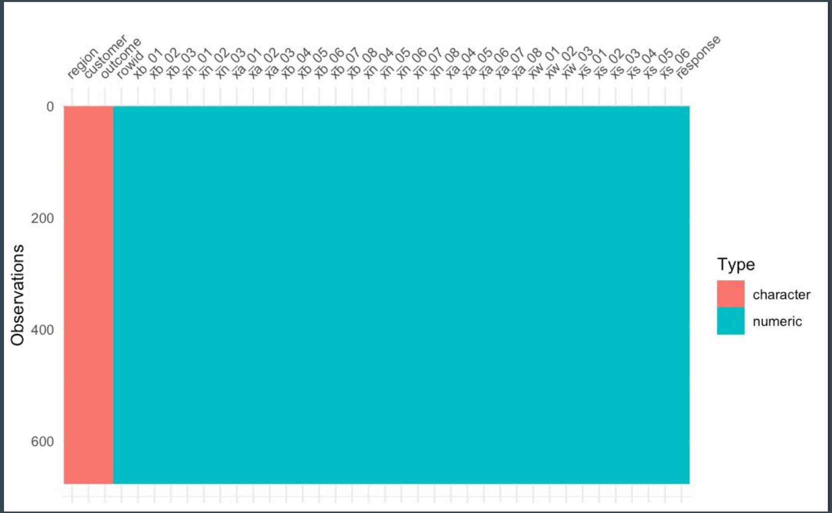
PPG CS 1675 Final Project

Sameera Boppana

Exploratory Data Analysis

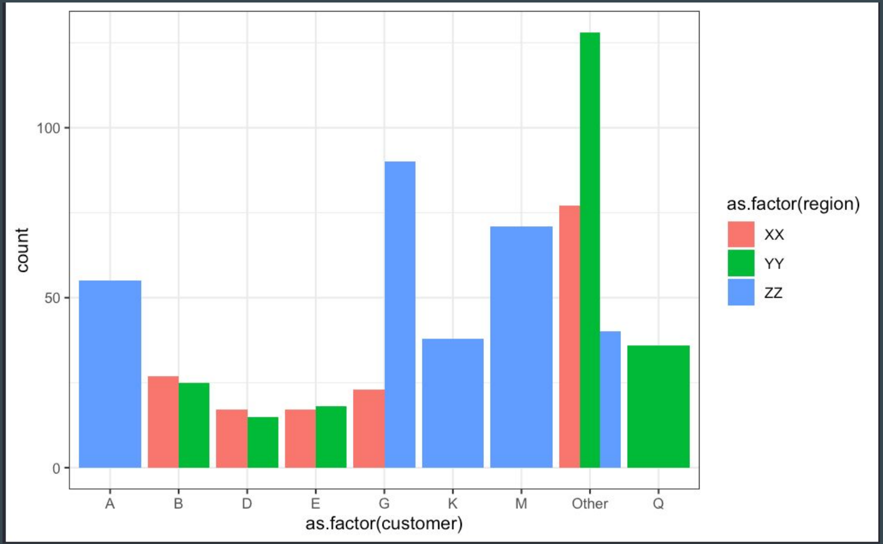
Input Data Types

- Inputs found within the training data are either characters or numeric
- Character Inputs: region, customer, outcome
- Numeric inputs include all sentiment derived features



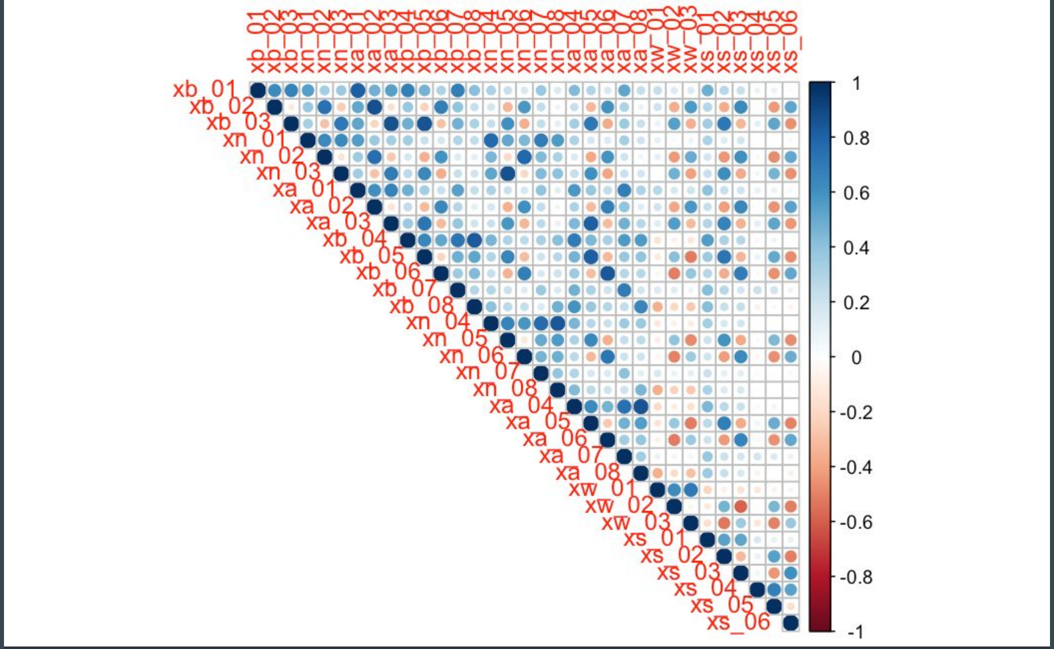
Customer vs. Region

- Visualizing the breakdown of the number and type of customer found in each region
- Customer A, K and M are only located in region ZZ
- Customer Q is only located in region YY



Correlation of Continuous Inputs

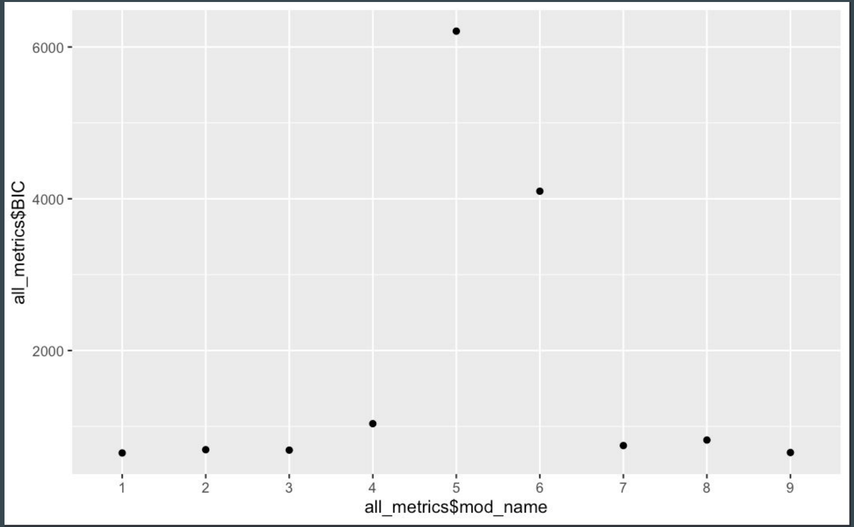
- Visualizing the correlation between all the continuous sentiment derived features
- Appears to be many inputs that are highly correlated with each other
- This may cause problems when determining important variables



Regression Modeling

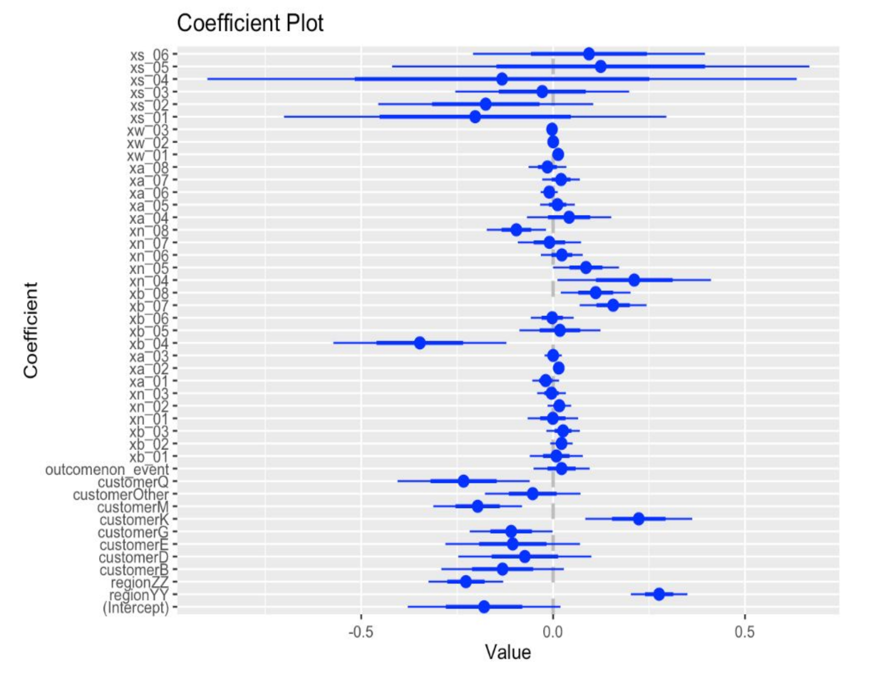
Selecting Best Regression Model

- Looking at the BIC values of the each of the 9 trained regression models
- Model 3 (linear additive with categorical + continuous inputs) has the lowest BIC value and is selected as the best



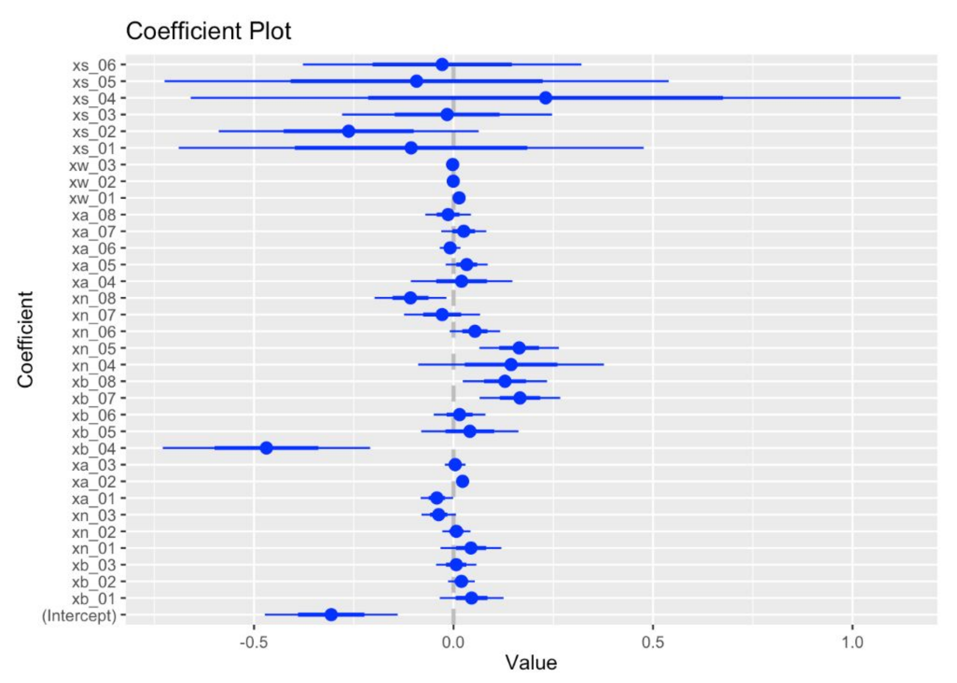
Visualizing Coefficients of the Best Model

- Looking at the coefficients of the linear additive model with both the categorical and continuous inputs
- Statistically significant inputs: regionYY, regionZZ, customerB, customerG, customerK, customerM, customerQ, xa_02, xb_04, xb_07, xb_08, xn_04, xn_05 , xn_08, xw_01, xw_03



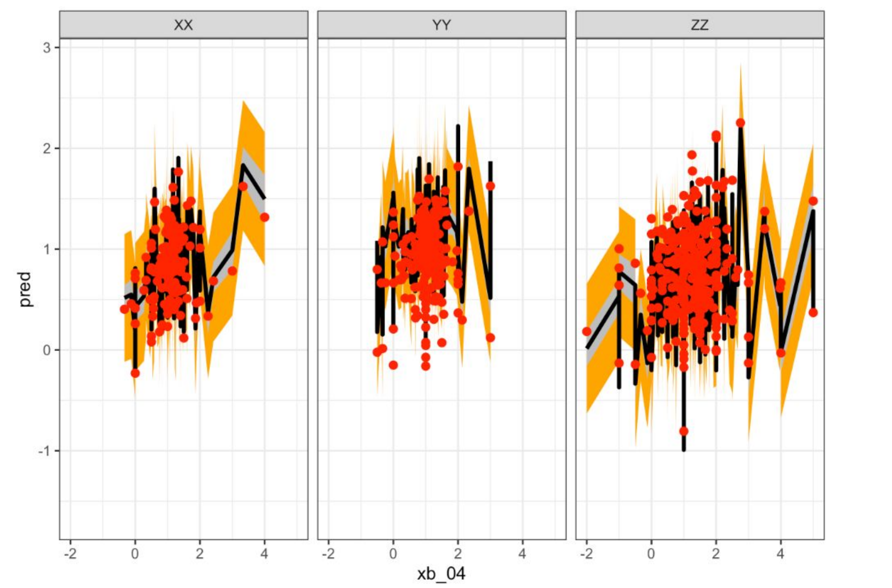
Visualizing coefficients of Continuous Input Model

- To compare the results of the best model selected by the lowest BIC value, the model with only the continuous inputs was further examined
- Significant coefficients: xn_03, xa_01, xa_02, xb_04, xb_07, xb_08, xn_08, xw_01
- Less continuous inputs are significant when categorical variables are not present



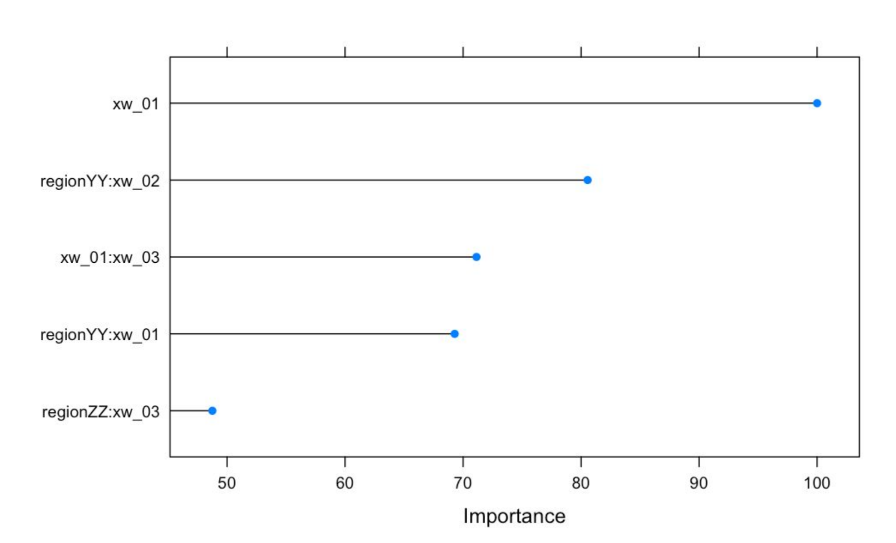
Model Predictions

- Looking at the different regions and against input “xb_04”
- There high concentration of values between 0 and 2 for xb_04
- Predictions amongst all regions concentrated around 0 - 2 hours per week
- These predictions represent log-transformed hours



Variable Importance

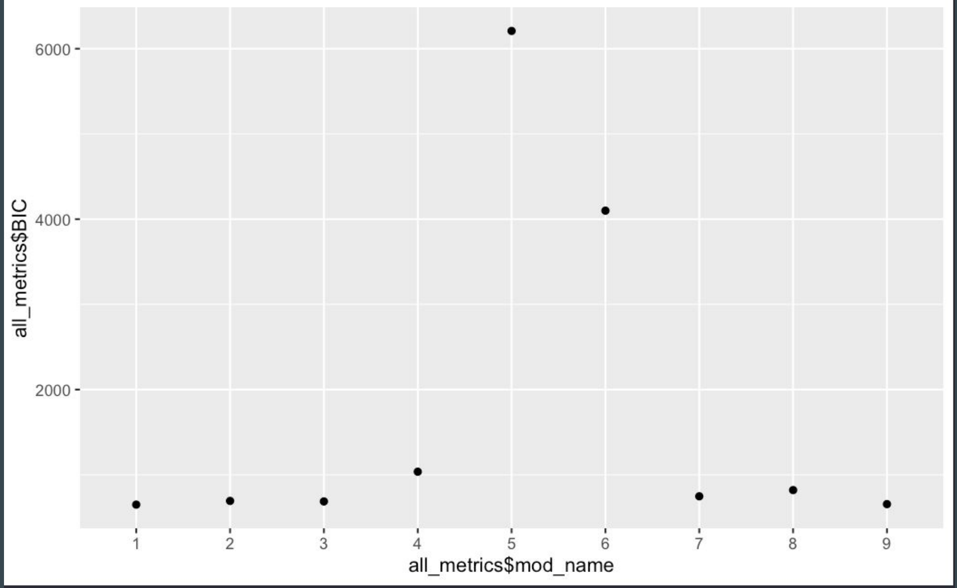
- After training, tuning, and evaluating the model with the pairwise interaction terms trained through elastic net had the best performance determined through RMSE values
- The variable with the highest contribution is xw_01



Classification Models

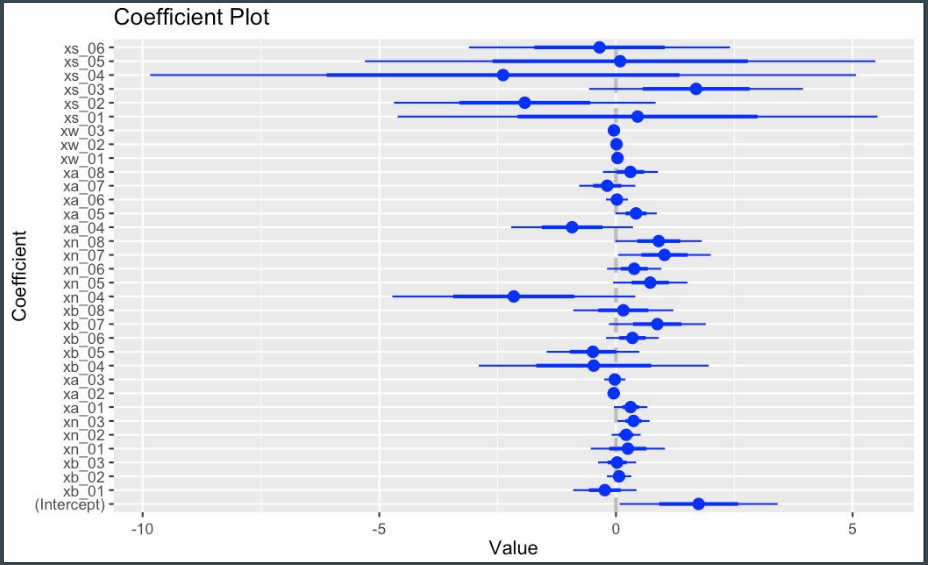
Selecting Best Classification Model

- Looking at the BIC values, the best classification model is model 1 (model with only continuous inputs)
- Selected due to lowest BIC value



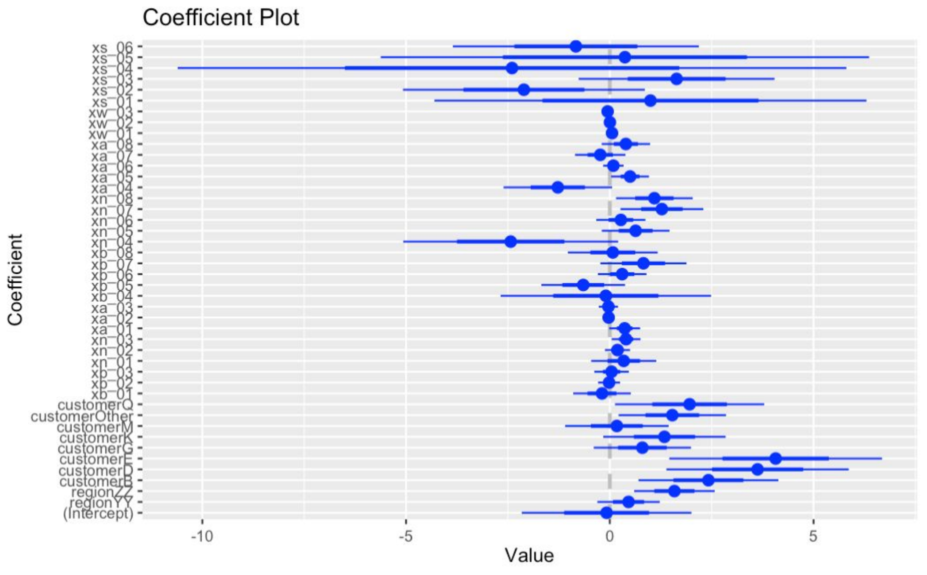
Visualizing coefficients of the Best Model

- Now the statistically significant coefficients are xw_03, xa_05, xn_07, xn_08, xn_04, xn_05, xb_07, xa_01, xn_03



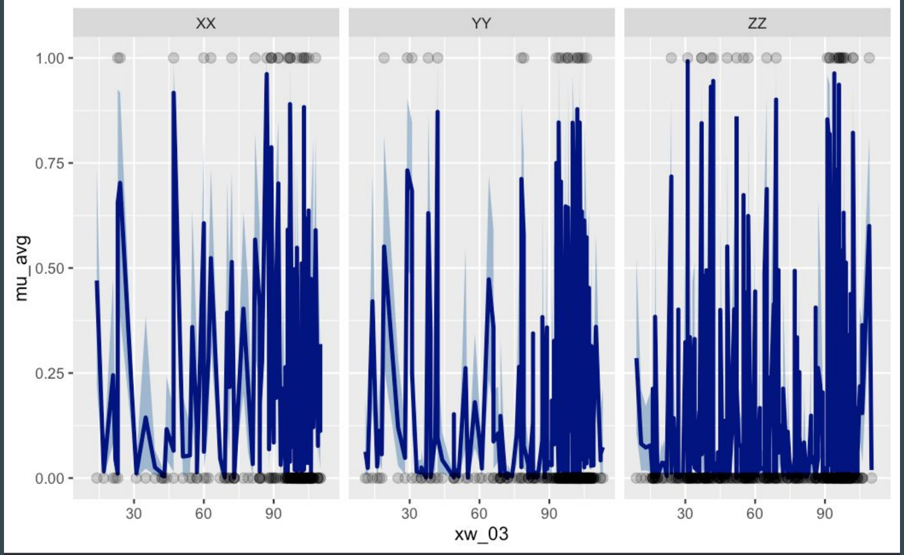
Visualizing coefficients of Categorical + Continuous Inputs

- Statistically Significant Inputs: xw_03, xw_01, xa_05, xa_04, xn_08, xn_07, xn_04, xa_01, xn_03, CustomerQ, CustomerOther, CustomerK, CustomerE, CustomerD, CustomerB, Region ZZ
- Compared to Regression, there are more categorical inputs that are significant



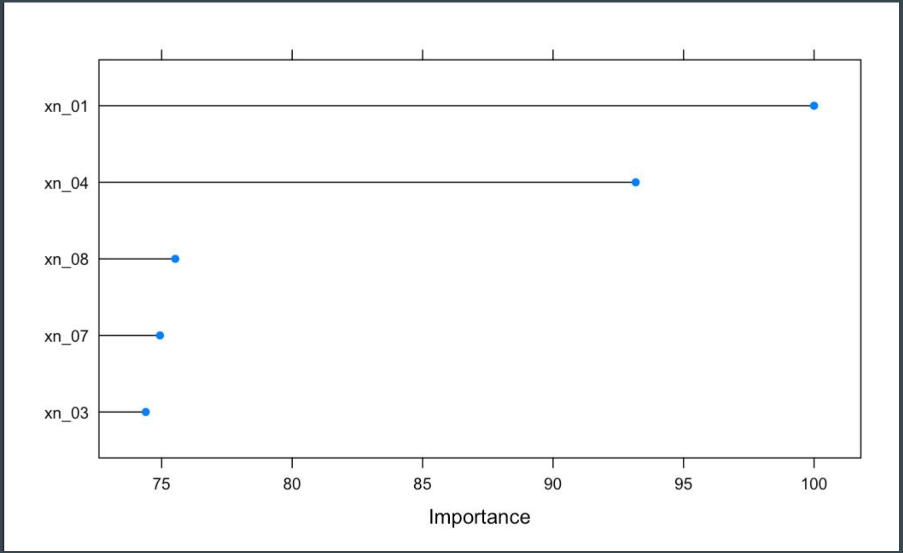
Predictive Trends of Categorical + Continuous Model

- Regions plotted against continuous input xw_03
- A common trend between all three regions is that there is a higher concentration of non-events than events in the higher values of xw_03



Variable Importance

- After training, tuning, and evaluating the model with the categorical and continuous inputs trained through Partial Least Squares was selected as the best via Accuracy and AUC ROC performance metrics
- The variable with the highest predictive contribution is xn_01


xn_01
xn_04
xn_08
xn_07
xn_03
Importance