In this project, I developed a model to classify activities (swimming, biking, running, and walking) using a collection of GPS-tracked activities to create covariates. The covariates consisted of variables detailing speed, distance, elevation change, etc., and there were ~700 runs, ~600 bikes, ~70 swims, and ~30 walks total.

The R-package *glmnet* was used for multinomical regression with ridge regression. The package runs an iterative scheme to select a λ (penalization coefficient) that yields the lowest cross-validation error (use: *cv.glmnet*); I used 10-folds cross-validation. The objective function that is *minimized* is the -log of the *likelihood*. See below for model details.

The *training* accuracy of the multinomial regression model was 98.57%, the ***test* accuracy** **was** **83.24%**, but the baseline model (always guessing run as the activity, since it had the highest occurrence), was only 25% accurate.

|  |  |  |
| --- | --- | --- |
| **Model Name** | **Estimate of Prediction Error**  **(Accuracy, %)** | **Model** |
| Base | Total: 51.58%  Walk: 0 %  Swim: 0 %  Run: 100%  Ride: 0 %  Avg. Accuracy: 25% | Always guess: ‘run’ |
| Chosen Model | Total: 98.21%  Walk: 80 %  Swim: 91.23 %  Run: 99.01%  Ride: 98.57 %  Avg. Accuracy: 98.57% | Population model:  Likelihood:    Objective function with ridge regression to minimize: |

|  |  |  |
| --- | --- | --- |
| **Model Name** | **Estimate of**  **Prediction Error**  **(Accuracy, %)** | **Test**  **Error**  **(Accuracy, %)** |
| Base | Total: 51.58 %  Walk: 0 %  Swim: 0 %  Run: 100 %  Ride: 0 %  **Avg Accuracy: 25 %** | Total: 50.17 %  Walk: 0 %  Swim: 0 %  Run: 100 %  Ride: 0 %  **Avg Accuracy: 25 %** |
| Chosen model | Total: 98.21 %  Walk: 80 %  Swim: 91.23 %  Run: 99.01 %  Ride: 98.57 %  **Avg Accuracy: 98.57 %** | Total: 96.25%  Walk: 37.50 %  Swim: 100 %  Run: 98.64%  Ride: 96.85 %  **Avg Accuracy: 83.24 %** |