Diabetes Risk Prediction using Machine Learning

Francisco Montero

Project Overview

- The problem area of the project is reducing the risk of diabetes by identifying factors that lead to the disease and use that information to formulate prevention methods.

- The proposed solution is to develop machine learning models to identify the risk of diabetes using Logistic Regression, Decision Trees, and Random Forests.

- The expected impact of the project includes growing a healthier community, better health education to any affected communities, and healthcare workers can focus on developing resources for those who are at risk.

Data and Preprocessing

- Analysis includes a dataset of females from the Gila River Indian Community in Arizona.
- Key Variables: Age, Pregnancies, Glucose (mg/dl), Blood Pressure (mmHg), Skin Thickness (mm), Insulin, Body Mass Index (kg/m^2), Diabetes Pedigree Function, and Outcome.
- Preprocessing techniques include using sklearn library to impute missing values, split train/test sets with 25% test size and standardizing the sets.

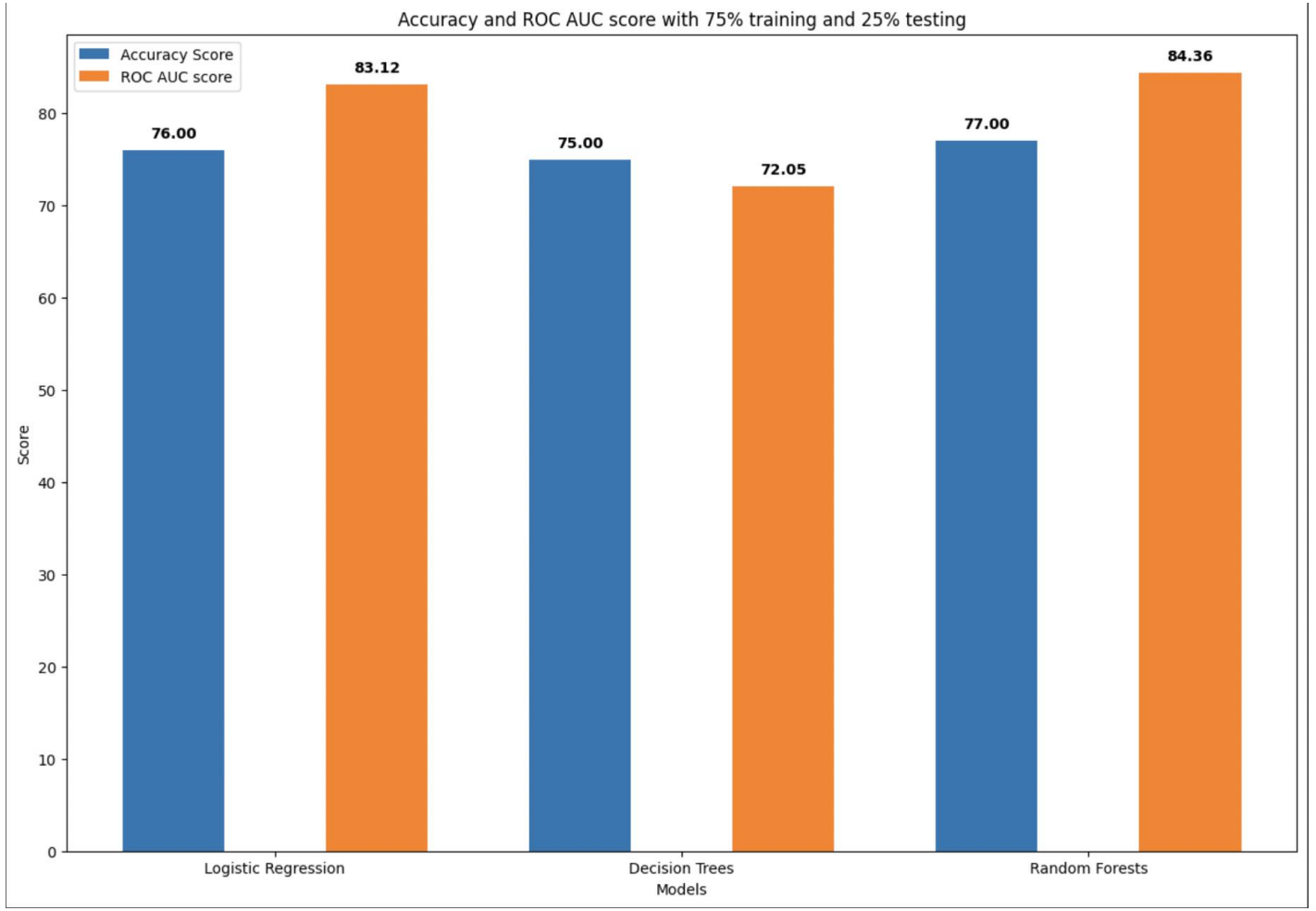
Key Insights

- Top findings from EDA that shaped our strategy was multicollinearity issues and outliers on some independent variables.

- For models that assumes linear relationships like Logistic Regression, this was going to be a huge roadblock when performing accurate predictions.

- This led to using more advanced models such as Decision Trees and Random Forest as they are non-parametric and overlook these issues.

Model Results and Interpretation



Model Results and Interpretation

Logistic Regression Classification Report

	precision	recall	f1-score	support
0.0	0.90	0.77	0.83	146
1.0	0.49	0.72	0.58	46
accuracy			0.76	192
macro avg	0.69	0.74	0.71	192
weighted avg	0.80	0.76	0.77	192

Decision Trees Classification Report

	precision	recall	f1-score	support
0.0	0.81	0.81	0.81	125
1.0	0.64	0.64	0.64	67
accuracy			0.75	192
macro avg	0.72	0.72	0.72	192
weighted avg	0.75	0.75	0.75	192

Random Forest Classification Report

	precision	recall	f1-score	support
0.0	0.79	0.87	0.83	125
1.0	0.70	0.57	0.63	67
accuracy			0.77	192
macro avg	0.75	0.72	0.73	192
weighted avg	0.76	0.77	0.76	192

Model Results and Interpretation

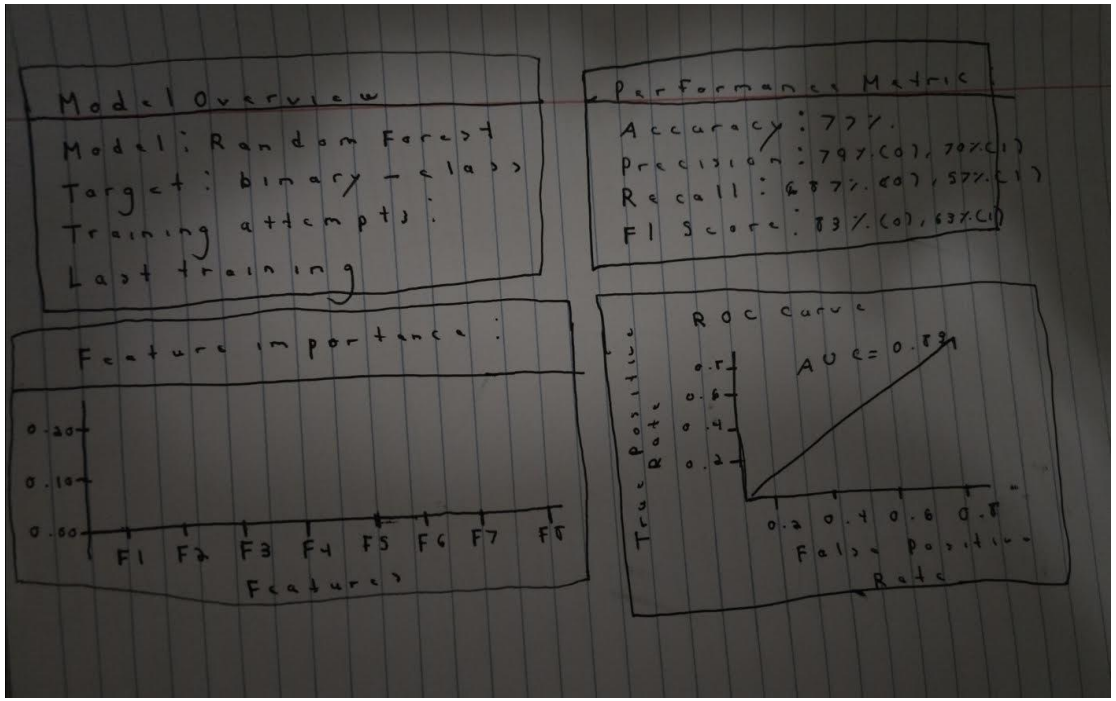
- Although Logistic Regression is higher than Decision Trees in terms of accuracy, ROC AUC and recall, both models do not take all features into consideration and their precision for both classes are unbalanced which can be risky when diagnosing a disease.

- The best model to use are random forests as it's the highest in accuracy, ROC AUC score, the precision for both classes are balanced and able to take all features into consideration.

- Using random forests in practice, doctors can see what features are crucial when determining if a patient has diabetes or not.

Next Steps and Dashboard Sketch

- Try more data pre-processing techniques.
- Better handling of class imbalance.
- Trying more model performance boosting techniques.



Thank you!