Notes:

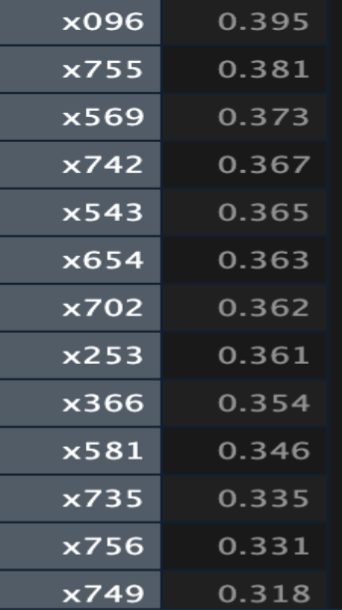
* Look at categorical variables
* Perform PCA on variables that are dropped
* Use different KNN imputation methods
* Research how to leverage deep learning and neural networks
* Kernel PCA
* Duplicate columns (92 duplicates)
* Basic Lasso model performing well
* I may be reducing the number of features to heavily
* 9.06 RMSE with lasso regression with standard scaler using median imputation
* Only keep columns with less than 5% missingness

See what is predicted to be above 25 remove it, then predict the below 25 values with the new training model, and then keep the original 25 predictions

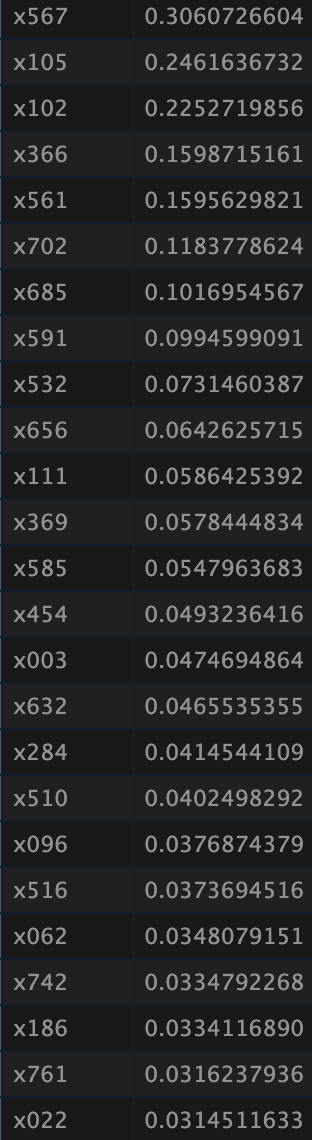
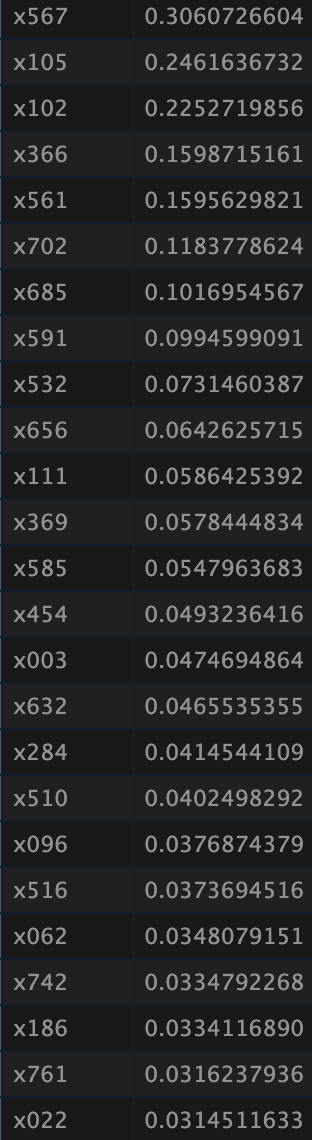
1. Filter Methods: These methods evaluate the relevance of features based on their statistical properties and are independent of the chosen model. Common techniques include:
   * Correlation-based methods: Identify features with high correlation to the target variable.
   * Variance thresholding: Remove features with low variance as they provide little information.
   * Mutual information: Measure the dependency between features and the target variable.
   * Chi-square test: Assess the relationship between categorical features and the target variable.
2. Wrapper Methods: These methods use a specific model to evaluate the performance of different feature subsets. They are computationally expensive but provide more accurate results. Common techniques include:
   * Recursive Feature Elimination (RFE): Iteratively removes features based on their impact on model performance.
   * Forward/Backward Selection: Starts with an empty set of features and gradually adds/removes features based on their impact on model performance.
3. Embedded Methods: These methods incorporate feature selection within the model building process itself. They automatically select the most relevant features during training. Common techniques include:
   * L1 Regularization (Lasso): Adds a penalty term to the loss function that encourages sparsity, effectively selecting the most relevant features.
   * Tree-based methods: Decision trees and random forests can provide feature importance measures that guide feature selection.
4. Dimensionality Reduction: Instead of selecting individual features, dimensionality reduction techniques aim to create a lower-dimensional representation of the data while retaining the most important information. Common techniques include:
   * Principal Component Analysis (PCA): Creates a new set of uncorrelated variables called principal components.
   * Linear Discriminant Analysis (LDA): Maximizes class separability while reducing dimensionality.
   * t-SNE: Visualizes high-dimensional data by mapping it into a lower-dimensional space while preserving local structure.

Variable Selection:

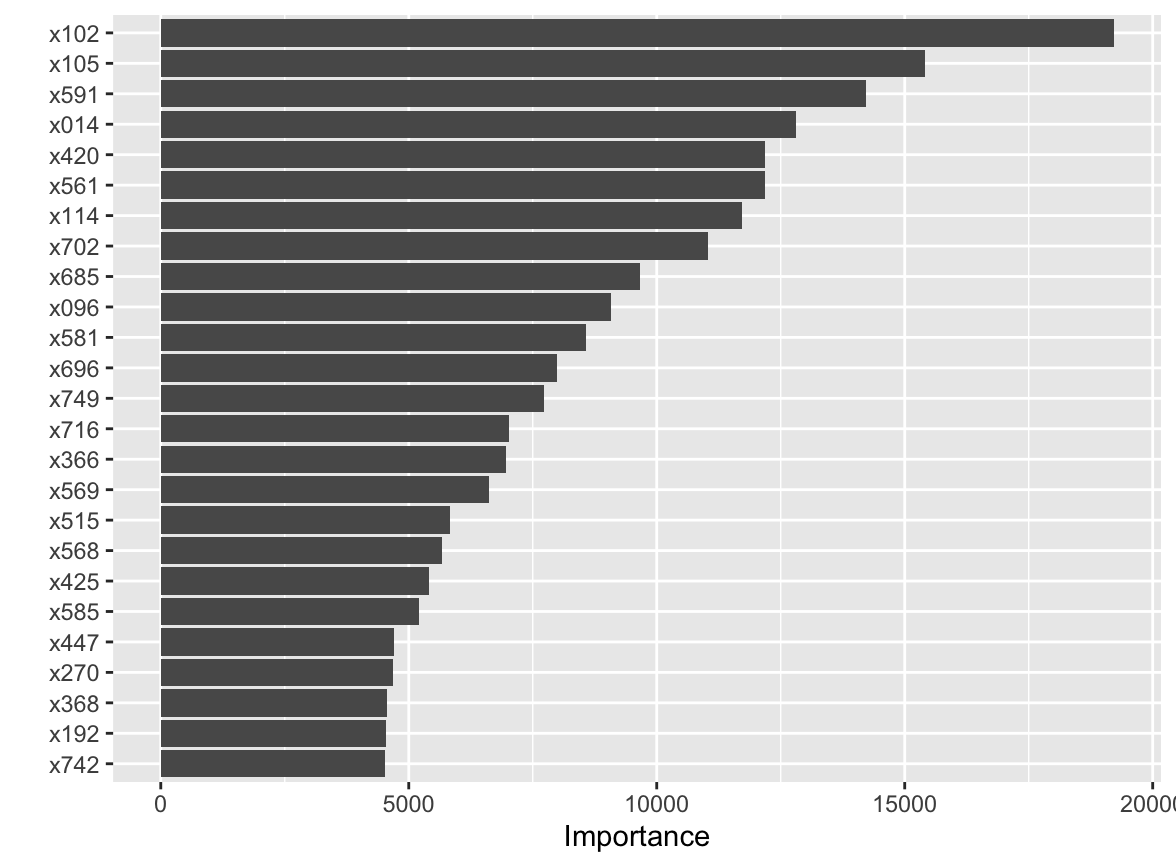
Correlation Matrix: [x102, x146, x118, x724, x619, x687, x670, x014, x561, x651, x696, x105, x096, x755, x569, x742, x543, x654, x702, x253, x366, x581, x735, x756, x749]

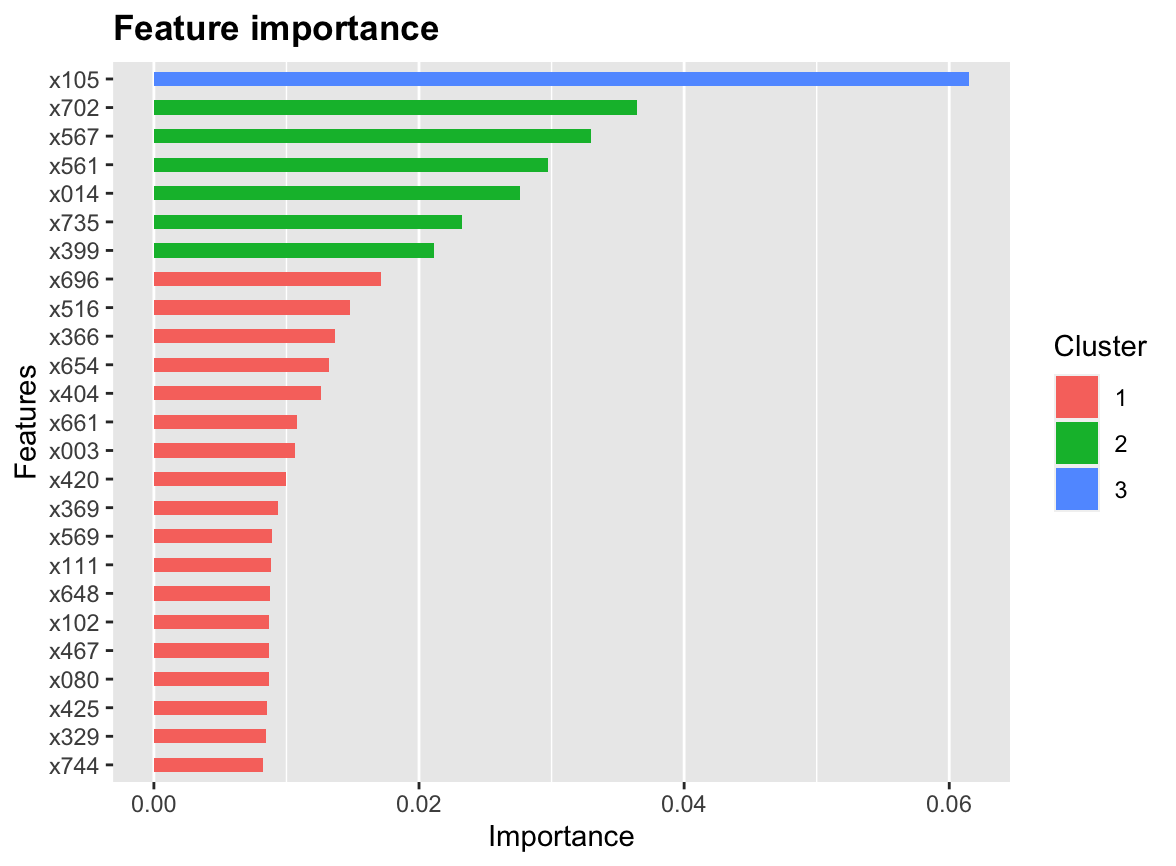


Lasso Feature Selection: [x567, x105, x102, x366, x561, x702, x685, x591, x532, x656, x111, x369, x585, x454, x003, x632, x284, x510, x096, x516, x062, x742, x186, x761, x022]

RF VIP: [x102, x105, x591, x014, x420, x561, x114, x702, x685, x096, x581, x696, x749, x716, x366, x569, x515, x568, x425, x585, x447, x270, x368, x192, x742]



BT VIP: [x105, x702, x567, x561, x014, x735, x399, x696, x516, x366, x654, x404, x661, x003, x420, x369, x569, x111, x648, x102, x467, x080, x425, x329, x744]

RF-BT Overlap: [x014, x102, x105, x366, x420, x425, x561, x569, x696, x702]

RF-CM Overlap: [x102, x014, x561, x696, x105, x096, x569, x742, x702, x366, x581, x749]

RF-Lasso Overlap: [x105, x102, x366, x561, x702, x685, x591, x585, x096, x742]

BT-CM Overlap: [x102, x014, x561, x696, x105, x569, x654, x702, x366, x735]

BT-Lasso Overlap: [x567, x105, x102, x366, x561, x702, x111, x369, x003, x516]

Lasso-CM Overlap: [x105, x102, x366, x561, x702, x096, x742]

RF-BT-Lasso Overlap: [x105, x102, x366, x561, x702]

RF-BT-CM Overlap: [x102, x014, x561, x696, x105, x569, x702, x366]

BT-CM-Lasso Overlap: [x105, x102, x366, x561, x702]

RF-CM-Lasso Overlap: [x105, x102, x366, x561, x702, x096, x742]

Total Overlap: [x105, x102, x366, x561, x702]

Variables: c(x105, x102, x366, x561, x702, x096, x742, x014, x696, x569, x567, x111, x369, x003, x516, x654, x735, x685, x591, x585, x581, x749, x420, x425)

Notes:

* Variable x735 is categorical
* Variable x516 is categorical
* Variable x742 is categorical

Imputation Method:

Didn’t matter

Model Selection:

Neural Network

Random forest

SVM\_rbf

Elastic Net

Ensemble model of these 4