A Predictive Machine Learning Pipeline for Large-Scale Fitness Data

University of California San Diego

Master of Advanced Study, Data Science & Engineering

UC San Diego Jacobs School of Engineering

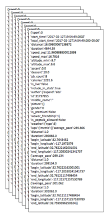
David Doerner, Jason Gilberg, Patrick Mulrooney, Masashi Omori

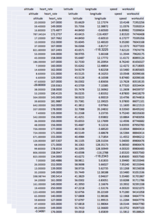
Advisor: Julian McAuley, PhD

SDSC SAN DIEGO SUPERCOMPUTER CENTER

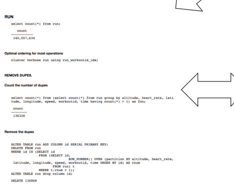
Machine Learning Pipeline

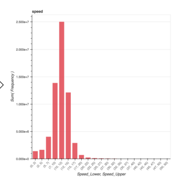
Raw Data **Database**



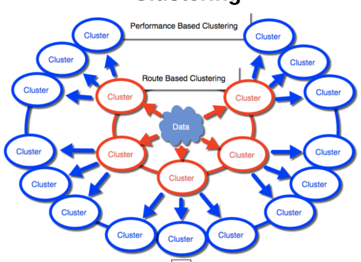


Clean Data **Visualize**

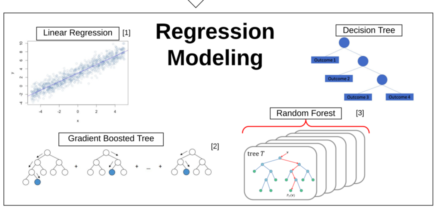




Clustering



Regression Modeling



Stacking (Ensemble)

Linear Regression, Random Forest

$$Y = c_1 r_1 + c_2 r_2 + c_3 r_3 + c_4 r_4$$

Gradient Boosted Tree, Decision Tree

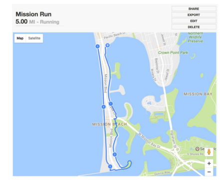
Abstract

A three-part machine learning pipeline was created to determine the level of accuracy attainable in predicting large-scale consumer fitness data on a commercial application, which tracked workouts using wearable devices and smartphones. The pipeline was built using a sequence of distributed unsupervised and supervised learning algorithms. A two-step sequence of bisecting K-Means clustering separated the dataset first into route clusters based solely on route information to identify targets for regression. Then, those route clusters were split into performance clusters based on the temporal and biometric data from the user, which allowed for the generation of new features that capture how well the user performs on particular routes. The route information and the user's historical performance were fit with four different regression models using the workout duration as the target value. The resulting regression models were assembled into an ensemble to make predictions for a specific user on a given route input. Despite inconsistencies due to malfunctions of the devices providing sensor data and the dependency on inherently variable user data, the pipeline was successful in creating predictions using a very limited number of features with accuracy levels that could be used for a novel fitness application.

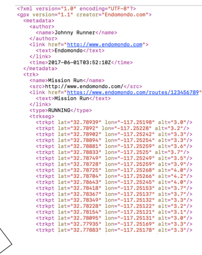
Prediction Pipeline

Prediction workflow:
1. User maps desired route
2. Route attributes are extracted: latitude, longitude, and altitude
3. Data factored into cluster selection
4. User's previous workout performance statistics assigned to cluster
5. Two clusters are fed into the model
6. Stacking is used to further refine the prediction, which is returned to the user

1 Route plan

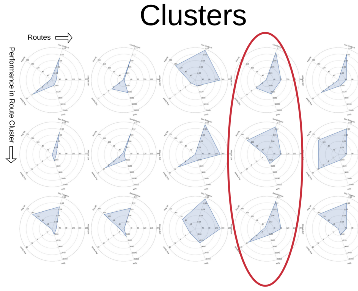


Route attributes



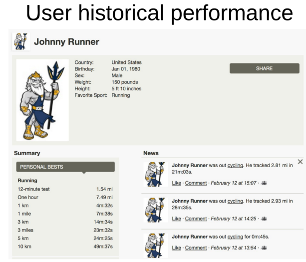
2

Clusters



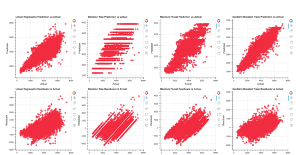
3

User historical performance



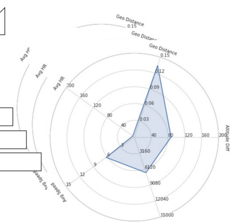
4

Single Model Predictions and Residuals

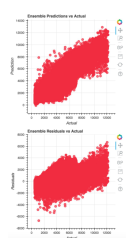


5

Identified Clusters



Ensemble Predictions and Residuals



6

Final Predictions

Route Cluster	Perf Cluster	Actual Value	Predicted Value	Absolute Error	% Error
0	0	4845	3820	1024	21%
	1	4845	3777	1067	22%
	2	4845	4180	665	14%

Findings:

The analysis of the existing data shows that future effort should be undertaken to improve the data quality at the source level. Several different techniques could be useful, but the first steps in identifying the sources of inaccuracy are finding trends between the different sources of data and the frequency of inaccuracy in related data. The various derived fields are extremely effective as a tool in identifying inaccuracy. Removing the restriction of representing the time series over 500 points should improve the finding of errant data. Incorporating external data sources, and building up an on-prem database of the information would provide a cost-effective lookup for data validation. Using these additional derived fields in the workout summaries should also enhance the accuracy of models produced with this data; although, it can be shown that a reasonable level of accuracy is achievable with a small number of fields by using a sophisticated, well-tuned machine learning pipeline.

The machine learning pipeline uses seven features to predict the duration of approximately 350,000 workouts within 13.2% using Mean Absolute Error or 19.4% using Root Mean Squared Error. The Mean Absolute Error is 23.4% of the standard deviation of the duration of the run workout duration, and the Root Mean Squared Error is 34.4% of the standard deviation of the run workout duration. The ensemble explained 88.2% of the total sum of squared, (R^2 = 0.882). This level of accuracy was achieved despite substantial amounts of missing data for altitude, speed, and heart rate, which made up ⅗ of the features used for clustering and ½ of the features used for regression. With more diligent data collection and curation as well as additional features generated from the original time series, prediction accuracy can increase, and the pipeline can be applied to the other types of workouts in the dataset that utilize geospatial data.

Acknowledgements:

amazon web services, APACHE Spark, jupyter, PostgreSQL, matplotlib, pandas, python, KanbanFlow

We would like to thank our friends, family, and employers for their support over the last two years. We would like to thank all those that wrote and contributed to the software packages that made this project possible. We would like to thank the outstanding faculty and staff that make the UCSD Data Science and Engineering program possible. And finally our advisor who provided indispensable advice and guidance throughout the project.

[1] https://3.bp.blogspot.com/...
[2] http://arogozhnikov.github.io/images/gbdt_attractive_picture.png
[3] https://dimensionless.in/wp-content/uploads/RandomForest_blog_files/figure-html/intro.jpg