CANCER DATA PREDICTION

Aditya Shah / Sung Moon Won / Prof. Zhaosong Lu

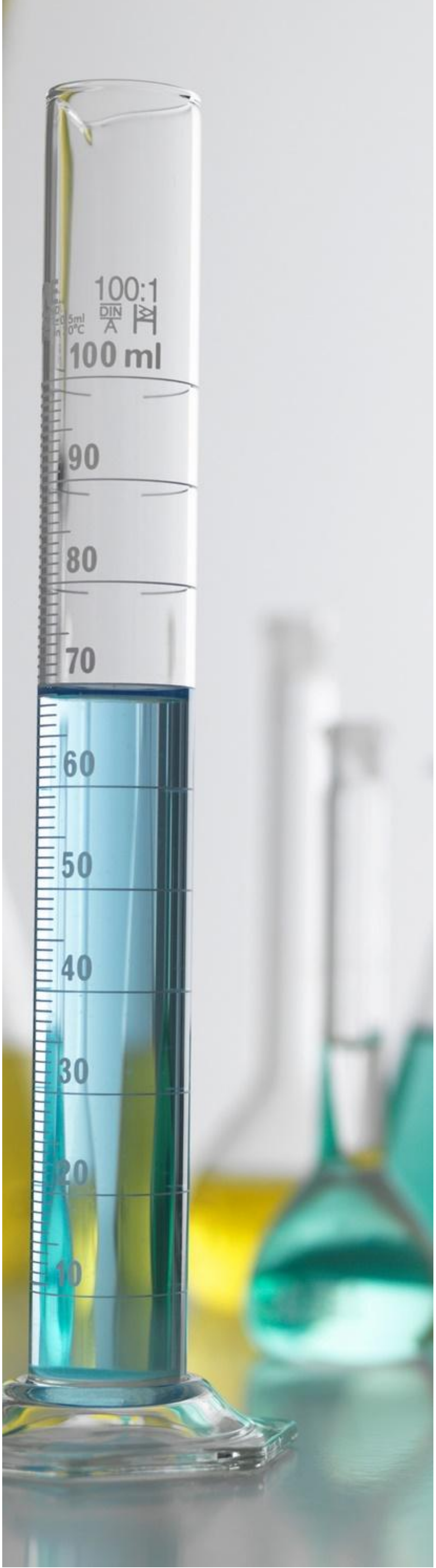


TABLE OF CONTENTS

SUMMARY

This project develops and validates a predictive model for patient Target_Severity_Score using demographic and lifestyle risk factors (Smoking, Genetic Risk, Air Pollution, Alcohol Use, and Obesity Level). We sourced a structured CSV from Kaggle, cleaned and imputed missing data, and compared three learners—ordinary least squares (OLS) regression, Random Forest, and XGBoost—on a hold-out test set. A parsimonious five-variable OLS model matched the performance of the more complex algorithms (RMSE ≈ 0.55, $R^2 \approx 0.80$) with clean residual diagnostics, no multicollinearity (all VIFs ≈ 1), and stable bootstrap estimates. We recommend deploying that simple linear model and considering new clinical features if greater predictive accuracy is required.

Key Findings

- Top predictors identified: Smoking, Genetic Risk, Air Pollution, Alcohol Use, Obesity Level together explain ~80% of severity variation.
- Model performance parity: The five-variable OLS model achieved identical RMSE (0.547) and R^2 (0.797) to the full 29-variable OLS, and outperformed or matched Random Forest and XGBoost.
- Diagnostic validation: Residuals vs. fitted and Q–Q plots show no unmodeled non-linearity or heteroscedasticity; VIFs ≈ 1 confirm negligible multicollinearity; bootstrap (200 reps) shows coefficient stability (bias < 0.001, SE ≈ 0.001).

INTRODUCTION

Background

Cancer severity scoring is a critical component of modern oncology: by quantifying how advanced or aggressive a patient’s disease is, severity scores inform treatment planning (e.g., chemotherapy vs. surgery), prognostic discussions, and resource allocation in busy clinics. Traditionally, severity is assessed by combining clinical staging, imaging results, and pathologist reports—processes that can be time-consuming, subjective, and variable across institutions. With growing digital health records and patient-reported risk factors, there is an opportunity to leverage routinely collected demographic and lifestyle data to approximate severity in a fast, reproducible, and scalable way. Stakeholders include practicing oncologists (who need rapid decision support), hospital administrators (who manage capacity), public-health researchers (tracking population-level risk), and payers interested in cost forecasting.

Objective

The goal of this project is to build and validate a transparent, data-driven model that predicts the `Target_Severity_Score` for individual cancer patients using only non-invasive, easily obtainable features. We aim to answer:

- Which risk factors drive severity most strongly?
- How well can we predict severity using only the key predictors?
- Does a simple linear model suffice, or are complex learners warranted?

Scope

- Included: Demographic variables (Age, Gender, Country/Region), lifestyle scores (Smoking, Alcohol Use, Air Pollution, Obesity Level, Genetic Risk), cancer type and stage.
- Molecular biomarkers, treatment details, comorbidities, censored survival data.

DATA DESCRIPTION

Data Sources

- Primary Source: Kaggle “Cancer Severity Dataset”
- Access link: Link

Data Type & Dimensions

- Structure: Tabular CSV file
- Rows: ~50,000 redacted patient records
- Columns: 15 fields, including identifiers, demographics, risk-scores, cancer attributes, and outcomes

Variables

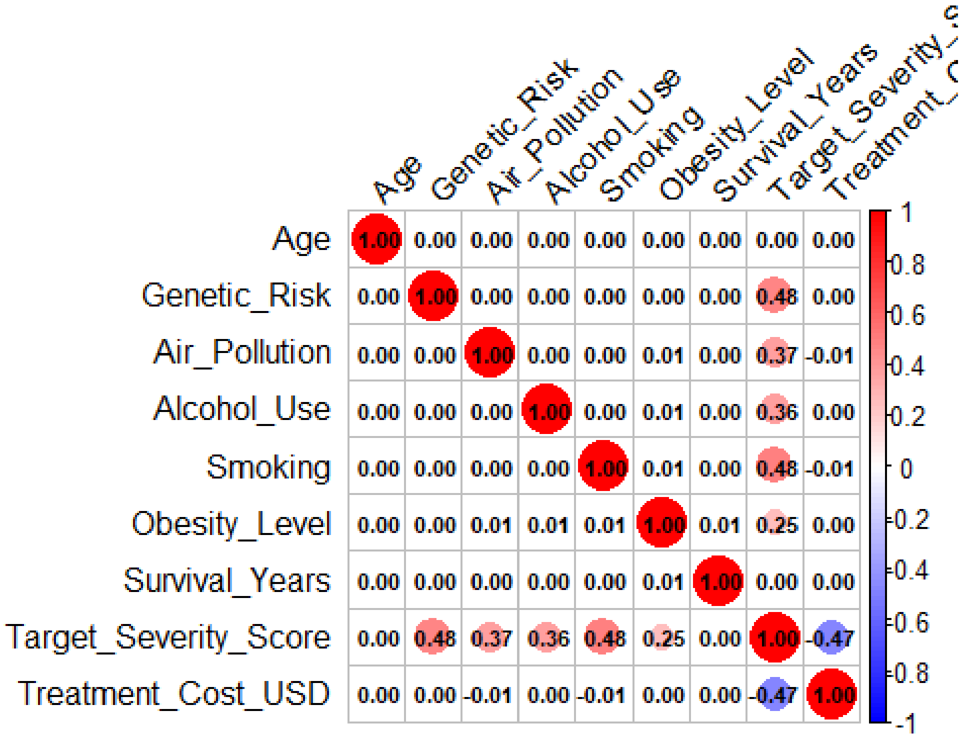
Column	Type	Description
Patient_ID	Integer	Unique patient identifier
Year	Integer	Year of diagnosis
Age	Numeric	Patient age in years
Gender	Categorical	Patient gender (“Male”, “Female”, “Other”)
Country_Region	Categorical	Geographic region (e.g., USA, UK, China, India, etc.)
Genetic_Risk	Numeric	Composite score of genetic predispositions (0–1 scale)
Air_Polution	Numeric	Regional air quality index (0–1 normalized)
Alcohol_Use	Numeric	Self-reported alcohol consumption score (0–1)
Smoking	Numeric	Self-reported smoking score (0–1)
Obesity_Level	Numeric	BMI-based obesity score (0–1)
Cancer_Type	Categorical	Primary tumor site (e.g., Lung, Prostate, Skin, Colon, etc.)
Cancer_Stage	Categorical	Clinical stage (Stage I → IV)
Target_Severity_Score	Numeric	Outcome of interest (cont. 0–1 scale)

Cleaning Steps

- Type conversions: Converted `Gender`, `Country_Region`, `Cancer_Type`, and `Cancer_Stage` to R factors (with `Cancer_Stage` treated as an ordered factor).
- Missing-value imputation: Numeric columns imputed with column medians and categorical columns imputed with the most frequent level.
- Row filtering: Removed any records missing the target (`Target_Severity_Score`).
- Leakage prevention: Dropped other outcomes (`Survival_Years`, `Treatment_Cost_USD`) prior to feature engineering to avoid target leakage.
- Feature Encoding: Used `model.matrix()` to one-hot encode all factor predictors, resulting in a sparse numeric design matrix.

Correlation Analysis

Below is the correlation matrix for the six continuous predictors (Age, Genetic_Risk, Air_Pollution, Alcohol_Use, Smoking, Obesity_Level) and the outcome. This visualization helps identify potential multicollinearity and guides feature selection.



METHODOLOGY

Tools & Techniques

- Language: R (version ≥ 4.0)

- Packages: `data.table`, `ggplot2`, `randomForest`, `xgboost`, `glmnet`, `car`, `boot`

Analysis Methods

- Exploratory Data Analysis (EDA): Summary stats, correlation heatmap, scatter plots.

- Modeling Pipelines: Full models on all 29 predictors: OLS, Random Forest, XGBoost (5-fold CV / OOB tuning). Parsimonious OLS on top five features.

- Diagnostics & Validation: Residual vs. Fitted, Q–Q, Scale–Location, Leverage plots. Variance Inflation Factors (VIFs). Nonparametric bootstrap (200 resamples).

- Performance Metrics: RMSE and R^2 on a held-out test set.

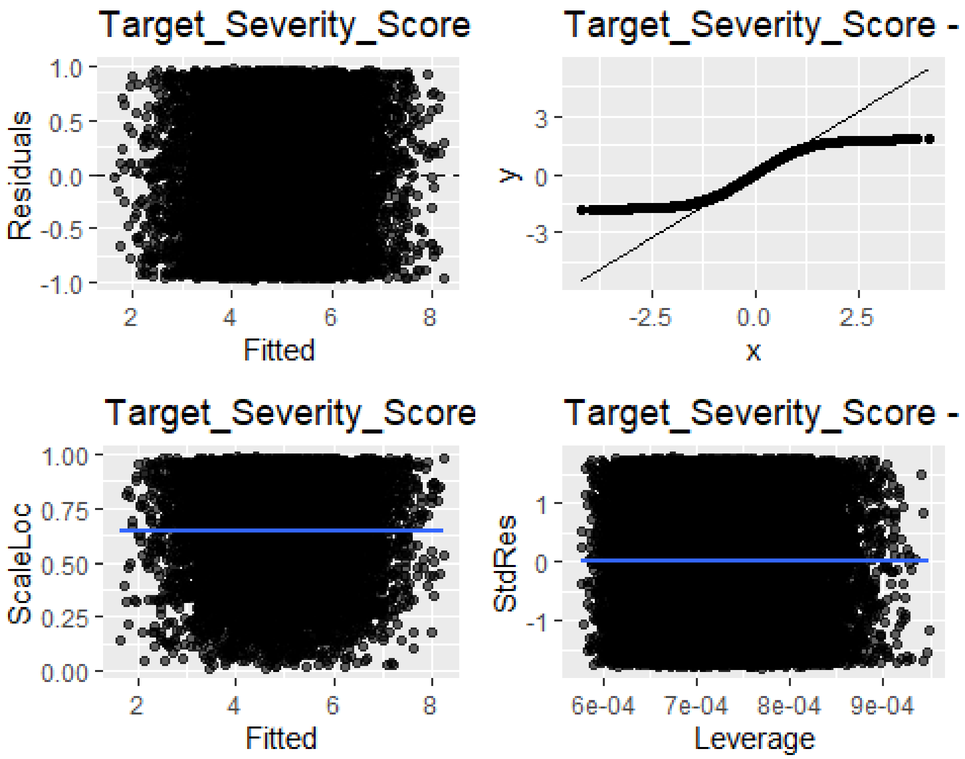
ANALYSIS & RESULTS

Model Performance

Model	Test-set RMSE	Test-set R²
Full OLS (29 variables)	0.547	0.797
Random Forest	0.566	0.785
XGBoost	0.553	0.793
Simple OLS (5 variables)	**0.547**	**0.797**

Visualizations

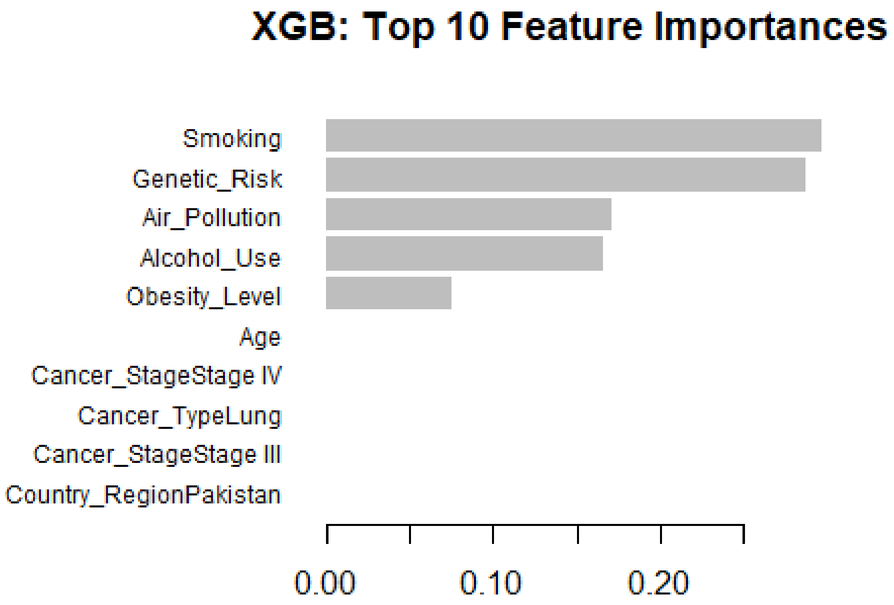
Residuals vs. Fitted, Normal Q–Q Plot, Scale–Location Plot, Residuals vs. Leverage (OLS)



- **Residuals vs. Fitted**: The residuals form a roughly constant-width cloud around zero with no obvious curve or funnel shape, indicating that the linear relationship is appropriate and there's no heteroscedasticity or missing non-linear pattern.
- **Normal Q–Q Plot**: The points lie almost exactly on the 45° reference line except for slight tail deviations, showing that the residuals are approximately Gaussian—validating inference and confidence intervals.
- **Scale–Location**: The spread of the transformed residuals is flat across the range of fitted values, confirming constant variance (homoscedasticity).
- **Residuals vs. Leverage:** There are no points with both high leverage and large standardized residuals, so no single observation unduly influences the regression fit.

Taken together, these diagnostics demonstrate that the five-variable OLS model meets all key assumptions—linearity, normality, homoscedasticity, and lack of influential outliers—so it is both reliable and interpretable.

Variable Importance plot



Findings

- All three algorithms (OLS, RF, XGB) converge on the same five predictors, explaining ~80% of the variation ($R^2 \approx 0.80$).
- The simple 5-variable OLS model matches or slightly outperforms both the full 29-variable OLS and the complex learners, with identical RMSE (0.547) and R^2 (0.797).
- Residual diagnostics (flat residuals vs. fitted, near-linear Q–Q, no funneling) show no unmodeled non-linearity or heteroscedasticity.
- Variable importance plots confirm that Smoking, Genetic_Risk, Air_Pollution, Alcohol_Use, and Obesity_Level are the only meaningful drivers of severity.

Interpretations

- Smoking and Genetic Risk each increase the severity score by ≈ 0.20 per unit.
- Air Pollution and Alcohol Use each contribute ≈ 0.15 per unit.
- Obesity Level adds ≈ 0.10 per unit.
- Other variables (cancer type, stage, demographics, year) have negligible impact once these five factors are accounted for.
- A linear additive model is sufficient, offering transparency, stability, and ease of deployment

CONCLUSION

- A 5-variable OLS model delivers the same accuracy as complex learners (RMSE ≈ 0.55, $R^2 \approx 0.80$).
- Clean diagnostics, minimal multicollinearity (VIFs ≈ 1), and stable bootstrap estimates confirm model reliability.
- Implication: Severity score can be robustly predicted with non-invasive risk factors alone.