30M records)

Each probability of Days Between was calculated and it was decided to eliminate 15% of the events that had the lowest probability based on the outputs and based on previous experience with the underlying data. Initially 10% had been selected but after viewing this distribution of Days Between it was clear that a 5% of events were not as common as experience suggests. To test this the same process was executed for the full dataset and each claim was flagged as an outlier or not in the original data.

Finding Correlations between Independent Variables & Results

Originally the following fields were selected as the ones suspected of having more correlation with the Days Between variable: Service Date, Days Supply, GPI Group, GPI Class, Medical Indication, OTC Indicator, Maintenance Drug Indicator, Age, Gender, Line of Business Type, Patient Paid Amount. A linear correlation was calculated using the R correlation function between each individual field against Days Between using a random sample of one hundred thousand records that did not include outliers. When the correlation was completed it was observed that none of them has a significant correlation with the Days Between variable. The highest correlation found was between the Maintenance Drug Indicator with a value of -0.119165144 which was deemed a non-significant correlation as displayed in Table 1.

Table 1. Days Between correlation with Independent Variables

Independent Variables	Days Between Correlation
Service Date	0.036651238
Days Supply	-0.108608754
GPI Group	-0.009889088
GPI Class	0.023802496
Medical Indication	-0.010771842
OTC Indicator	0.02046733
Maintenance Drug Indicator	-0.119165144
Age	-0.035700316
Gender	-0.020483037
LOB Type	0.003862907
Patient Paid	0.001324059

Prediction Models

Before running any kind of prediction, we must first decide what kind of algorithm must be used. Since the Days Between variable is continuous and we have data that may have identifiable patterns, the methodology selected was Regression of the Supervised Learning Algorithms [4]. Within it the Linear was chosen to attempt to establish a linear relationship between the independent variables and the dependent variable and Boosted Decision Tree algorithm was chosen to improve accuracy. Using the variables from Table 1 a random sample dataset of 100,000 records for the period between 2016-01-01 and 2017-03-31 was selected and a linear model was generated using the Linear Model (LM) function. Since the LM R function only supports 56 unique categorical values, all categorical values were converted to numerical values to be able to be used.

Table 2. Correlation with Ranked Variables

Variable	Correlation
Service Date Month	-0.000281359
Days Supply	-0.096335411
GPI4 Rank	0.233986702
Medical Indication Rank	0.243384629
Maintenance Drug Indicator	-0.117709123
Brand/Generic Indicator	-0.002938256
Route	-0.104177553
Age	-0.029167594
Gender	-0.00785906
Claim History	-0.157269337
Average Days Between History	0.21486376
Average Fill Days Between History	0.209407359

The data used to train the LM function was from 2016. The resulting model was used to predict Days Between on the test data and generate a Prediction (P) based on each record in the test dataset which consisted of the remaining data from the sample. To verify if P was correct the function Postresample was employed, this function returns from your prediction against the original value a R2 which determines how successful the prediction was compared to the original value, the closer to 1 the more successful the prediction would be and demonstrates if the model fits the data. LM resulted in an R2 of 0.02192908. Since this was not satisfactory another model was chosen, Gradient Boosted Machine (GBM), based on the same sample and its' R2 value was 0.06542075 which is more favorable than the LM value but nonetheless too close to zero.

Due to both the LM and the GBM low R2 values, the variables used were adjusted. The GPI Group and GPI Class were eliminated and replaced by a new variable named GPI 10 and Brand/Generic Indicator, Dose Form and Route variables were included. A correlation was performed based on these new variables which resulted in that none of the correlations were significant, against with a slight exception of the Maintenance Indicator. In order to add a historical component to each Patient claim, it was determined that 3 new variables would be created based on calculations obtained from the Patient utilization information, the drug history consumption of the Patient. These variables are Claim History, which is the quantity of claims for the Patient before the current claim, the Average Days between History, which was calculated as the name implies, and the Average Fill Days Between which is the difference in days between the Service Date and the Actual Next Claim Date. However, these new variables did not present meaningful correlations in our trained dataset.

To attempt to improve the correlation of GPI and Medical Indication it was decided to create an ID for the GPI using a rank based on the Average Days Between for each GPI4 and use this as the identification for GPI4 instead of using the GPI4 variable. As a result, the GPI10 was eliminated. A similar ranking was also executed for Medical Indication. The output correlation resulted in that the ranking for Medical Indication and GPI4 improved by a factor of 3.

As a final exercise a correlation was executed using a sampling of 3 months of data, from Quarter 1 (Q1) 2017, and the Service Date variable was changed and replaced by the Service Date Month.

Methodology

The variables for this final correlation were: Service Date Month, Days Supply, GPI4 Rank, Medical Indication Rank, Maintenance Drug Indicator, Brand/Generic Indicator, Route, Age, Gender, Claim History, Average Days Between History, Average Fill Days Between History which resulted in the correlation in Table 2.

Results & Discussion

Using the Q1 2017 data the LM and GBM models were generated with the final variables and tested with data for April 2017. LM generated an R2 value of 0.0974037537657968 while GBM generated a R2 value of 0.0712577142179457. Figure 4 demonstrates the prediction using the LM vs. the original Days Between values for the test dataset, while Figure 5 demonstrates the prediction using the GBM vs. the original Days Between values for the test dataset.

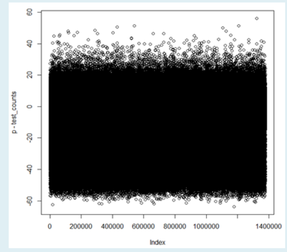

Figure 4. LM Prediction vs Original Data Set of Days Between

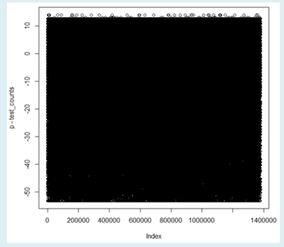

Figure 5. GBM Prediction vs Original Data Set of Days Between

As it is represented the output is not ideal, for an accurate prediction model the output would have displayed all values around the zero axis which would have demonstrated that the predicted value and the original value were similar.

Another output of the process involved the LM model generating an equation based on the coefficients that form part of the linear equation used to generate the graphical predictions displayed in Figure 4. The output of the Intercept (B0) and the rest of the equations coefficients are displayed in Table 3.

Table 3. Final Coefficient Results

Variable	Value
Intercept	4.3484174
Service Month	0.1226453
Days Supply	-0.0628202
GPI4 Rank	0.0037999
Medical Indication Rank	0.0306193
Maintenance Drug Indicator	-0.4498464
Brand/Generic Indicator	-0.2918423
Route	-0.0853789
Age	-0.0007751
Gender	-0.081935
Claim History	-0.0350313
Average Days Between History	-0.0163794
Average Fill Days Between History	0.0656391

This generates the following equation which is the final model equation (4):

P = 4.3484174 + 0.1226453(Service Month) – 0.0628202(Days Supply) + 0.0037999 (GPI4 Rank) + 0.0306193(Medical Indication Rank) - 0.4498464(Maintenance Drug Indicator) - 0.2918423(Brand/Generic Indicator) - 0.0853789(Route) - 0.0007751(Age) - 0.081935(Gender) - 0.0350313(Claim History) - 0.0163794(Avg. Days Between History) + 0.0656391(Avg. Fill Days Between History) (4)

Conclusion

Even though the final model provided better results than the original version, it is still not enough to provide accurate predictions of patient behavior. In contrast from the first LM and GBM models, the final LM model had a better R2 than the GBM model due to the fact that part of the model was linearized when the GPI4 Rank and the Medical Indicator Rank values were created.

A predictive model cannot be created only using demographic data, it is established that a minimum of historical data per patient is required to generate an accurate predictive model.

It was also established that R language is a memory intensive environment which hindered the usage of the full extracted data that was originally selected to create an ideal model. In the end, to generate a model a total of 5 million records were used from the original 20 million records.

References

[1] Docs.microsoft.com. (2018). Install SQL Server 2017 Machine Learning Services (In-Database) on Windows [Online]. Available: https://docs.microsoft.com/en-us/sql/ advanced-analytics/install/sql-machine-learning-services-windows-install?view=sql-server-2017. [Accessed: May 4, 2018].
[2] Findacode.com. (2018). GPI Codes - Generic Product Identifiers from Wolters Kluwer - Drugs and Pharmaceuticals [Online]. Available: https://www.finda code.com/drugs/gpi-codes.html. [Accessed: May 5, 2018].
[3] F. Berhane. (2018). Anomaly detection with R [Online]. Available: http://datascience-enthusiast.com/R/anomaly_detection_R.html. [Accessed: May 4, 2018].
[4] Docs.microsoft.com. (2018). How to choose machine learning algorithms [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning /studio/algorithm-choice. [Accessed: May 4, 2018].