Melanoma

Tumor Size Prediction

Michael Son

Melanoma Facts

- Evolves from melanocytes
- Most dangerous type of skin cancer
- Prevalent among individuals with low melanin level
- Appears in various shapes/sizes/colors

Problem

- **Diagnosis/Prognosis Difficult**
- **High Prevalence**

Solution

- **Size as an early indicator**

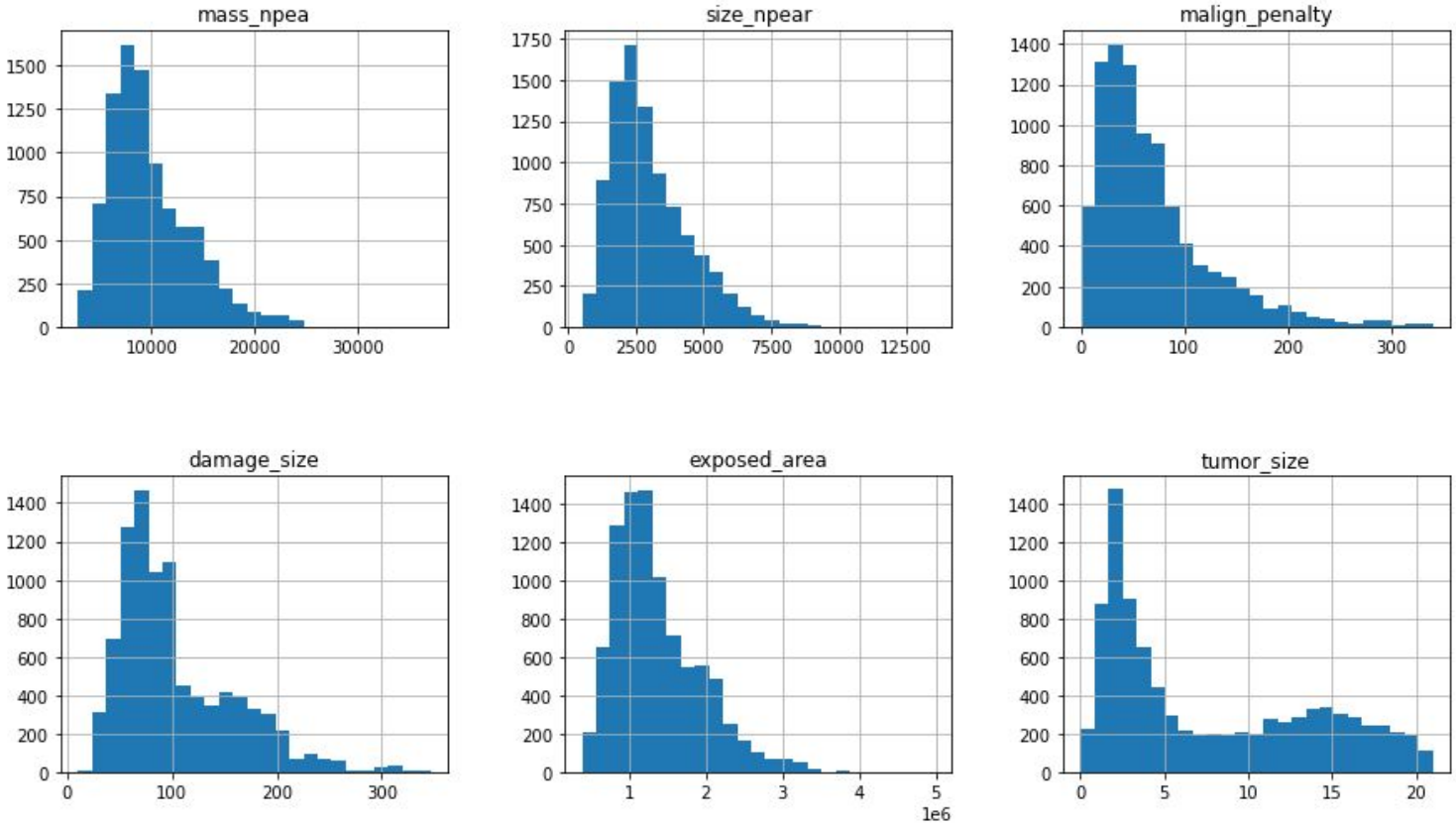
Stakeholders

Data Information

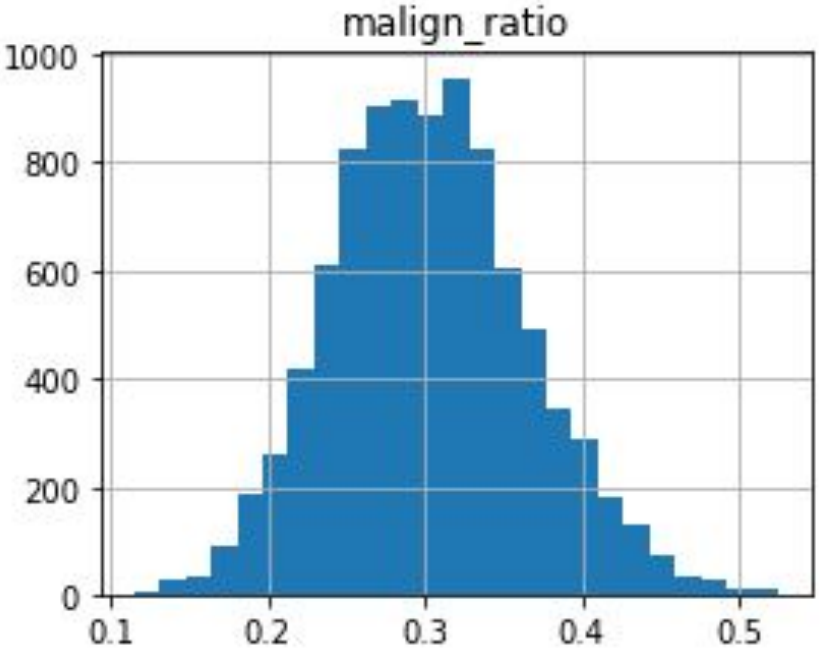
- Source: Machine Hack: Melanoma Tumor Size Prediction
- Training Set + Test Set
- **Tumor Size** and relevant attributes
- Number of entries: **9,146**
- Number of features: **10**
- Number of null values: **0**

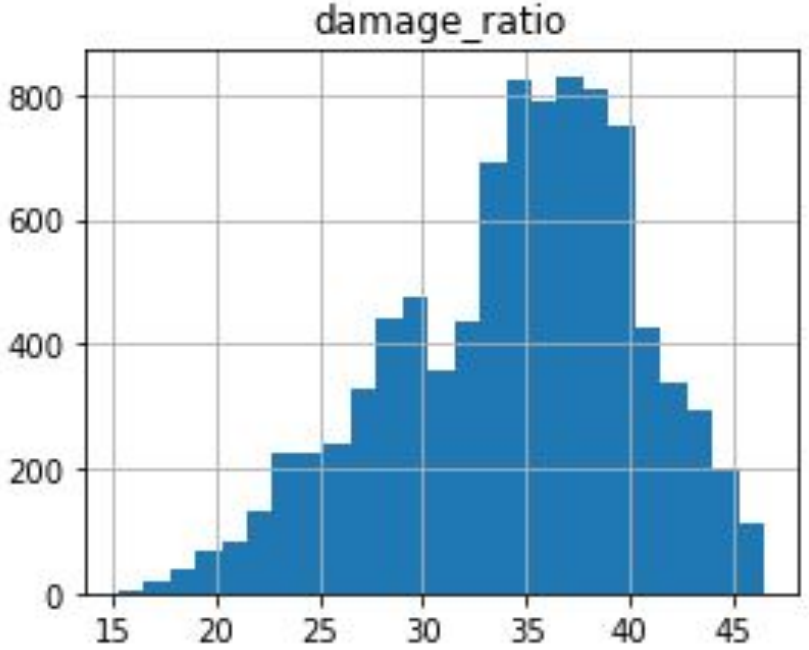
Features

- Mass of the Area
- Size of the Area
- Normal/Malign Surface Ratio
- Damage Size
- Total Area Exposed to the Tumor
- Standard Deviation of Malign Skin Measurements
- Error in Malign Skin Measurements
- Penalty Imposed due to Measurement Error
- Damage/Total Spread Ratio
- Tumor Size

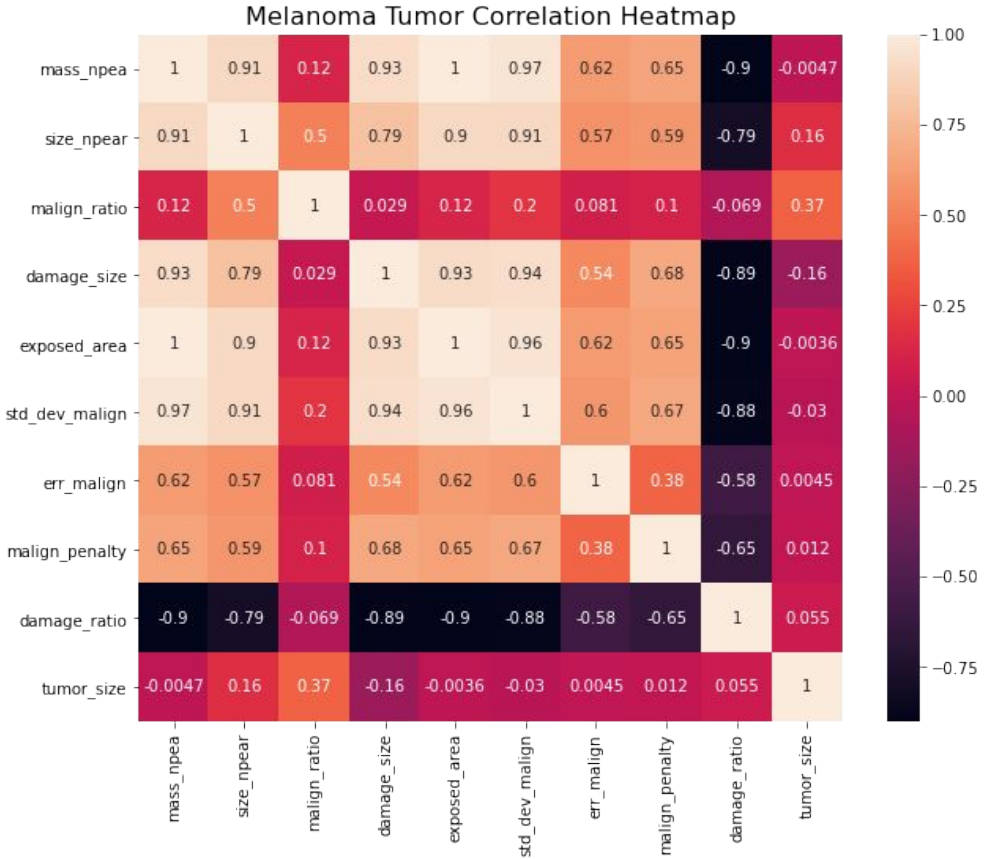


Apparent correlations due to the inherent proportionality between mass/volume





- Mean of **malign_ratio** (normal/malign) ≈ 0.3 → malignancy prevalent in the dataset
- Left-skewed distribution of **damage_ratio** supports this notion



- Notable **tumor_size** correlations
  - **size_npear**
  - **malign_ratio**
  - **damage_size**

- Other notable correlations
  - **mass_npea/damage_ratio**
  - **size_npear/malign_ratio**

Modeling Overview

- Supervised Learning
- Regression
- Machine Learning Tools: Scikit-Learn, Keras

Modeling Procedure

I. Data Preprocessing

1. Training Validation Split (70%: 30%)
2. Feature Standardization

II. Randomized Search

- cv = 3
- n_iter = 50
- scoring = “neg_mean_squared_error”

III. Training with Tuned Parameters

IV. Performance Evaluation

- Evaluation Metric: R2, MSE
- Separate Test Dataset

Techniques/Algorithms Used

Hyperparameter Tuning Technique

- Randomized Search from Scikit-Learn

Regression Algorithms

1) Multiple Linear Regression
2) Random Forest
3) Support Vector Machine
4) Multi-Layer Perceptron
5) Keras Regression

Model Comparison

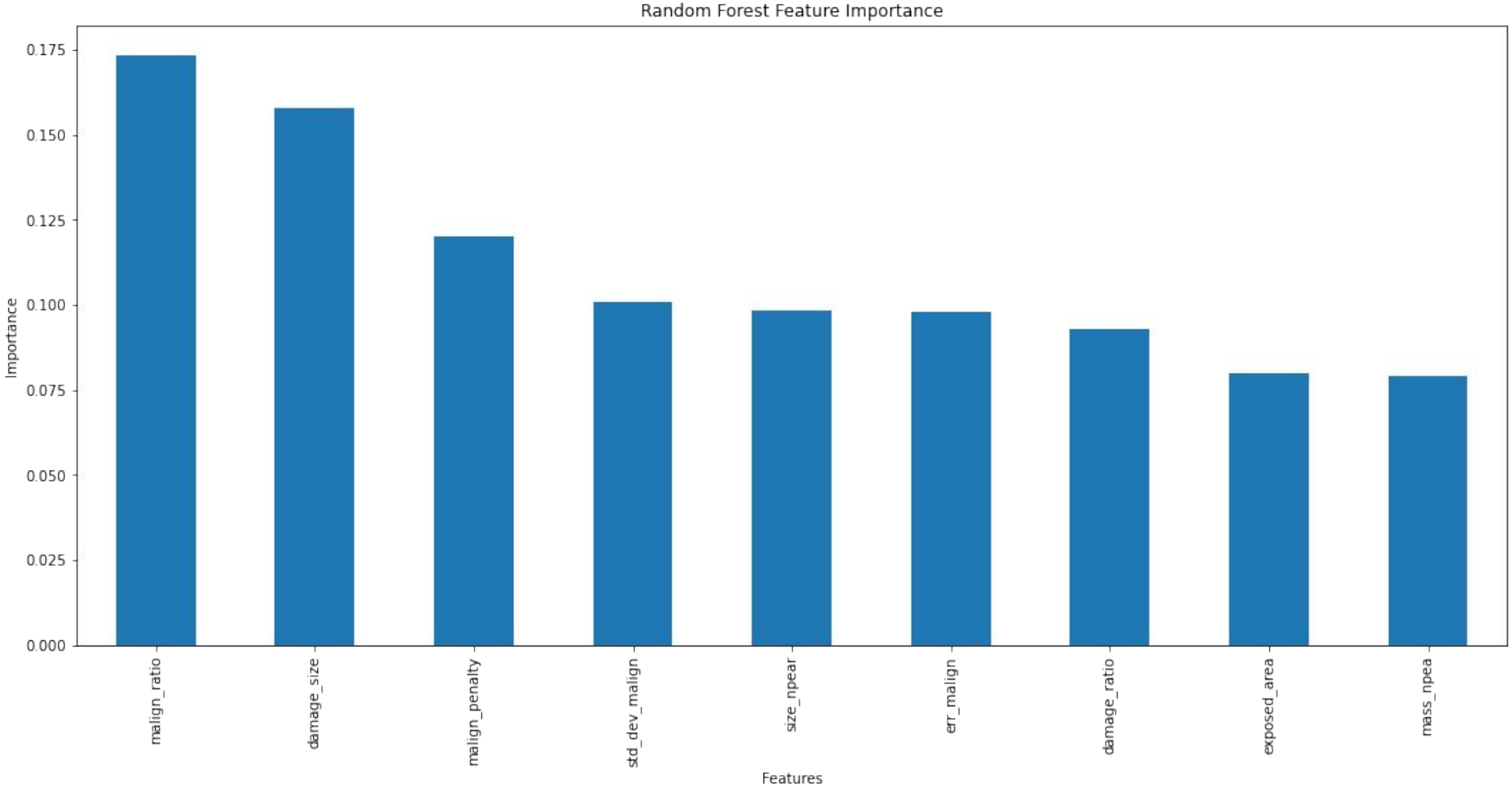
Model	MSE	R2
Multiple Linear Regression	26.43	0.29
Random Forest	16.21	0.57
Support Vector Machine	21.53	0.42
Multi-Layer Perceptron	29.62	0.21
Keras Regression	18.69	-

High Performance Models: Random Forest, Keras Regression

Performance on Test Set

Model	MSE	R2
Random Forest	8.25	0.23
Keras Regression	12.12	-

Best Model: Random Forest


Random Forest Feature Importance
Importance
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
malign_ratio
damage_size
malign_penalty
std_dev_malign
size_npear
err_malign
damage_ratio
exposed_area
mass_npea
Features

Assumptions/Limitations

- Primary Assumption
  - All measurements obtained with sufficient accuracy/precision

- Limitations
  - Insufficient number of entries for training set compared to those of test set (mere 9146, while 36584 for test set)
  - Low number of features

Conclusion

- All features used
- Best model: Random Forest
- Primary features of importance: **malign_ratio**, **damage_size**, **malign_penalty**
- Prospective improvement
  - Larger volume of data for training set
  - Hyperparameter tuning with Grid Search
  - Further experiment with neural network/deep learning models