Otabor, Clark, Ankem, Howard

Michael Shuckers & Ilieva Ageenko

DTSC 2302-002

Project 1 Report

(A)

The data we've analyzed contains 6 metrics from 48,842 individuals, recorded by the U.S. Census. Using the data about the individuals, a Support Vector Machine model was developed to predict whether an individual earned more than $50,000, or not. The predictors are listed below:

- Age (Age in years)
- Education-num (Years of education)
- CapitalGain (Amount of capital gains in the last year)
- CapitalLoss (Amount of capital loss in the last year)
- Hours-Per-Week (Hours per week worked)
- SalaryClass (Whether an individual makes more than 50k annually)

The SalaryClass indicates whether an individual makes more than $50,000 dollars annually or not, measured in the U.S. dollar. The model we chose for this portion of the project was the Support Vector Machine model with a radial basis function kernel. With this model, we aimed to identify patterns and make accurate predictions on income based on the predictors. The model was trained and tested on the data, making it able to classify individuals based on the probability of earning above or below the amount. We also tried other classification models, like

logistic regression, linear discriminant analysis, and support vector machines with a linear kernel, but the model we chose outperformed the others with the highest accuracy (82.99%), a strong precision (77.50%), and a solid F1 score (51.57%). Despite the chosen model's recall of 38.65%, its combination of accuracy, precision, and overall performance makes it the most useful model for predictions.

In Appendix A, we can see that the SVM-RBF model has the highest accuracy with the blue bar, followed closely by logistic regression and SVM-Linear. This shows that the SVM-RBF model made the most correct predictions. Precision was fairly high across all models, with SVM-Linear performing the best with the green bar. This means that both the SVM models were good at predicting high income individuals. Recall is where the models differ significantly. SVM-Linear has the highest recall, while LDA(Linear Discriminant Analysis) and logistic regression lag behind, with the yellow bar. High recall is important when missing positive cases is a big problem (ex, missing fraud cases). SVM-Linear and SVM-RBF are the strongest contenders for the F1 score. The F1 score gives a single number that captures both precision and recall. It's useful for seeing overall performance when both false positives and false negatives matter.

While the SVM-Linear model performed well, there are important reasons why SVM-RBF is the better choice for this portion of the project. First, the radial basis kernel is known to better capture complex, non-linear relationships between predictors compared to the linear kernel, which assumes a straight-line relationship. Income classification can be influenced by multiple factors that can interact with each other (e.g., education, capital gains/losses, age), which are unlikely to have a simple linear relationship. SVM-RBF is better suited to modeling these interactions, thereby leading to a more adaptable classification method. Second, SVM-RBF

achieved the highest accuracy, meaning overall it makes fewer classification errors. Third, SVM-RBF is better than the other models at differentiating classes where there may be overlap. For instance, there may be overlap between income groups. SVM-Linear might struggle if the data is not perfectly separable by a straight line. If the relationships in the data were truly linearly separable, SVM-Linear would be a strong choice, but given the complexity of the factors of income we were given, SVM-RBF remains the best choice for capturing the nuances.

According to the Support Vector Machine (radial basis kernel) model, an individual who is 39 years old, with 10 years of education, 0 capital gains and losses, and works 40 hours per week is predicted not to make more than $50,000 annually. An individual who is 45 years old, with 14 years of education, 2000 capital gains, 0 capital losses, and works 35 hours per week, is also predicted not to make more than $50,000 annually.

(B)

In this portion of the project, we analyzed a dataset of 30,000 credit card holders to predict whether an individual will default on their next payment. The dataset includes features such as credit limit (LIMIT_BAL), age (AGE), gender (SEX), education level (EDUCATION), marital status (MARRIAGE), payment history (PAY_0 to PAY_6), bill amounts (BILL_AMT1 to BILL_AMT6), and payment amounts (PAY_AMT1 to PAY_AMT6). The target variable, DEFAULT, indicates whether the individual defaulted last month. By analyzing these features, we aimed to predict future defaults and better understand financial risks.

We compared two models: Decision Tree and RBF Kernel SVM. The Decision Tree model, after pre-pruning with a maximum depth of 5, achieved an accuracy of 82.08%, with a cross-validation score of 81.68%. The model was stable and interpretable. The RBF Kernel SVM slightly outperformed the Decision Tree in test accuracy (82.17%), but its lower cross-validation score of 77.88% indicated overfitting, suggesting poorer generalizability.

The Decision Tree was chosen as the final model for its stability, interpretability, and strong accuracy. Although the RBF Kernel SVM had a slightly higher accuracy, its tendency to overfit made the Decision Tree the more reliable choice for understanding the decision-making process.

Visuals for the Decision Tree's performance and decision-making process are provided in **Appendix B**, which includes both the confusion matrix and the Decision Tree plot.

(C)

The model was made to predict the probability of O-ring damage based on the air temperature during launch. In appendix C the model shows at 50° the probability of O-ring damage is 0.1536, at 40° the probability is 0.5869, at 30° the probability is 0.9176. The decision boundary is marked at 0.5, temperatures below this threshold are more likely to result in O-ring damage. The model highlights that lower temperatures significantly increase the risk of O-ring damage. This finding aligns with historical observations of O-ring failures in cold conditions. The graph provides a clear visualization of the logistic regression curve, emphasizing the critical temperatures where the probability of damage becomes substantial.

Ethics Questions:

2. From an ethical perspective, what does the phrase "data scientists should never surrender their professional judgment to organizational pressure" mean to you? Can you think of consequences when this principle is violated?

To me, the phrase "data scientists should never surrender their professional judgment to organizational pressure" means standing with your ethical values at work even when you run into situations where it would be better or more convenient for you to throw

away your values. It's really important to be objective and authentic when you are analyzing and interpreting data so that your conclusion is honestly represented and not tampered with in any way by stakeholders to push their narrative or to meet a deadline. This phrase involves you prioritizing things like morality and fairness over an organization's goals that may not align with these values.

When you violate these principles, like what the engineers did with the Challenger case where they downplayed warnings, you end up with a tragedy that can cost lives. In other cases, like that from a data scientist's context, surrendering your professional judgment could lead to things like biased algorithms that further push existing biases, like what was seen with Amazon's algorithm, which was biased against women. In the long term, this can lead to a loss of trust and reputability of an organization.

3. What ethical issues arise from asking data scientists to prove the launch is safe rather than prove it is unsafe? How does this relate to confirmation bias and responsibility in risk communication?

Asking data scientists to prove a launch is safe rather than proving that the launch is unsafe leads them to try to ignore the potential dangers and focus mainly on the positives rather than the negatives since the question can subtly encourage overlooking or downplaying uncertainties and data gaps because the data scientists are pressured into coming to a conclusion that proves the safety of the launch. When you switch the question around and ask a data scientist to prove the launch is unsafe, they look at evidence-based reasons for concerns. This ensures all potential flaws are carefully

examined, which is very important during a rocket launch where everything has to be perfect.

This relates to confirmation bias and responsibility in risk communication because when people get tasked with providing safety, they can unconsciously look more into data that prove safety and dismiss data that shows risk. Your responsibility would be to come to a safe verdict rather than look into potential flaws. This was seen in the case of Challenger, where the engineers presented data that showed that the O-ring had problems at lower temperatures, but instead of looking at this, management looked at how there wasn't enough evidence to prove that this was that big of a problem. This all comes to show how important it is to convey potential risks and to avoid folding to organizational pressure to come to a conclusion that supports their narrative.

4. How reliable or "proven" should technology be before it's used in high-risk missions like spaceflight? What ethical considerations guide that threshold?

The threshold should be very high when it comes to the reliability of the technology used in space flight since many factors need to go right for a rocket launch to be successful; engineers should aim to be as close to perfect as possible. The key ethical principles that are guiding this threshold are honesty, to not harm, and empathy, which in this scenario is the astronauts who are risking their lives. Engineers should be empathetic to the astronauts and prioritize their safety over the demands of their organization.

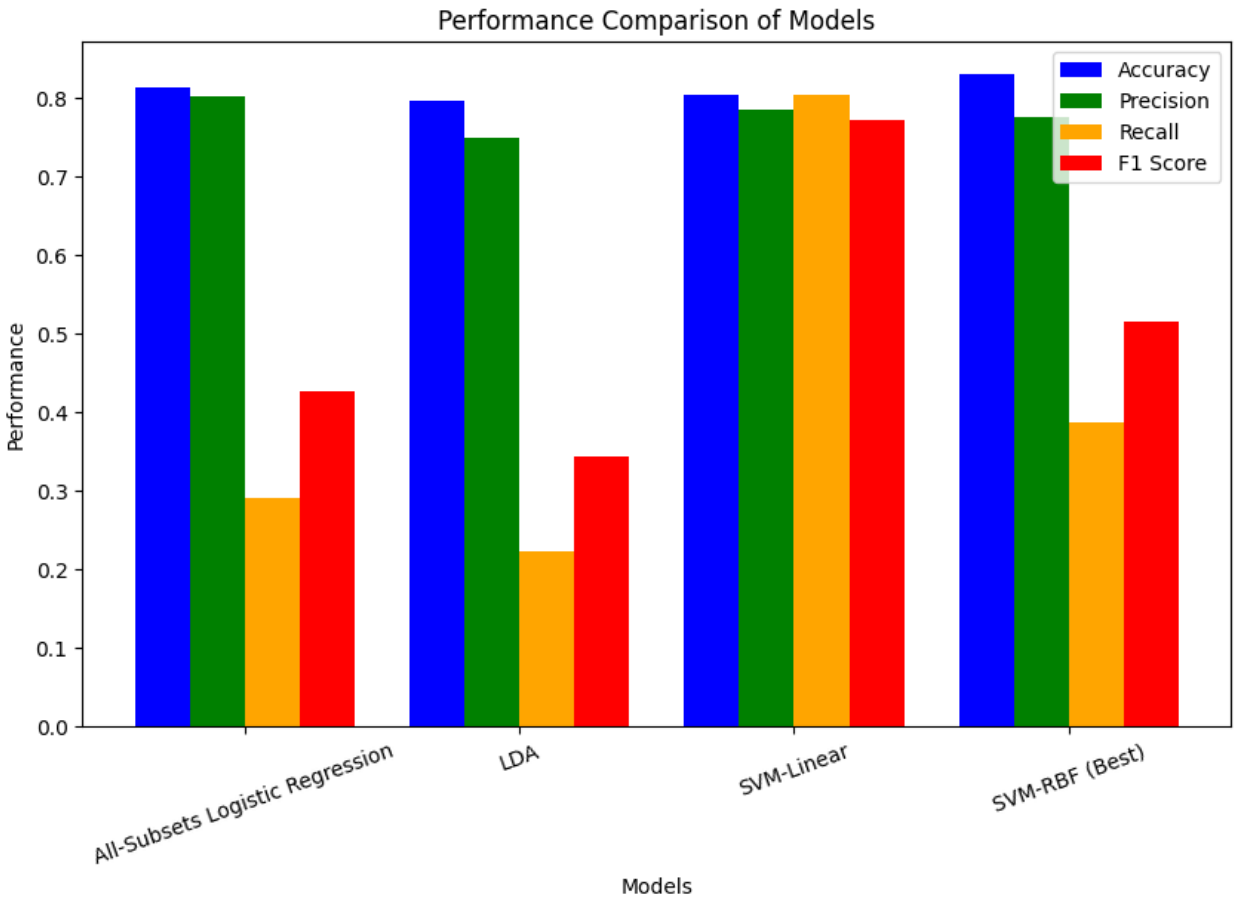
The threshold is guided by a balance of prioritizing the goals of the organization with the safety of the astronauts. We should try to avoid risk as much as reasonably

possible. The Challenger disaster highlights what happens when you break this balance and the engineers have to overlook certain problems because of deadlines. Technology should be proven to the highest reasonable degree, and there should be a clear understanding of uncertainties.

Appendix

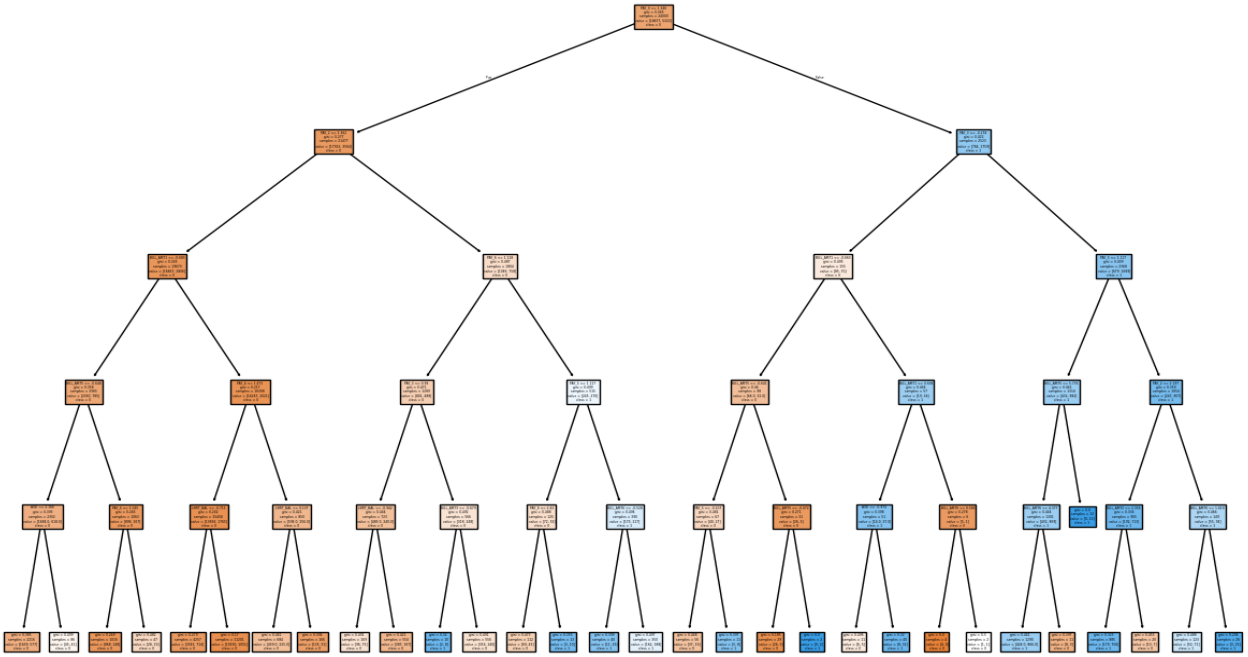
Appendix A

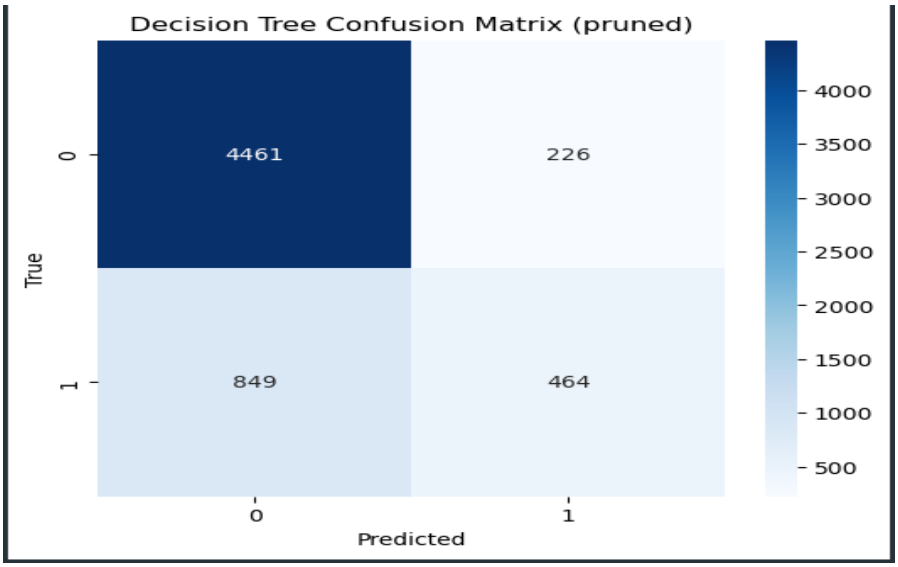
A bar chart visualization that compares the performance of four models that were tested, including the SVM-RBF(Radial Basis Function) model we chose.



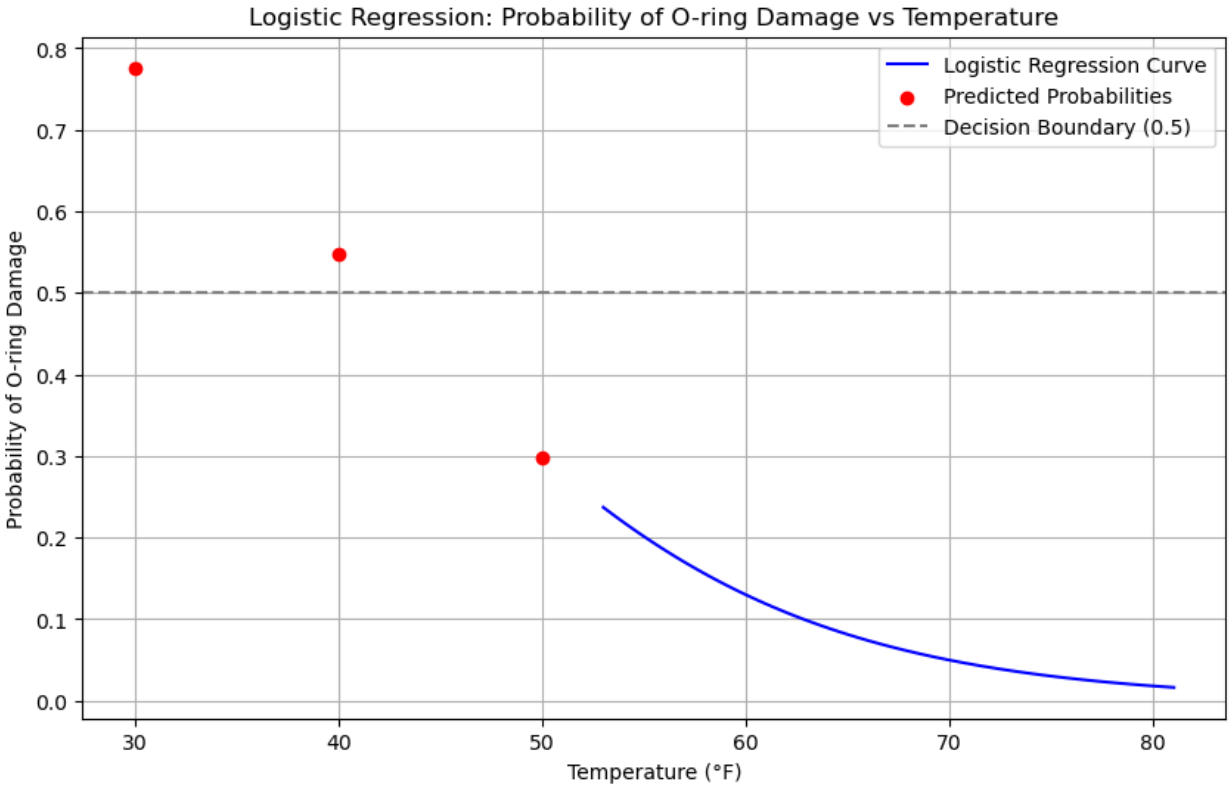
Appendix B

Confusion matrix and Decision Tree plot. These visuals show the model's performance and its decision-making process.





Appendix C


Logistic Regression: Probability of O-ring Damage vs Temperature
Logistic Regression Curve
Predicted Probabilities
Decision Boundary (0.5)
Probability of O-ring Damage
Temperature (°F)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
30
40
50
60
70
80