Doesn’t Keep Appointments (DKA): Model Validation and Global Sensitivity Analysis

David J. Wong

Under supervision of Stephen Zuniga

University of California, Los Angeles; Kaiser Permanente

**Introduction**

The goal of our study is to recommend a prognostic model1 that balances the highest levels of performance and practicality to aid hospital operational logistics in efficiency by predicting which patients will be at high risk of not showing up to their appointment(s), what we call DKA, and further to suggest what they can do to reduce the number of DKAs. We used model validation to compare regression models with regards to performance and practicality. We also performed global sensitivity analysis to identify high impact areas (what will give you the most bang for your buck if given focused attention). Machine learning techniques were initially considered as possible models but were ruled out after examining their poor performance and data-specific restrictions.

**Data**

The data collected came from 375,398 encounters (9,069 patients) during the time from January 2010 to August 2014.

**Model Validation**

We compared eight different models in our model validation. We compared the models based on calibration2and discrimination3 measures. We had two sets of four models, with the first four created without multiple imputation4, and the latter four with multiple imputation. The first and fifth model only differ by multiple imputation, as do the second and sixth model and so on. Table 1 shows the model variables.

|  |  |
| --- | --- |
| Predictors | |
| age | visit\_provider\_group |
| age2\*\* | Specialty |
| gender | med\_center |
| language | Prior\_dka\_rate |
| Copay | Prior\_dka\_rate2\*\* |
| copay2\*\* | prior visits |
| Weekday | prior visits2\*\* |
| Hour | Age x gender\*\*\* |
| hour2\*\* | Weekday x hour\*\*\* |
| days\_wait | dka\* |
| days\_wait^2\*\* | race\* |
| **Table 1**  \* Variables used only in the multiple imputation. Outcome is DKA.  \*\*These are squared term variables (age2 is the squared term of the linear term age)  \*\*\*These are interaction terms | |

***Model Descriptions***

First model (Full linear model): this model was our gold standard model that we compared all other models’ performance and practicality against. It includes only linear terms.

Second model (Full interactions model): this model includes linear terms and the two interaction terms (Age x gender, Weekday x hour).

Third model (Full non-linear splines model): this model includes linear terms and adds non-linear splines.

Fourth model (Full quadratic model): this model includes linear terms and adds the squared terms (variables followed by a 2).

The fifth, sixth, seventh, and eighth models include the same terms as the first, second, third, and fourth models respectively, but are created after using multiple imputation to address missing data.

**Results**

Table 2 displays the results of our model validation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Discrimination | | Calibration |
| Model | Model C-index | C-index | R2 | MAE | MSE |
| Full (linear) | 0.806 | 0.805 | 0.217 | 0.008 | 0.00031 |
| Full (interactions) | 0.806 | 0.805 | 0.217 | 0.008 | 0.00036 |
| Full (non-linear) | 0.820 | 0.819 | 0.247 | 0.002 | 0.00003 |
| Full (quadratic) | 0.813 | 0.812 | 0.237 | 0.002 | 0.00002 |
| MI Full (linear) | 0.805 | 0.786 | 0.188 | 0.008 | 0.00026 |
| MI Full (interactions) | 0.805 | 0.786 | 0.188 | 0.008 | 0.00026 |
| MI Full (non-linear) | 0.820 | 0.814 | 0.237 | 0.002 | 0.00002 |
| MI Full (quadratic) | 0.813 | 0.812 | 0.237 | 0.002 | 0.00600 |
| Winning model (reduced) | 0.806 | 0.805 | 0.227 | 0.003 | 0.00002 |
| Note: The total R2 and C-index of the complete model used to develop the approximated model is 0.237 and 0.813. The predictors of the reduced model are prior DKA rate, prior DKA rate^2, days wait, days wait^2, and specialty. Bootstrap Validation, split-sample, and ten-fold cross-validation resulted in similar model assessment values (e.g., C-index= 0.81). Subsequently, cross-validation was used throughout the analysis. The full linear, full non-linear, and full quadratic models were all significantly different based on a likelihood ratio test (p < 0.0001). The full linear model and the full reduced model are not significantly different. | | | | |
| MAE: Mean absolute error (mean [Y - Y hat]) | | |  |  |  |
| MSE: Mean squared error (mean (Y - Y hat)^2) | | |  |  |  |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

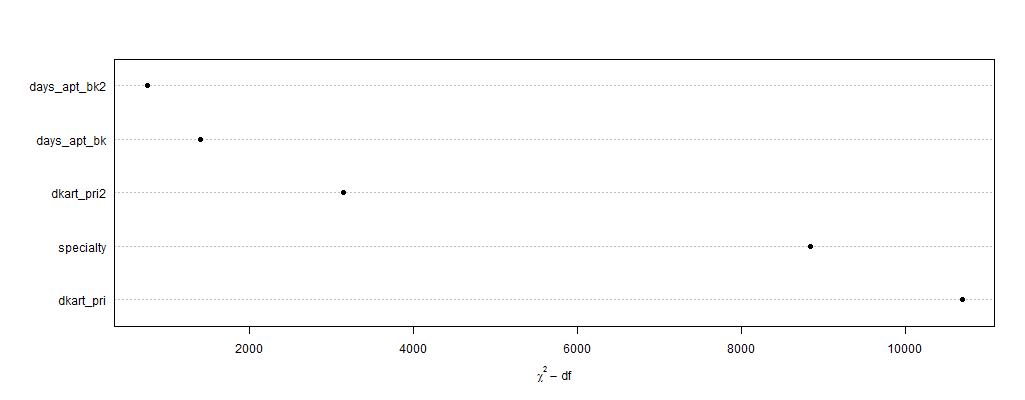
**Table 2**

***Results Interpretation***

Our winning model is the full quadratic model without multiple imputation. Though the multiple imputation full non-linear splines model performed the best based on discrimination and calibration, it was also the most complex model and therefore not the most practical. The winning model was selected because it performs at similar levels with the best model but with less complexity. The reduced model performs as well as the full linear model but with only 5 predictors. Imputation helped with calibration and non-linear splines increased performance. Using model approximation, we reduced the quadratic model and found that it performed just as well as our full linear model looking at discrimination but with better calibration; this reduced model is even simpler than the full model and still retains 96% of the R-squared as we can see in table 3.

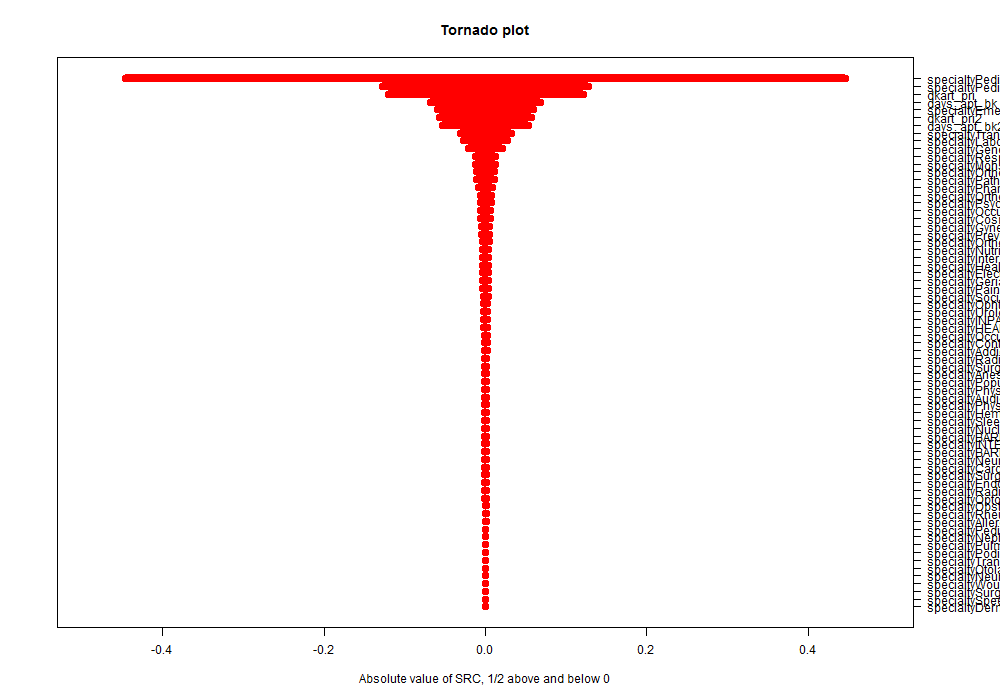
|  |  |  |
| --- | --- | --- |
| Deleted X | Proportion of remaining R2 | Total Remaining R2 |
| age2 | 1.0000000000 | 0.237 |
| vst\_pri | 0.9999998000 | 0.237 |
| gender | 0.9999984000 | 0.237 |
| vst\_pri2 | 0.9995935000 | 0.236904 |
| copay2 | 0.9991274000 | 0.236793 |
| copay | 0.9990607000 | 0.236777 |
| weekday | 0.9984463000 | 0.236632 |
| hour | 0.9977642000 | 0.23647 |
| hour2 | 0.9974693000 | 0.2364 |
| language | 0.9960991000 | 0.236075 |
| med\_ctr | 0.9937943000 | 0.235529 |
| visit\_prov\_group | 0.9902274000 | 0.234684 |
| age | 0.9761841000 | 0.231356 |
| days\_apt\_bk2 | 0.9604290000 | 0.227622 |
| days\_apt\_bk | 0.9551189000 | 0.226363 |
| dkart\_pri2 | 0.9337603000 | 0.221301 |
| dkart\_pri | 0.8045697000 | 0.190683 |
| specialty | 0.0000000000 | 0 |

**Table 3:** Reduced quadratic model has only 5 predictors and retains 96% of original R-squared.

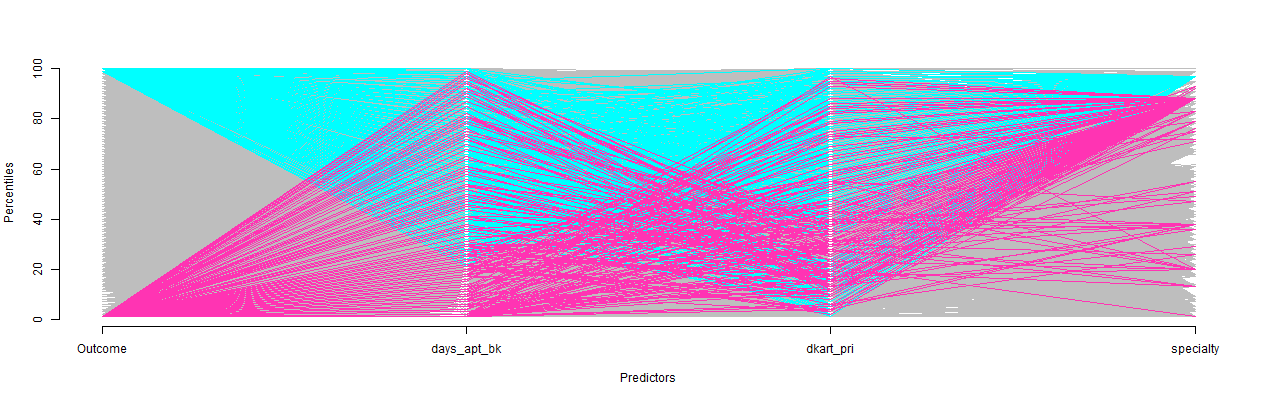
***Predictor Importance***

**Importance Plot:** shows us how relatively important (χ2- df) each predictor is in our model with prior DKA rate and specialty being the most important.

**Global Sensitivity Analysis**

The purpose of global sensitivity analysis is to help ask and answer the “what if” questions and to help focus our attention on specific areas of impact. We look at 10,000 Monte Carlo simulations and their predicted scores.

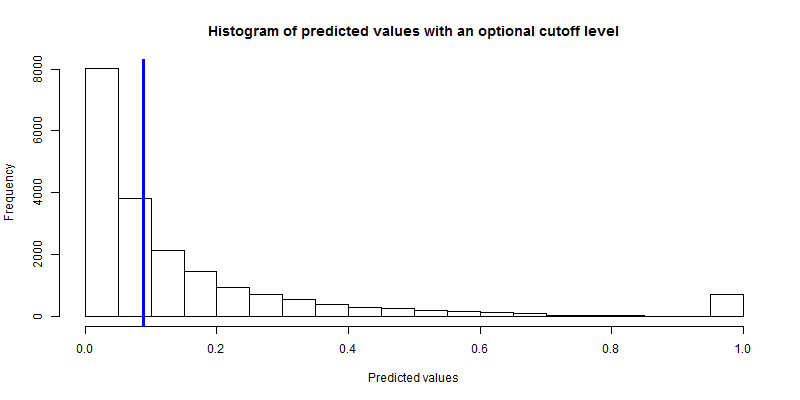
**Tornado Plot**: This shows which predictors the DKA outcome is most sensitive to. The width of the bar indicates importance. DKA is sensitive to 2 specialties and prior DKA rate and days wait.



**Cobweb Plot:** This shows the top 1% and bottom 1% of predicted scores and the concentrations around each predictor. Those most likely to DKA (cyan) are concentrated around psychiatry visits, have a longer days wait, and have a higher DKA rate. Those least likely to DKA (magenta) have a slightly greater concentration in their prior DKA rate and are found in multiple specialties.

**Conclusion**

What does this analysis mean for the operational logistics team? In our model validation, we compared models using calibration and discrimination measures and ended with one that both performs well and is simple to implement. Non-linear terms and quadratic terms were extremely helpful in improving the performance of the models and multiple imputation improved calibration. We reduced the model using model approximation ending with a model with only 5 predictors making it even simpler. In our global sensitivity analysis, we looked at important variables and areas of impact and asked our “what if” question. We found that days wait had a lot of impact on DKA so we consider this variable to be the focus of our attention following this study. The mean days wait is 9.78 days with a standard deviation of 16.1 days. We set a target cutoff at the current DKA rate and sought to reduce it from the current 8.8%. The graph below shows us that if we reduce the mean days wait to 5 days, we can expect 4.6% more cases below the target, which would mean fewer DKAs.



***Definitions***

1Prognostic model: a combination of multiple predictors from which risks can be assessed for individuals. Also known as a predictive model.

2 Calibration: refers to the reliability of a model; this is the ability of the model to predict future observations as well as it predicted the current observations. A perfectly calibrated model will predict the exact number of DKAs in our dataset. Measures used: MAE, MSE

3Discrimination how well a model assigns higher probabilities of risk to patients actually at higher risk. A perfectly discriminating model assigns the highest probabilities to those who get DKAs. Measures used: C-index, R-Squared

4Multiple Imputation: a statistical technique for analyzing missing entries in datasets.