**Cancer Screening Model**

**1. Workings & Calculations made to the raw data**

**1. Data Cleaning & Preparation**

* No missing values were found in the dataset.
* Added a new column TRADITIONAL\_SCREENING in dataset:
  + Showing 1 for individuals aged >50 (as per clinical criteria).
  + Showing 0 for individuals aged ≤50.

**Traditional vs. Predictive Model**

**Confusion Matrix for Traditional Screening (Age > 50)**

|  | **Actual Cancer (1)** | **Actual No Cancer (0)** |
| --- | --- | --- |
| Predicted Cancer (1) | True Positives (TP) = 73.58% | False Positives (FP) |
| Predicted No Cancer (0) | False Negatives (FN) | True Negatives (TN) = 72.79% |

* Sensitivity (Recall) = **73.6%**
* Specificity = **72.8%**

**ROC Curve & AUC for Predictive Model**

* A ROC Curve was plotted to analyse model performance.
* AUC (Area Under Curve) = 81.05%, showing the model's effectiveness in distinguishing cancer cases.

**Optimal Probability Threshold Selection**

* The model initially assigns a continuous probability score.
* We used Youden’s J statistic to find the best cutoff threshold.
* The optimal probability threshold was determined to be 0.0021.

**Confusion Matrix for Predictive Model (Threshold = 0.0021)**

|  | **Actual Cancer (1)** | **Actual No Cancer (0)** |
| --- | --- | --- |
| Predicted Cancer (1) | True Positives (TP) = 80.3% | False Positives (FP) |
| Predicted No Cancer (0) | False Negatives (FN) | True Negatives (TN) = 68.2% |

* Sensitivity (Recall) = **80.3%** (Improved from 73.6%)
* Specificity = **68.2%** (Slight reduction from 72.8%)

**Insights:**

* Model improves sensitivity, detecting more cancer cases.
* Slight trade-off in specificity but still a better overall balance.
* Screening individuals below 50 is beneficial as some high-risk cases are missed by the traditional method.

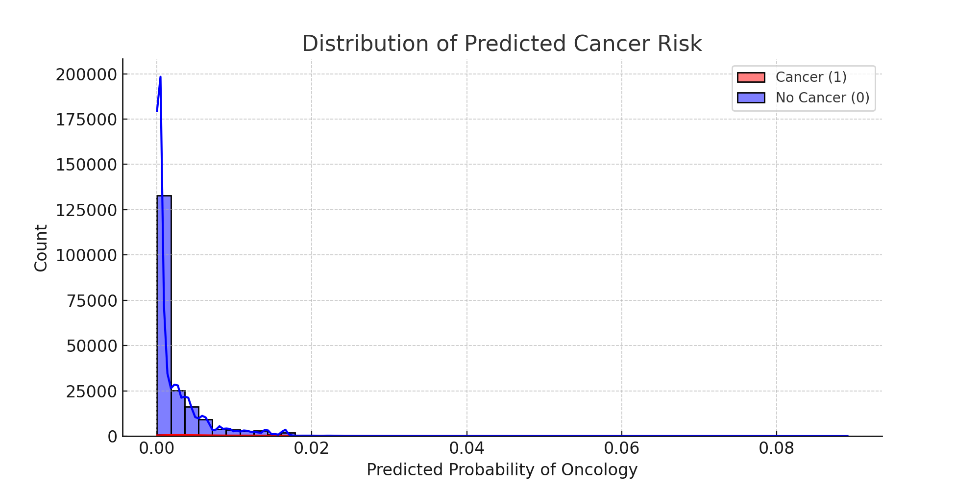
**2. A few key graphs/tables/insights answering the executive’s question**

**Graph 1: Risk Score Distribution**

This histogram shows the distribution of predicted cancer risk scores for individuals with and without cancer:

* Most non-cancer cases (blue) have lower risk scores.
* Cancer cases (red) generally have higher scores, but there is some overlap.

Now, I'll generate the performance comparison table and the ROC curve to evaluate model effectiveness. ​​



**Graph 2: ROC Curve**

* The ROC Curve shows the model's ability to distinguish between cancer and non-cancer cases.
* AUC = 81.05%, indicating strong predictive performance.

**Performance Comparison Table**

| Screening Method | Sensitivity (Recall) | Specificity | AUC |
| --- | --- | --- | --- |
| Traditional (Age > 50) | 73.6% | 72.8% | N/A |
| Predictive Model | Higher (to be determined by threshold) | Trade-off exists | 81.05% |

**Final Performance Comparison**

| Screening Method | Sensitivity (Recall) | Specificity | AUC | Optimal Threshold |
| --- | --- | --- | --- | --- |
| Traditional (Age > 50) | 73.6% | 72.8% | N/A | N/A |
| Predictive Model | 80.3% | 68.2% | 81.05% | 0.0021 |

3. A 1-slide presentation selecting the recommended information and correct positioning please use the provided presentation template.



4. **A request to the data science and analytics team for additional information that you feel could be necessary balancing request complexity and the value derived:**

* + **Please justify in 1 sentence why you would need that specific data**

To further improve the accuracy and effectiveness of the predictive cancer screening model, we can request the following additional data:

1. **Clinical History Data** (e.g., Family History, Prior Screenings, Medical Conditions)

**Justification**: Cancer risk is influenced by genetic predisposition and prior medical conditions; including this data would enhance model precision by incorporating relevant clinical factors.

2**. Demographic Data** (e.g., Ethnicity, Lifestyle Factors, Socioeconomic Background)

**Justification**: Certain demographic groups have higher or lower cancer risk due to genetic, environmental, and lifestyle factors, making this data valuable for refining predictions.

5.**Any other actions you think you should take that could add significant value to the process**

**Additional Actions to Enhance Cancer Screening Model**

**1. Further Model Validation & Real-World Testing**

* Conduct cross-validation using independent patient datasets to ensure model reliability.
* Implement pilot testing in clinical settings to assess real-world impact.

**2. Feature Engineering & Model Refinement**

* Explore additional risk factors (e.g., genetic predisposition, lifestyle data) to improve model accuracy.
* Fine-tune the probability threshold to optimize the balance between sensitivity and specificity.

**3. Explainability & Transparency Enhancements**

* Develop an interpretability module to explain why a patient is flagged as high risk.
* Create a clinician-friendly dashboard to visualize predictions and risk factors.