CSCI E-63 Big Data Analysis - Final Project (Fall 2018)

Title: Reducing U.S. Commute Time with Machine Learning and Graph Analysis
Author: Walter Yu, Graduate Degree Candidate

Abstract: The average commute time within each U.S. census division has a large impact on its economy, productivity, infrastructure and environment. Longer commute times cause lost wages for workers with longer commute times, additional wearing of highway infrastructure and environmental impacts. As a result, this project evaluates commute patterns with the National Household Transportation Survey (NHTS) dataset provided by the Federal Highway Administration (FHWA) and whether public transportation or additional transportation planning could reduce commute times based on data analysis.

Introduction: This project continues the analysis developed for the 2017 NHTS Data Challenge. Whereas the contest entry focused on exploratory data analysis and visualization, this project continues analysis with machine learning and graph analysis.

Spark ML: Spark ML was used to continue the 2017 NHTS Data Challenge to evaluate relationships within the NHTS dataset tables; specifically, relationships between average annual trips, miles traveled and commute time. Analysis was completed with the Spark ML decision tree and gradient-boosted tree algorithms.

Spark GraphX: GraphX was used to evaluate relationships between the households and trips tables; specifically, relationships between census division and trip distance (miles driven) as follows: 1) create graph between household and trip tables, 2) analyze graph to identify additional relationships, and 3) evaluate In and out-degree relationships validated ML analysis results.

Benefits and Drawbacks:

Spark ML Benefits	Spark ML Drawbacks
1. Analyze large datasets quickly 2. Algorithm choices are sufficient for basic analysis 3. Supports data pipelines within Spark/Python environment	1. Algorithm choices may be limited for future analysis 2. Data visualization is limited (Matlibplot)
GraphX Benefits	**GraphX Drawbacks**
1. Examine graph relationships between data 2. Supports dataframes; minimal integration 3. Supports data pipelines within Spark/Python environment	1. Not as full-featured as Neo4J or graph databases 2. Data visualization is limited (Matlibplot)

Results: The analysis reinforces the intuitive notion that living within urban areas with access mass transit can help in reducing vehicle count and usage. As a result, additional transportation planning or outreach to raise awareness for reducing vehicle usage may help reduce commute times. Public awareness may lead to support for future transportation planning and vehicle usage reduction efforts.

Links:
Github: https://github.com/walteryu/e63-final
YouTube Summary (2 Minutes):
YouTube Technical Presentation (15 Minutes):