Feature Engineering Report

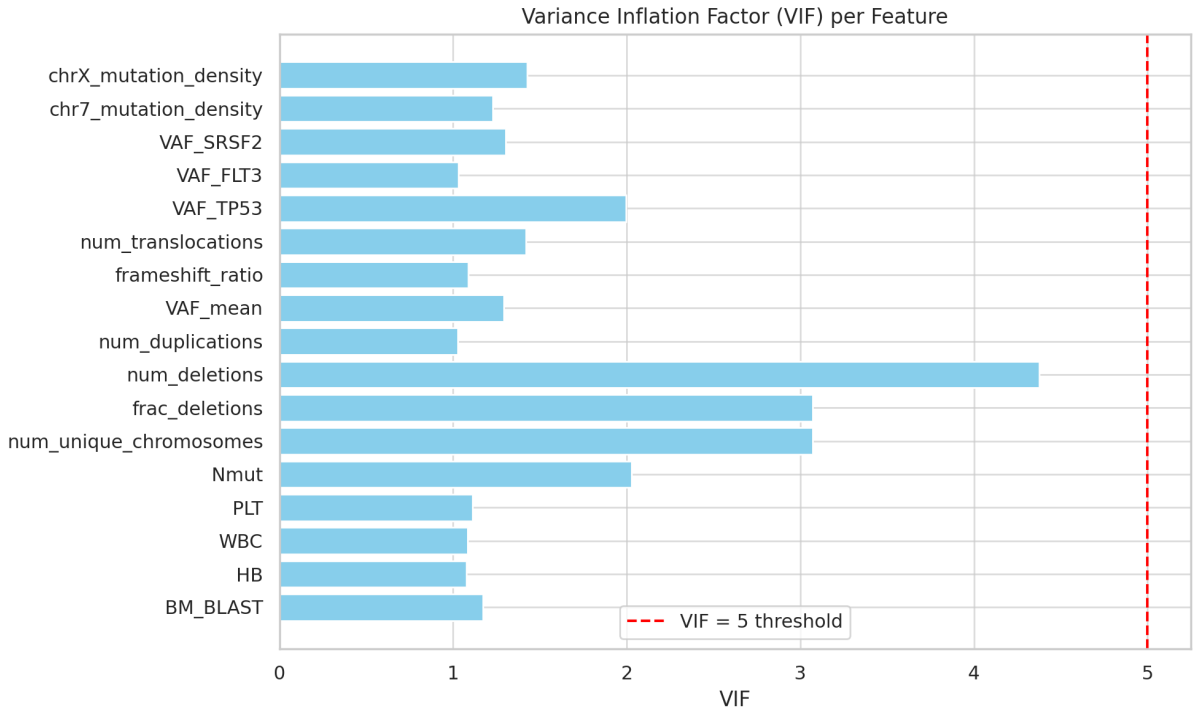
Survival Prediction for Adult Myeloid Leukemia

This report presents an in-depth exploratory analysis and feature engineering process conducted as part of the 2025 QRT Data Challenge in collaboration with Institut Gustave Roussy. The challenge aimed to build predictive models for overall survival (OS) in patients diagnosed with adult myeloid leukemia, using a dataset composed of clinical and molecular data. The dataset includes: - Clinical variables (e.g., blood counts, bone marrow blasts, karyotype abnormalities) - Molecular data from somatic mutations (e.g., affected genes, variant allele frequency) From these raw data, several informative features were derived to capture disease severity and progression patterns. Notable features include: - Mutation burden: number of mutations, unique genes, average VAF - Key gene-specific VAFs (TP53, FLT3, SRSF2) based on their relevance in hematologic malignancies - Chromosome-specific mutation densities (chr7 and chrX) - Cytogenetic abnormalities: deletions, duplications, translocations, and affected chromosomes The goal of this analysis is to visualize these features, assess their predictive potential, and ensure they are suitable for use in survival models.

Variance Inflation Factor (VIF) Analysis

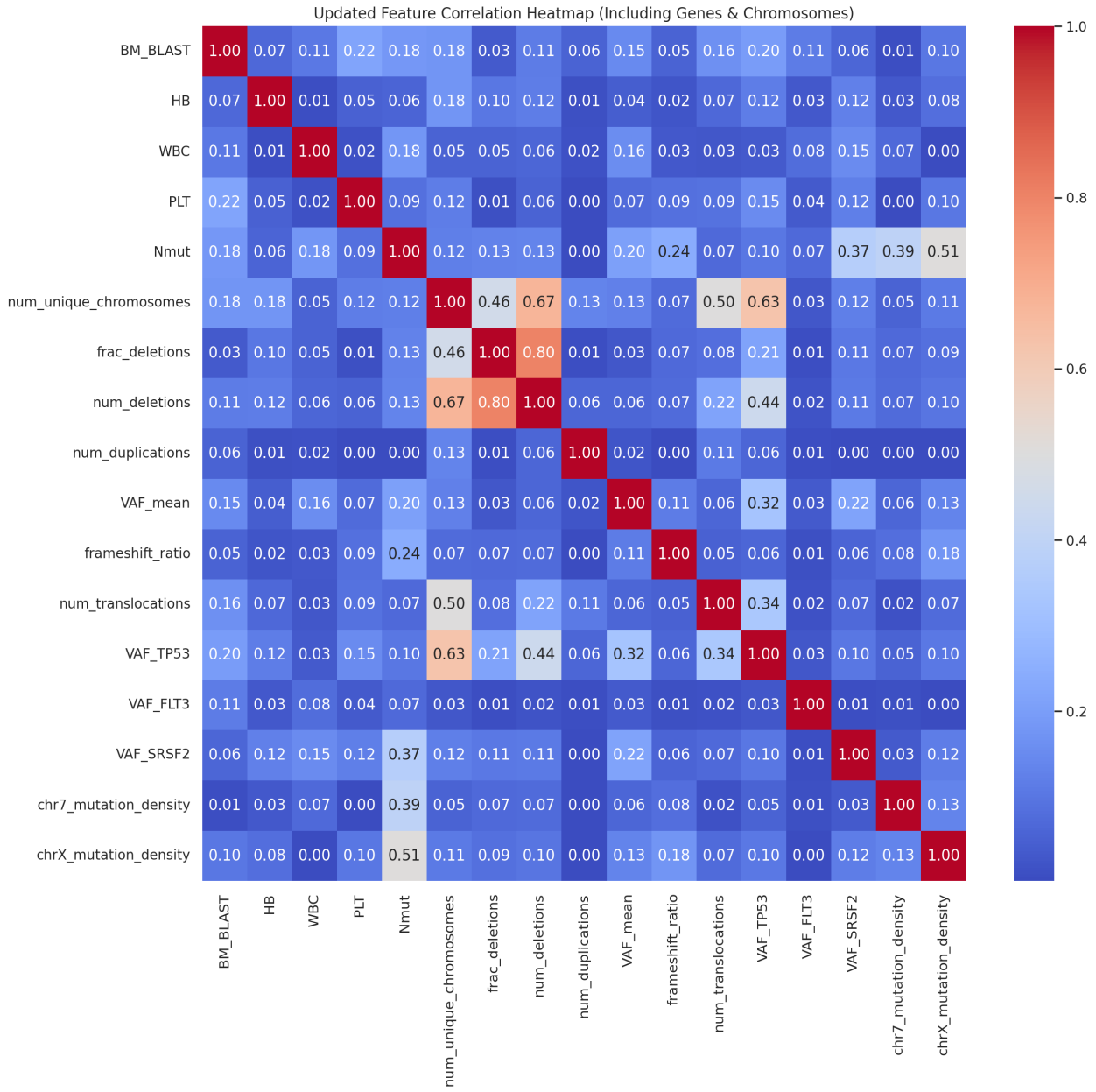
Feature	VIF
BM_BLAST	1.17
HB	1.08
WBC	1.08
PLT	1.11
Nmut	2.03
num_unique_chromosomes	3.07
frac_deletions	3.07
num_deletions	4.38
num_duplications	1.03
VAF_mean	1.29
frameshift_ratio	1.09
num_translocations	1.42
VAF_TP53	1.99
VAF_FLT3	1.03
VAF_SRSF2	1.3
chr7_mutation_density	1.23
chrX_mutation_density	1.43

The table above displays the Variance Inflation Factor (VIF) for each feature. Features with VIF less than 5 are considered free from multicollinearity issues.



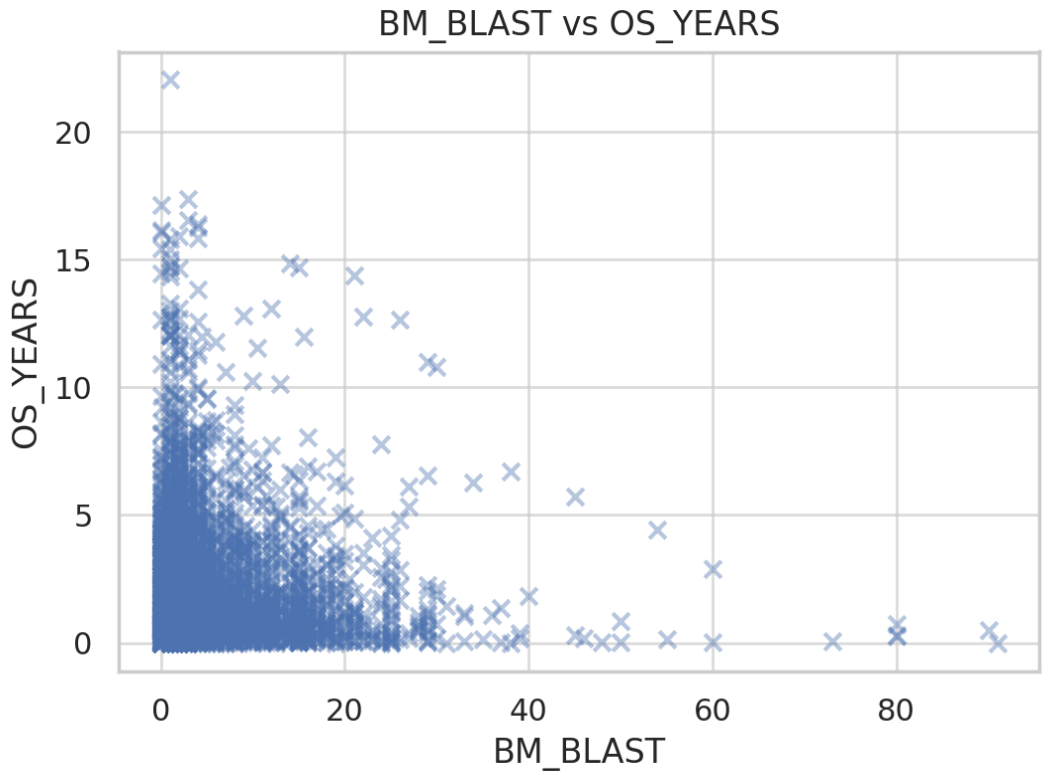
The bar chart illustrates the multicollinearity among features. VIF values above the red line at 5 suggest potential redundancy.

Feature Correlation Heatmap



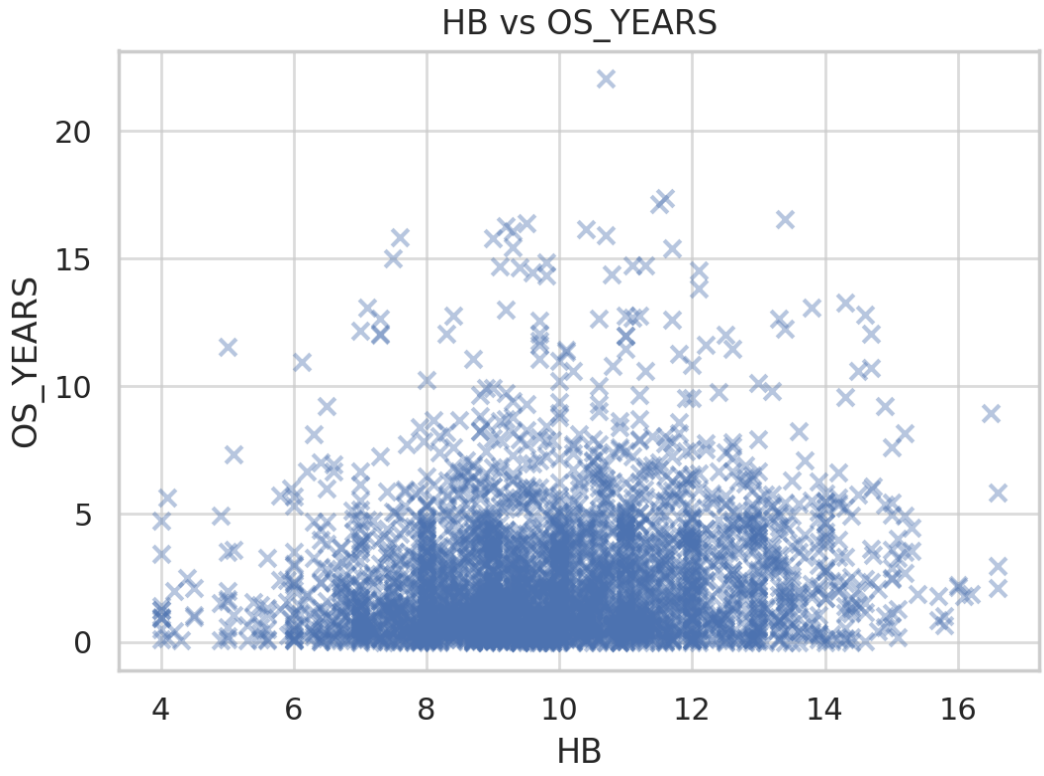
This heatmap displays pairwise correlations between engineered features. Strong correlations (|corr| > 0.8) are highlighted, indicating redundancy.

BM_BLAST vs OS_YEARS



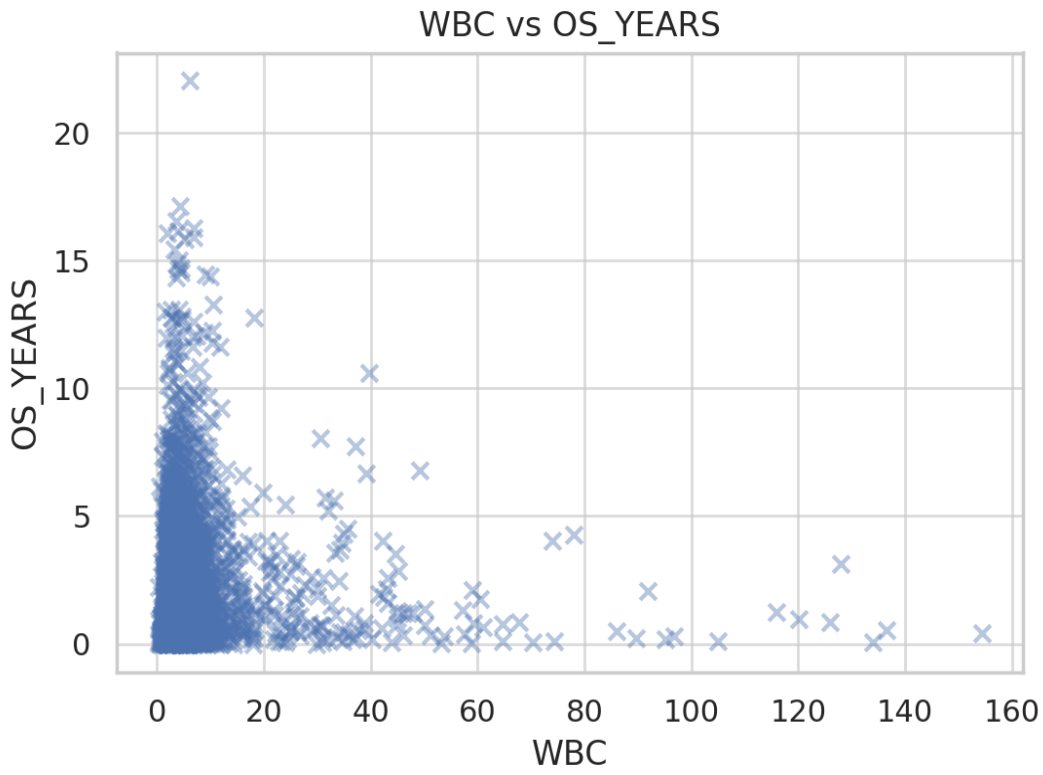
The plot above illustrates the distribution of the feature 'BM_BLAST' against overall survival in years. Patterns can highlight clinical relevance.

HB vs OS_YEARS



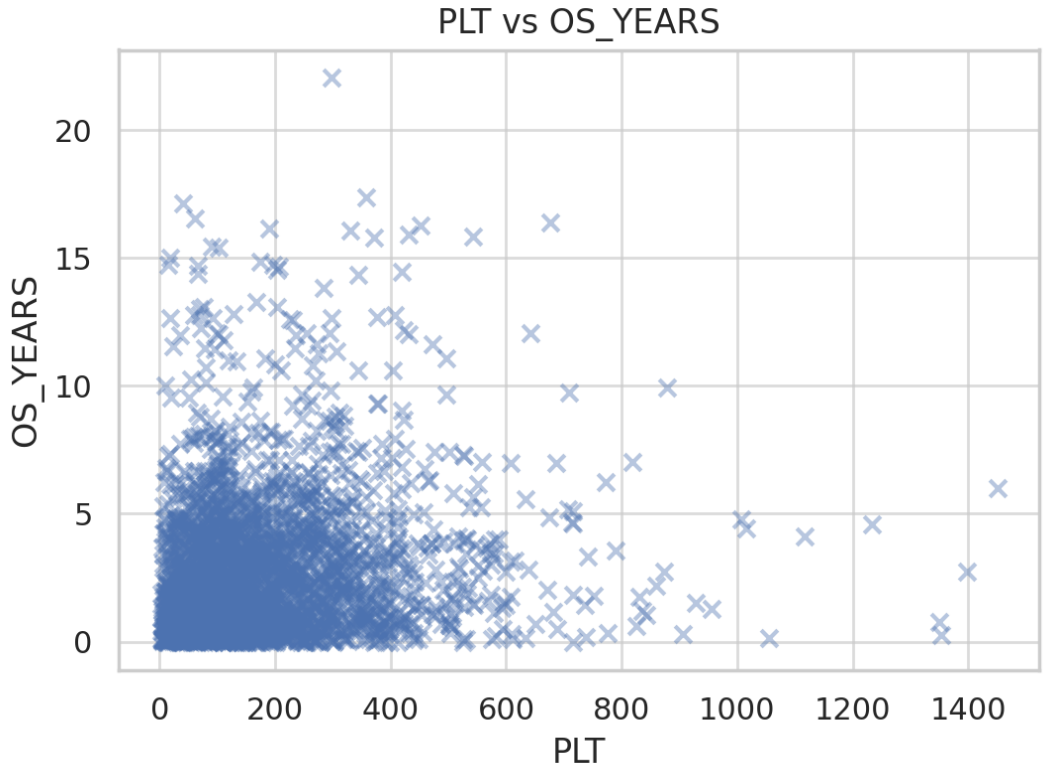
The plot above illustrates the distribution of the feature 'HB' against overall survival in years. Patterns can highlight clinical relevance.

WBC vs OS_YEARS



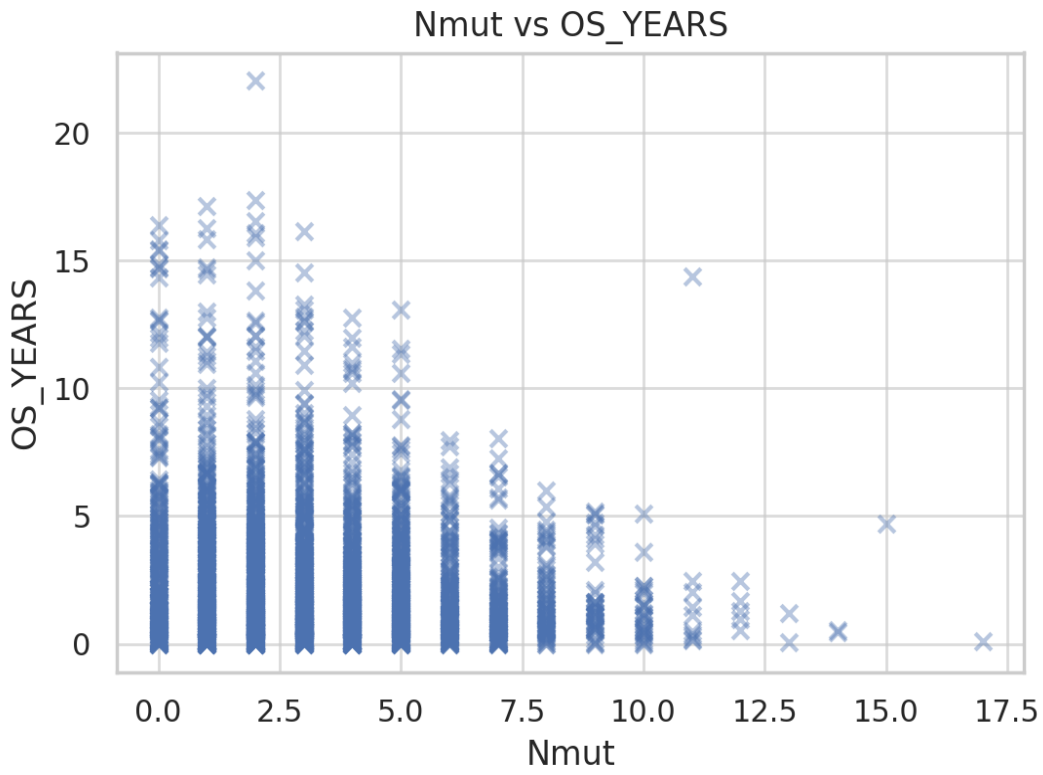
The plot above illustrates the distribution of the feature 'WBC' against overall survival in years. Patterns can highlight clinical relevance.

PLT vs OS_YEARS



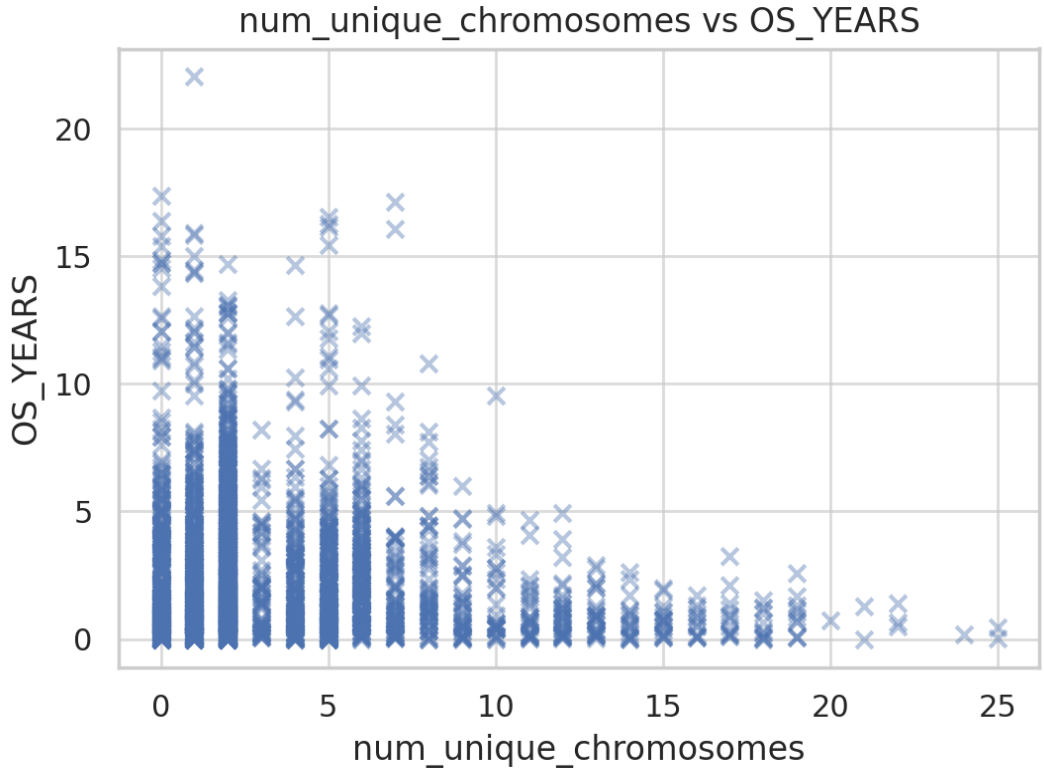
The plot above illustrates the distribution of the feature 'PLT' against overall survival in years. Patterns can highlight clinical relevance.

Nmut vs OS_YEARS



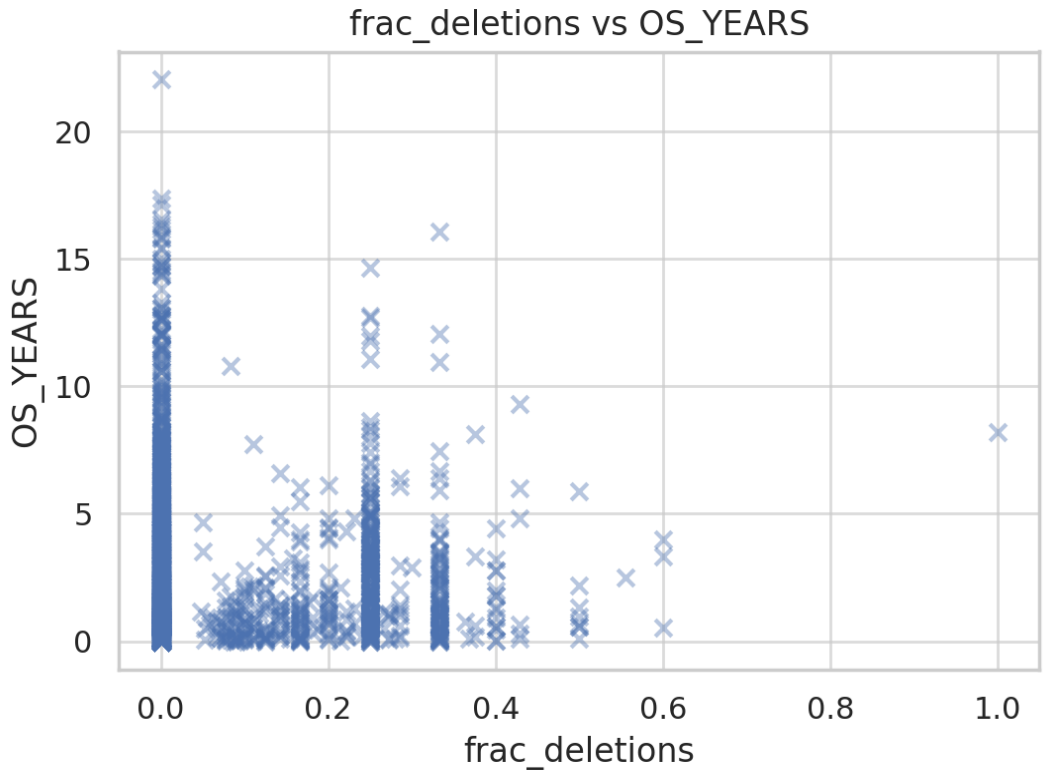
The plot above illustrates the distribution of the feature 'Nmut' against overall survival in years. Patterns can highlight clinical relevance.

num_unique_chromosomes vs OS_YEARS



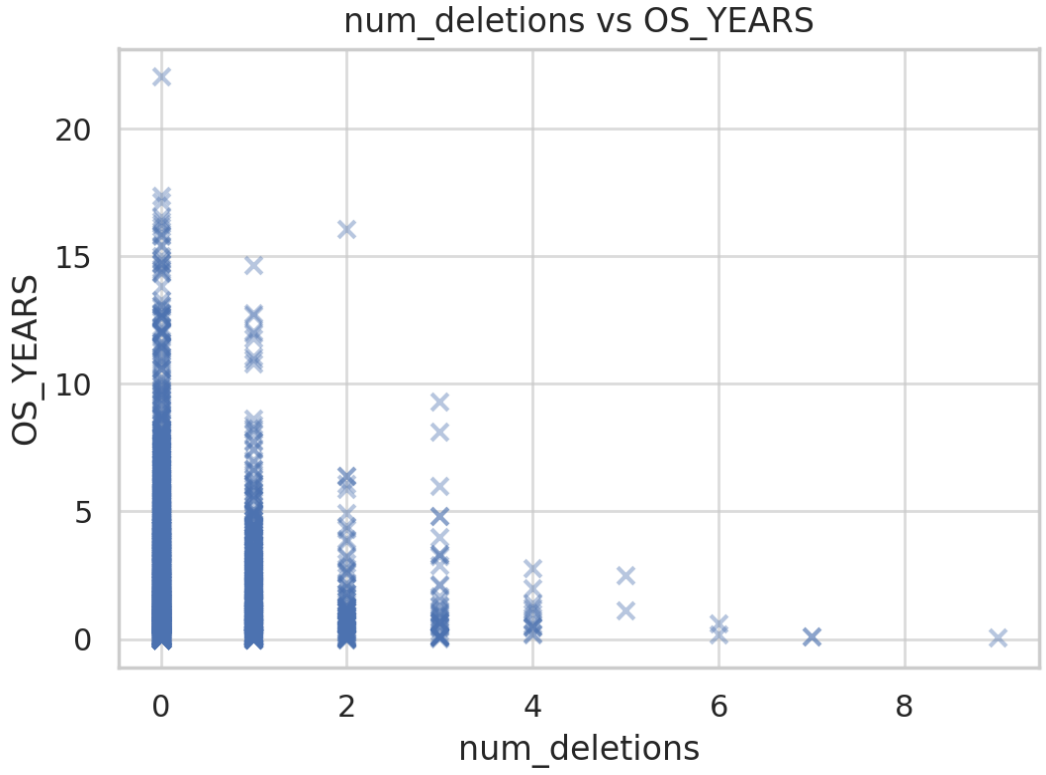
The plot above illustrates the distribution of the feature 'num_unique_chromosomes' against overall survival in years. Patterns can highlight clinical relevance.

frac_deletions vs OS_YEARS



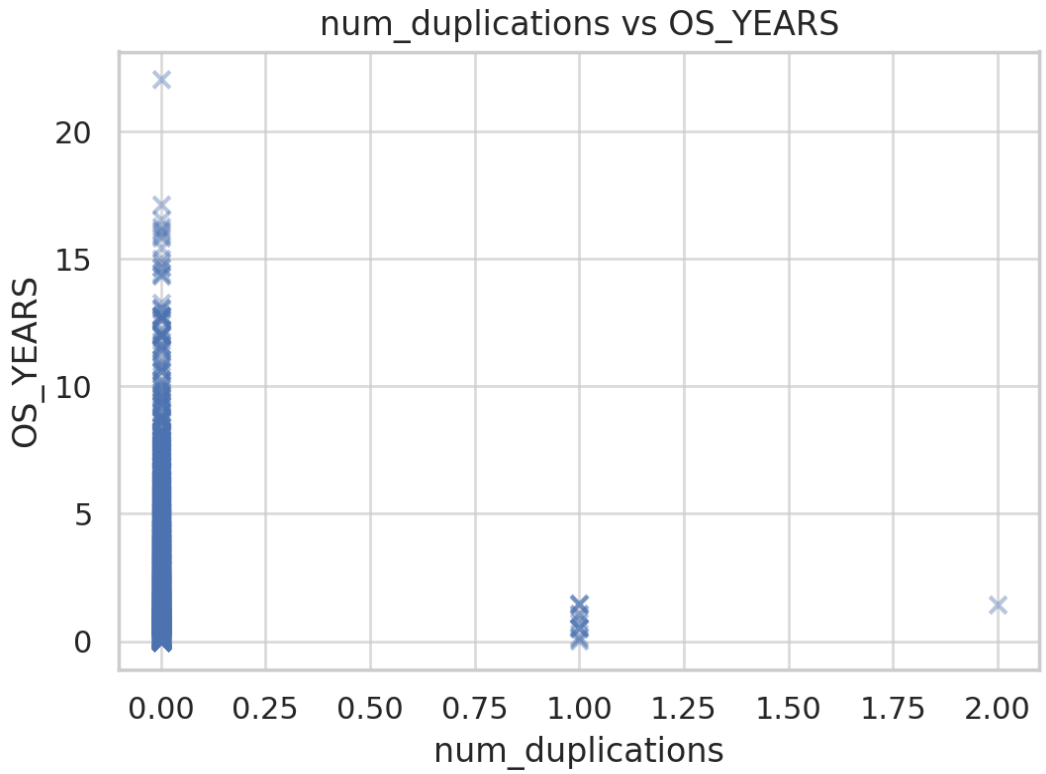
The plot above illustrates the distribution of the feature 'frac_deletions' against overall survival in years. Patterns can highlight clinical relevance.

num_deletions vs OS_YEARS



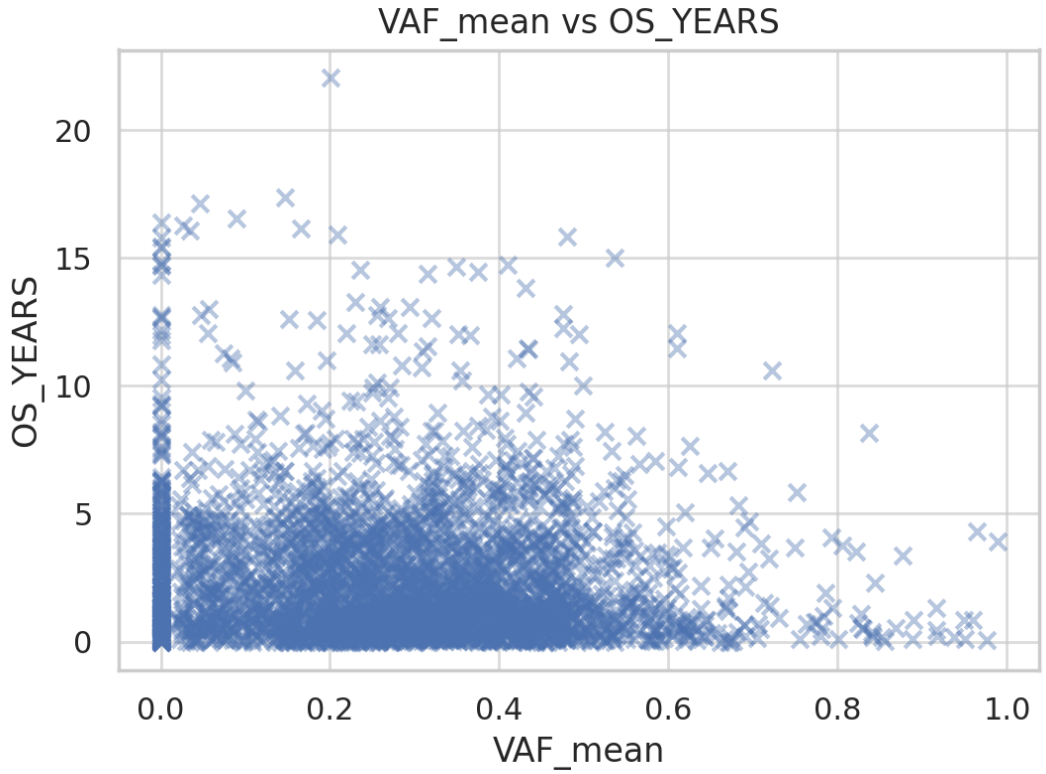
The plot above illustrates the distribution of the feature 'num_deletions' against overall survival in years. Patterns can highlight clinical relevance.

num_duplications vs OS_YEARS



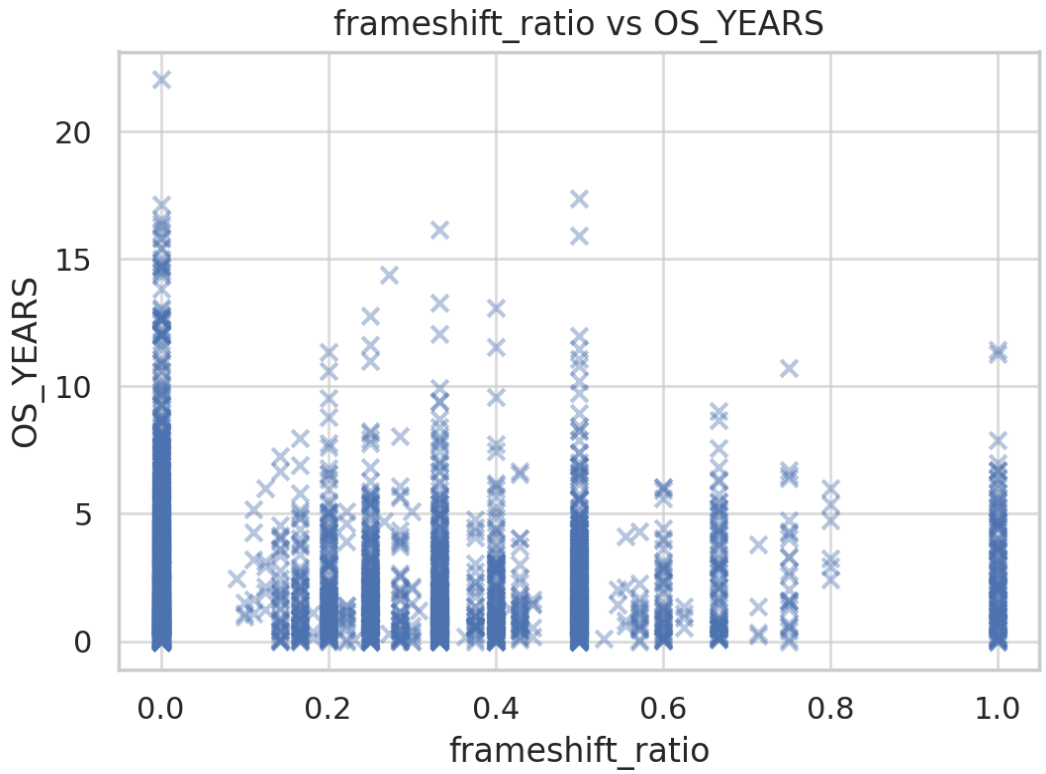
The plot above illustrates the distribution of the feature 'num_duplications' against overall survival in years. Patterns can highlight clinical relevance.

VAF_mean vs OS_YEARS



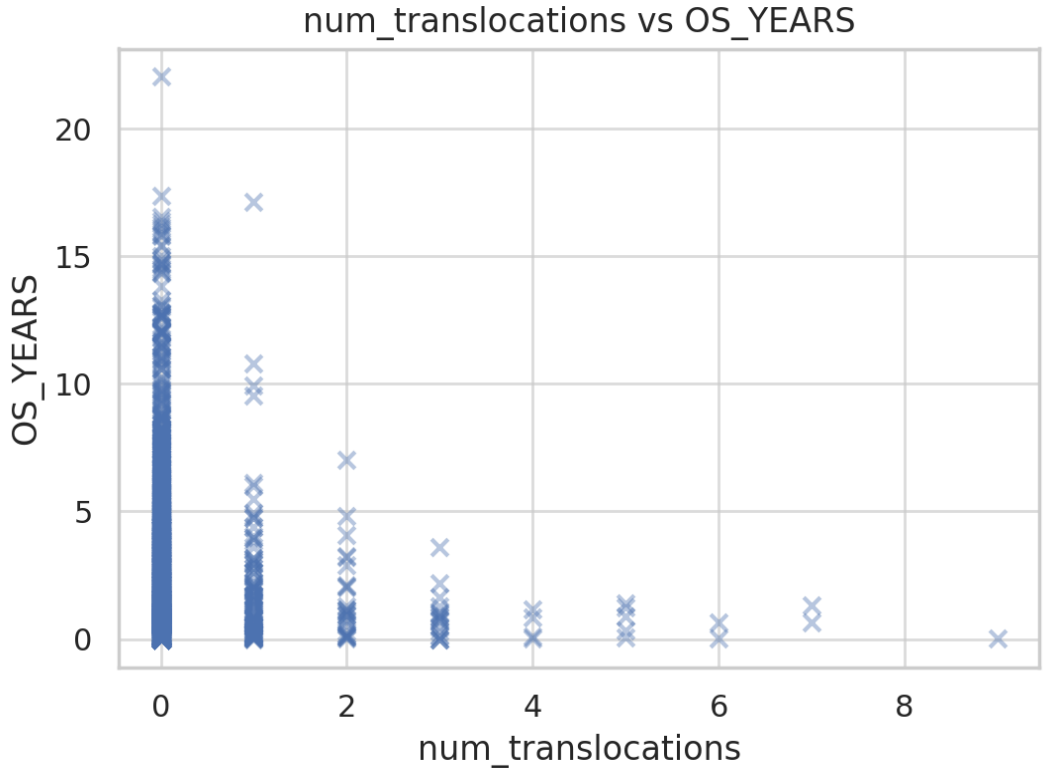
The plot above illustrates the distribution of the feature 'VAF_mean' against overall survival in years. Patterns can highlight clinical relevance.

frameshift_ratio vs OS_YEARS



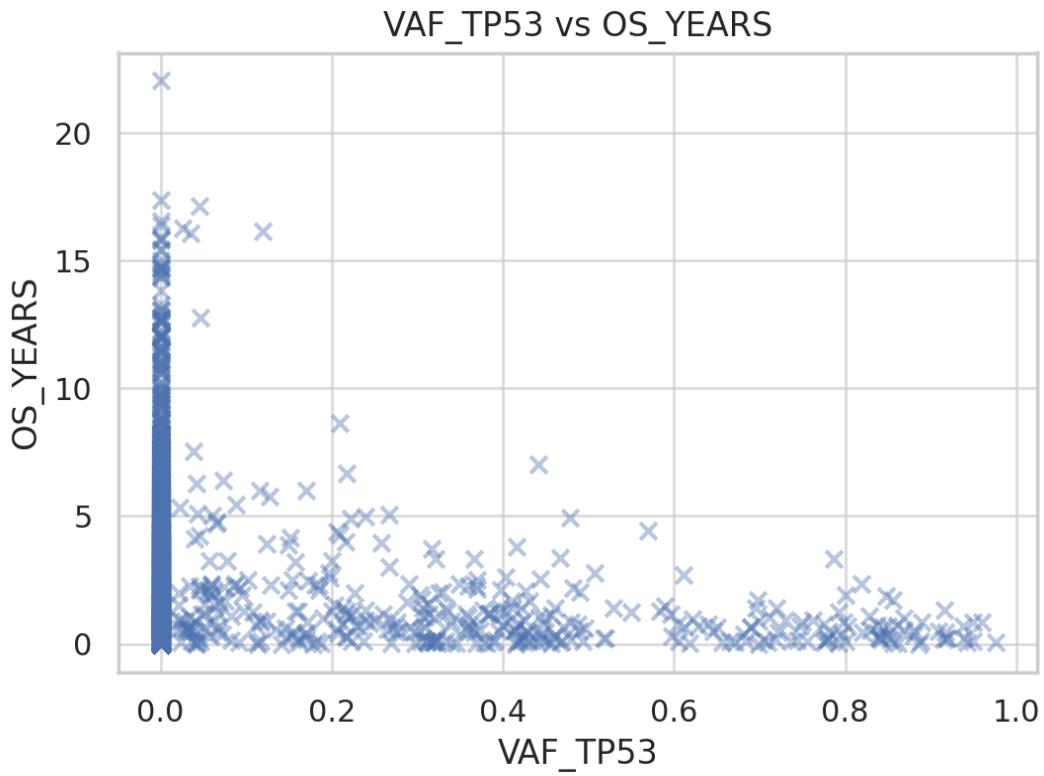
The plot above illustrates the distribution of the feature 'frameshift_ratio' against overall survival in years. Patterns can highlight clinical relevance.

num_translocations vs OS_YEARS



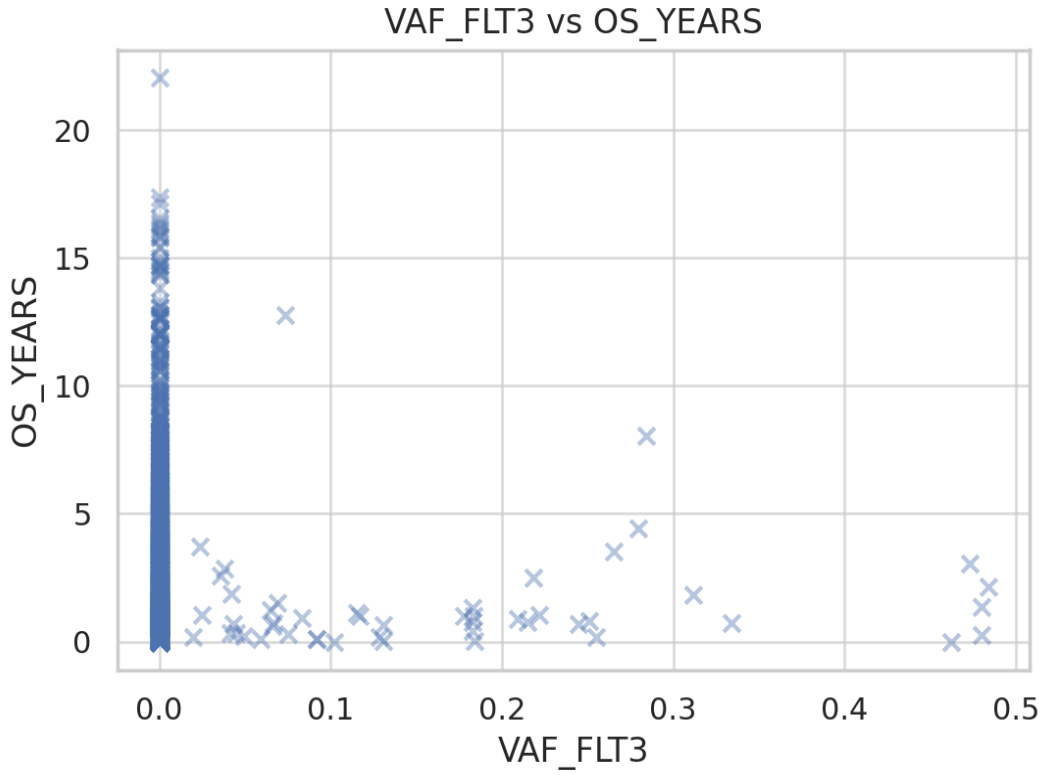
The plot above illustrates the distribution of the feature 'num_translocations' against overall survival in years. Patterns can highlight clinical relevance.

VAF_TP53 vs OS_YEARS



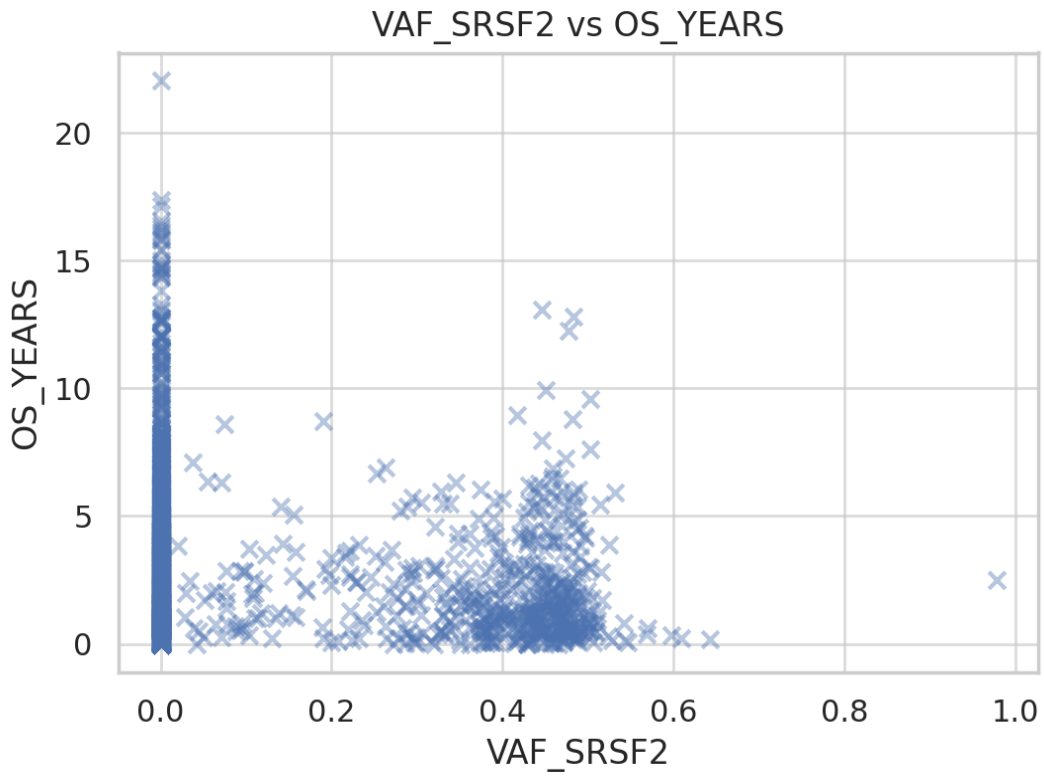
The plot above illustrates the distribution of the feature 'VAF_TP53' against overall survival in years. Patterns can highlight clinical relevance.

VAF_FLT3 vs OS_YEARS



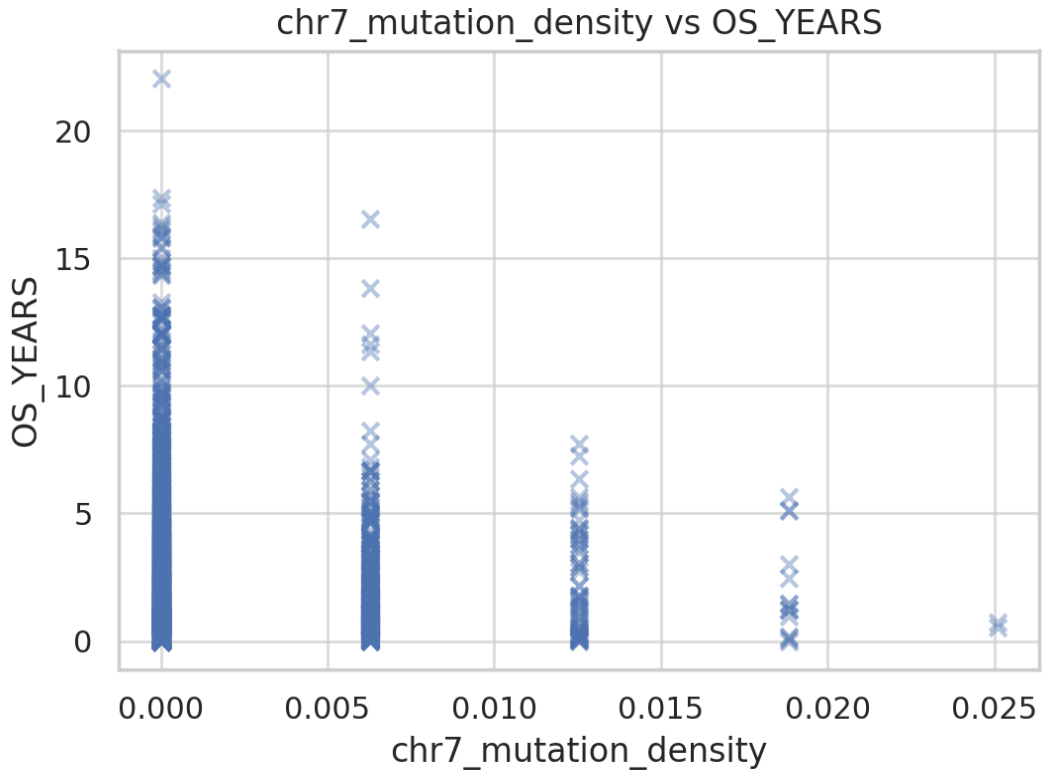
The plot above illustrates the distribution of the feature 'VAF_FLT3' against overall survival in years. Patterns can highlight clinical relevance.

VAF_SRSF2 vs OS_YEARS



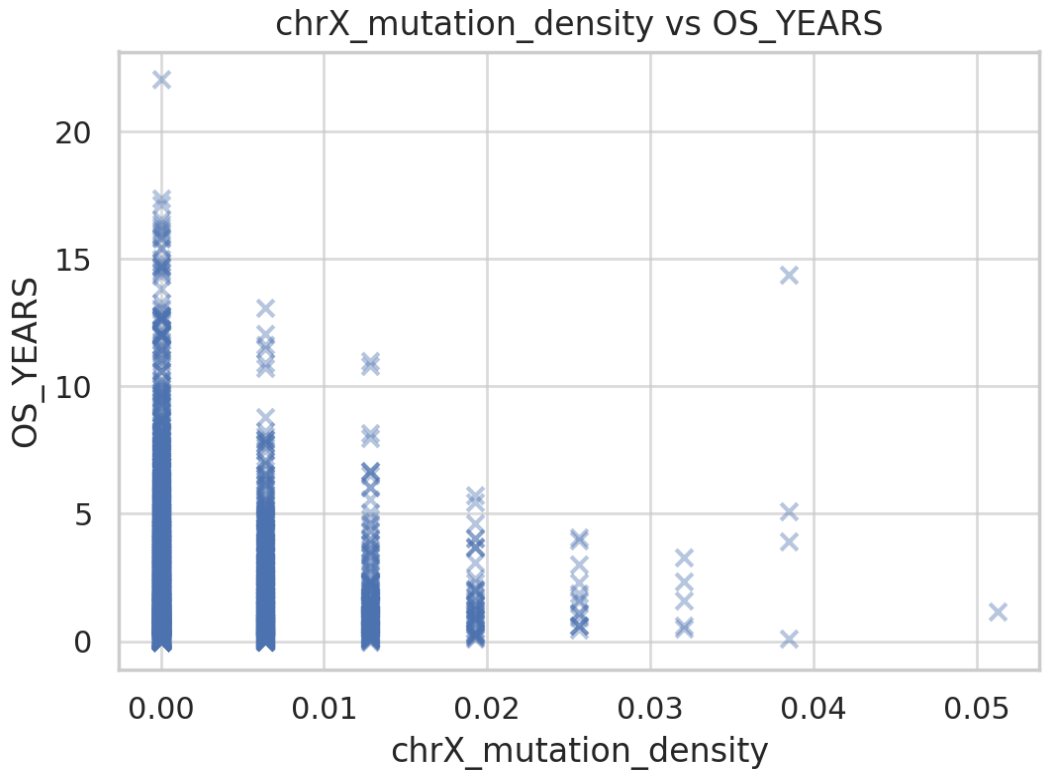
The plot above illustrates the distribution of the feature 'VAF_SRSF2' against overall survival in years. Patterns can highlight clinical relevance.

chr7_mutation_density vs OS_YEARS



The plot above illustrates the distribution of the feature 'chr7_mutation_density' against overall survival in years. Patterns can highlight clinical relevance.

chrX_mutation_density vs OS_YEARS



The plot above illustrates the distribution of the feature 'chrX_mutation_density' against overall survival in years. Patterns can highlight clinical relevance.

Conclusion

In conclusion, this exploratory analysis has highlighted several engineered features with strong clinical relevance and potential predictive value for survival in adult myeloid leukemia. The inclusion of gene-specific VAFs and chromosome-specific mutation densities provides a genomic dimension that complements clinical observations. The derived features have demonstrated acceptable multicollinearity and largely low inter-feature correlation, making them well-suited for integration into survival models such as Extra Survival Trees or Cox Proportional Hazards. These findings will guide the subsequent model development phase to improve personalized prognosis and treatment planning in hematologic oncology.