Predicting Ridership in Public Transit Systems: A Case Study

Joshua Morgan 038
jmorgan63@gatech.edu

April 21 2020

ISYE 7406: Data Mining and Statistical Learning

Abstract

Public Transit is hurting with costs rising and ridership falling. Tax payers are baring more of the burden for providing transit infrastructure. In Atlanta, Marta funds over half of its operating expenses with tax revenue. In 2018, the Atlanta Regional Commission invested in MARTA enabling the More MARTA initiative which will invest in MARTA and expand its capabilities. While more investment is good, transit investments have generally done with simple heuristic methods or intuition, adding some data driven insights to this could help inform MARTA of things to improve on. By taking data from the MARTA Origin Destination Transfer System, the American Community Survey, and General Transportation Specification Format a data set of the daily ridership and the demographics of census blocks is created. First a detailed feature engineering analysis was conducted to reduce the number of features from 56 to 14 that maintained the same amount of information. The data was then explored for patterns which highlighted the importance of the number of stops and rails stops. Finally, a series of machine learning models were fitted to be able to predict the effect of changes on the system. A Lasso Regression was selected as the best performing model. This model had three variables in the final model being the number of rails stops, number of stops, and the total population. Using this model and increasing the accessibility and reliability of the system, MARTA could see its ridership double. Besides this predictive power there were three main takeaways for MARTA. First, increasing reliability of the system will increase ridership. The superior performance of the rail system to the bus system suggest that MARTA should find innovative solutions to the last mile problem such as partnering with scooter companies. Second demographics had little to no impact on ridership. Traditionally, people have viewed public transit as a service for minorities, but the data suggests that race has little impact on ridership. Finally, MARTA's new experimental Origin Destination Transfer Matrix system provides adequate data for analysis, but this system could be improved to provide more powerful insights.

Contents

1 Introduction

1.1 Problem Description

According to the American Public Transit agency, nationwide transit ridership has remained constant or decreased over the last 30 years [6]. Public transit is a cheap good and is often substituted for a more convenient, accessible, reliable, and comfortable option, cars. In addition, to traditional competition from personal cars, new ride share marketplaces such as Uber and Lyft have entered the marketplace with affordable products such as Uber pool that have eaten into some of the market share of public transit. While volume has stayed constant or decreased, public transit agencies due to their public accessibility mandates cannot reduce service levels on a yearly basis, so operating losses have increased accordingly. Due to the environmental benefits and labor benefits of public transit, there has been renewed interest in expanding public transit, but transit planning has often done by rules of thumb and speculation. To help provide a more analytical approach to investing, we present a new data driven approach to forecasting ridership to potentially increase volume.

1.2 Motivation

Public transit is a public good providing essential transportation for individuals in the bottom quarter of the income distribution. For these individuals, transportation can be key to getting jobs and experience that allows them to move off of government aid. Not only does public transit have economic benefits but also environmental benefits [5]. These two primary benefits demonstrate that public transit has a significant role to play in creating a more equitable and sustainable society. Despite this important role, transportation agencies often themselves underfunded, with large capital cost, and consistent operating losses. In Atlanta, Georgia, the Metro Atlanta Regional Transportation Agency(MARTA) has been operating at a loss for the last 20 years and loses $1.50 per rider. While taxpayers cover this bill, an essential service like transit should be able to turn a profit. To help turnaround MARTA, the Atlanta Regional Commission (ARC) has funded the More MARTA initiative a ambitious growth strategy for the agency [4]. Traditionally transit expansion has been done using rules of thumb and intuition not data driven insight. While often this wildcatter approach to investment yields great returns, it does not mitigate the downside. By creating a data analysis of the current transit system, this expansion could be better informed.

1.3 Problem Solving Strategy

Since public transit is such a easily substituted good for more convenient, accessible, reliable, and comfortable options, a high level analysis to determine if there are any associations between different factors that could effect transit ridership. Three factor groups, accessibility, demographic, and reliability, have

been identified. Accessibility is the ability for individuals to interact with the system and use it to get from point A to point B. Demographic factors include both socio-economic factors and race and inform different preferences by social groups. Finally, reliability means how fast and consistently the service can get a individual to their potential destination. These three groups provide a firm foundation for analysis and emphasizes the interpret ability of the results.

1.4 Challenges

Unfortunately, analysis to help inform transit investments can be very difficult. Because of the general lack of funding sources for transit systems, there is a lot of technical debt with system existing in disparate databases that must be connected. Moreover, this data is very noisy since the system was designed for transaction accuracy not for analysis. Furthermore, the combination of transit system with socio-economic factors can be difficult because data sources such as the United States Census contain 100's of columns containing similar information. These features must be combined and filtered to help make results interpretable.

1.5 Approach

To tackle the problem, first a detailed description of the data, and a exploratory analysis will be conducted. With a firm understanding of the data, a analysis of each of the different features will be analyzed to create a set of features that are useful for further analysis. Following this multiple models will be created, and feature selection performed to optimize each model. Finally, key findings and results will be highlighted.

2 Data

The data for this report comes from three main sources: General Transportation Format Specification(GTSF) , MARTA's Origin Destination Transfer System (ODX), and the American Communicty Survery (ACS) [2][1].

2.1 GTFS

The GTFS dataset was extracted for the MARTA system and consists of multiple tables describing, stops, routes, schedules, trips, and other characteristics. The tables were combined using a procedure described in Appendix B.1. Following this combination procedure, a stable wiht a unique identifier, the name of a stop, the geolocation, and a indicator as a rail stop was created. There were 9170 stops of which 81 were rail stops(Appendix E, Figure 1).

2.2 ODX

The origin destination transfer system is a data set created by the Socially Aware Mobility lab at Georgia Tech. This system was created by a senior design team in 2019 and extract trips from a series of smart card entries. The system outputs multiple tables for different analysis, but the only table of interest for this project was the trips table (Appendix E Figure 2). The trips table consist of all of the records of a individuals trips throughout a day. The table was filtered for only entries in the morning and all columns but the start stop were dropped. For a detailed description of the filtering see Appendix A.2. The ODX data set consisted of only 12,849 trips form April 3, 2018 which is a fraction of the total trips in a given day, but since this system is experimental an analysis based on current information could inform improvements.

2.3 ACS

The American Community Survey is composed of two data sets conducted by the united states Census Bureau. There is a yearly ACS (ACS-1) and a every 5 years ACS (ACS-5). For the purpose of this report, we have selected the ACS-5 data set because it provides data to the level of a census block. For a detailed description of what a census block is and the extraction process visit (Appendix A.3). The ACS, also, comes with a series of shape files which describe the geographic regions in space and connect to latitude and longitude described before. The ACS data contained over 56 columns describing location, population, sex, race, and number of households.

2.4 Data Combination

These three data sets were combined together to create a single data table for investigation. First the stops table and the census shape file were combined using a spatial join function using the package geopandas [3]. Following this join operation aggregate statistics for each census block were collected about the number of stops and the presence of rail stops, and riders. Then the stops data table was joined with the ODX table and aggregate statistics were collected about the number of riders. Finally this table was joined with the ACS data table on the Census Block unique identifier. The final data set consisted of 945 rows and 65 columns.

2.5 Data Filtering

This data set consist of very sparse data with some census blocks not having any trips and some some having very large amount of trips. A initial problem in the analysis was the presence of the Atlanta Airport, where most system users are travelers or employees getting off of the night shift. Because these people are from out of town and cannot be traced this census block was removed form the analysis. Following this filtering, the final data set had 940 rows.

3 Proposed Methodology

3.1 Feature Engineering

With 940 rows and 50 variables that often are related to each other the data had a lot of information. A model with 50 variables could potentially be accurate, but it would not be very interpetable to an executive at MARTA. To help reduce the number of variables, first three types of analysis will be performed. The first type of analysis will see if there are any patterns between a single variable and the number of Trips. This analysis will inform whether some features should be dropped or combined from the data set. Then a random forest model will be fit onto the data, and the different trends variable importance's will be captured from the data. The variable importance will describe how often a variable is used and the influence of it. Finally a lasso regression will be done with a increasing penalty for more variables, so the model will have to estimate a increasingly sparse model. After this analysis, features emphasizing interpret-ability will be combined and others will be removed from the data set. The final data set will consist of a much smaller set of data that could potentially lead to more interesting, interpretable conclusions. For a technical explanation of each of these steps see Appendix B.1.

3.2 Exploratory Analysis

After feature engineering, a detailed exploratory analysis of the data will be performed. By understanding potential patterns in the data set, the model selection can be informed. First the correlation between each variable and response variable will be calculated. After the correlations have been analyzed, a principle component analysis will be performed and plotted to see if there are any clear clusters emerging in the data set. These clusters could be suggestive if there are different racial disparities such as more white people using transit than African Americans. Finally, a set of all the potential interaction terms and higher order terms will be calculated. These will then undergo the changing lasso regression to see if there are any important interactions. Finally, after using these different analysis techniques a subset of models will be selected.

3.3 Models and Model Selection

The problem of predicting ridership is a regression problem since there is a continuous response variable. The following models will be used to predict our the number of trips: K-Nearest neighbors (Regressor), Lasso Regression (LASSO), Ridge Regression (Ridge), Random Forest Regression (Random Forest), Elastic Net (Elast), Support Vector Machine(SVR), and Multi-Layer Perceptron (MLP) [7]. For a complete description of why different models were selected visit Appendix B.3. While each model performs well on different types of data, it also has to be optimized for this data set, so a exhaustive grid search was choosen to determine the optimal hyper parameters with a 10 fold cross validation using

root mean squared error as the metric. For a detailed description of the different parameters that were selected visit Appendix B.3. While machine learning models often have good performance on the training data, the data that is used to train these models introduces bias into the model. To control for bias in the model, a 10 fold cross validation procedure was selected using root mean square error as the selection criteria.

4 Analysis and Results

4.1 Feature Engineering

The feature engineering component of this project, led to a series of key findings. First all original variables were used through the feature engineering process described in the methodology section. The correlation analysis found the population of Asians, number of males 25 to 34, the number of rail stops, and the number of stops were positively correlated with the response term. The number of Asians seemed to be capturing potential train usage in the Buford highway area which is a predominantly Asian community with a train stop (Appendix E, Figure 7). Similarly, the population variables for males and females followed the same general trends suggesting these could be combined into simple population variables (Appendix E, Figure 7). The variable influence analysis indicated that most of the population variables outside of total population had little information due to low values (Appendix E Figure 6). Following these two analysis, the number of features was reduce by removing separate male and female population features and combining population features into smaller buckets. The final feature set was the number of stops, number of rail stops, total population, percentage of males, percentage of females, percentage white, percentage black, percentage Asian, percentage other race, number of households, population under 18, population 18 to 34, population 35 to 64, and population over 65. This smaller feature set provided a clearer set of features for analysis.

4.2 Exploratory Analysis

The exploratory analysis led to a two key findings. First the predominance of rail stops. In the correlation analysis, rails stops and stops were the most strongly correlated (Appendix E Figure 10). In the variable importance factor analysis, the number of rails stops was the strongest factor, and in the sparse estimation, the number of rail stops was the only predictor in all four models (Appendix E Figure 11). Second, the lack of importance of polynomial features and interaction terms. The sparsity analysis suggested that there were some interaction terms, but they all included rail stop meaning they could be adding little information(Appendix E Figure 8). A later variable importance analysis confirmed this theory as these features were still overpowered by the number of rail stops.

4.3 Modeling Results

While the modeling results are not exceptional, they are sufficient given the data limitations (Appendix E Figure 3). The MLP and random forest models performed poorly, which is likely due to limited amounts of data. The three regression models Ridge, Elast and LASSO models performed well. The Ridge and Elast converged to the same solution, but were outperformed by the LASSO. The LASSO performed well likely because it estimated a very sparse model which was good at generalizing to unseen data. The LASSO had all coefficients except Total Population, Rail Stops, and number of stops estimated at 0 (Appendix E figure 4).

4.4 Analysis

The Lasso Regression outperforming the random forest, neural network, and support vector machine indicates that most important relationship were likely linear and supports our exploratory analysis conclusions. The lasso coefficients indicated the most important variables were the total population, the number of stops, and the number of rail stops. This conclusion was similar to what was found in the exploratory analysis where the number of stops and rails stops were the most correlated and most influential variables consistently. In short, there is little evidence that demographics are a strong predictor for transit usage. While this might be counter intuitive, this likely highlights a confirmation bias among educated middle class Americans. A likely explanation of this is that public transportation is a easily substituted good along the dimensions of convenience, accessibility, and reliability. The inclusion of the number of stops in the final model highlights the influence of accessibility. If there is a bus stop nearby you are more likely to use it then if it is a mile away. Moreover, the presence of a rail stop highlights the dimension of convenience and reliability. Subways are often very convenient in dense urban cores like Midtown Atlanta, Buck head, and Downtown where they are close to most major attractions, business, and restaurants. Subways, also, are fast and reliable especially during rush hour traffic. A subway ride from midtown to the airport at 4 pm takes about 30 minutes, and a in a car takes an hour. This more reliable and quicker will always beat out a car if individuals have the option to take the subway. The lasso regression model was applied to data as if every census block had a rail stop in it. This would have doubled total ridership for the entire system. This increase highlights the importance of MARTA thinking creatively about redesigning its transit system.

5 Conclusion

This analysis suggest three main takeaways for MARTA. First ridership is a function of value. In a ever more competitive landscape, public transit can compete with other forms of transit where it is convenient, accessible, and reliable. MARTA's rail system normally performs well across these three dimensions

in surrounding neighborhoods, but the bus system generally performs poorly across these three dimensions. MARTA must find ways to strengthen its value proposition for riders that are not close to the rail system. Likely this is do to a poor solution to the last mile problem where to provide the publicly mandated accessibility convenience and reliability are sacrificed. A more dynamic mobility as a service system could provide the improvement along these dimensions.

Second ridership is not dependent on age and race. While public transit is stereo typically a realm of minorities, the number of white individuals in a community is not a predictor for lack of ridership. The prevalence of this notion is likely due to stereotypes and political campaigning against MARTA expansion in the 1970's from segregationist. While this is not a firm conclusion, further analysis of this question is warranted.

Finally, the origin destination transfer system provide adequate data to conduct analysis; however, an analysis of a single day is hard to draw widespread conclusions from. Our analysis of the strong performance of regression based methodology suggests econometric techniques such as mixed effects modeling and instrumental variable analysis could be extremely useful in making deductions about the transit system.

This study is limited by the missing ridership captured in other data sources in the origin destination system. The addition of these data sources is crucial to providing a clear analysis of the overall system. Other limitations include the lack of income in the variables of interest, but these could be added given additional time and resources. Despite these limitations the study provides clear guidance for MARTA on ways to grow and solve their volume problem.

5.1 Class Feedback

The class has been a great experience. Coming into this class I was self taught in a lot of these topics and had a surface level understanding. Following this class, I have definitely learned a lot and have been able to do deeper and more comprehensive analysis. I have mastered sci-kit learn and multiple R packages. These will definitely help me in my future work in industry. I think switching some of the course work over into python would be more useful since the APIs in python are much simpler, easier to use, and more attractive on the job market. I think a early semester project focusing on a theoretical problem similar to the test could be useful, but without the time constraint which causes unnecessary stress. I think the textbook for this class is great and provides a easy to read explanation of most models for a Georgia Tech ISYE graduate student.

A Appendix: Data Set

A.1 General Transit Specification Format

The GTFS dataset was extracted for the MARTA system and consists of multiple tables describing, stops, routes, schedules, trips, and other characteristics.

The stops table provided the name of the stops and its geolocation or latitude and longitude of a stop. The routes table was joined with the stops table to determine if a given stop was on a rail route. While being on a given route could have had a interesting effect, the route of a stop was removed due to the analysis only occurring on a single days worth of data.

A.2 Origin Destination Transfer Matrix System

The origin destination transfer matrix system works by taking different smart card swipes and assuming that individuals next stop must be in walking distance of their previous stop. This system then calculates the entirety of a trip in multiple legs providing the actual system demand.

A.3 American Community Survey

A.3.1 Census Data

The United States Census Bureau divides geographical regions on many different levels. While most individuals are familiar with zip codes, these are denomination of the post office. Instead the Census department divides regions by State then county the census tract then census block group then census block. A census block is a small are that generally encompasses a entire block in a city like Atlanta. Census blocks capture very detailed information and are useful for analysis because in dense cities like Atlanta a mile can be the difference between million dollar homes and 700 dollar a month apartments.

A.3.2 ACS Data Extraction

To extract data from the ACS, the US Census API was used. The 5 year profile table was selected, and a series of columns were selected with information of interest. These results were filtered for all the counties that MARTA serves, Fulton, Clayton, and Dekalb. Also, from the US census a set of shape files were downloaded.

B Appendix: Methodology

B.1 Scaling

The data was scaled according to the code in Appendix D Listing 2. Two different scaling procedures were used. First a MaxAbsScaler was used to scale data between 0 and 1 for the neural network. Then Data was scaled using a python Standard scaler which assumes normality of the data.

B.2 Feature Engineering

B.2.1 Correlation Analysis

The Pearson's correlation was calculated for each variable with the response variables. See Appendix D Listing 6 for the implementation. By estimating, correlations with the response variables such as Total Population 45 to 54 years old and Total Population 65 to 74 years old could be seen to have similar correlations. While this does not mean they have the same information, they are likely related and thus could be combined. Through analyzing the correlation analysis it was clear that some features should be combined.

B.2.2 Sparse Model Analysis

A lasso regression penalty penalizes large variables while trying to estimate the model. As a result, models with large coefficients will estimate sparser models. To determine which variables contained useful information a series of series of sparse models with penalties of 0.25, 0.5,0.75, and 1. The results should suggest which variables contain little information. For the implementation see Appendix D Listing 3.

B.2.3 Variance Importance Analysis

To see how important variables were, a variable importance analysis was conducted. In this analysis a random forest, which randomly selects features from the data set to estimate many models, is fit. After fitting the models a variable importance is calculated describing how often a variable was used and its weights. This provides a understanding of how important features are for generating accurate predictions. For the implementation see Appendix D Listing 7.

B.3 Modeling

B.3.1 K-Nearest Neighbors Regression

A K-nearest neighbors regression(KNN) will be used because it performs well if the data has strong local trends. The exhaustive grid search methodology will be done for k values of 2,4,6,9,10, and 12, and the weighting methods of distance based and uniform. The model optimization procedure selected a neighbors score of 8 with a uniform weight.

B.3.2 Lasso Regression

A Lasso Regression (Lasso) will be used because it estimates a sparse model allowing it to find the important few predictors from a sea of features. The grid search methodology was done for alphas 0.01,0.1,0.25,0.5,0.75, and 1. The optimized lasso found a penalty of 0.25.

B.3.3 Ridge Regression

Unlike the lasso which use a L-1 penalty, the ridge regression use a L-2 penalty, and a result tends to have more small coefficients in the solution. The grid search methodology was done for alphas 0.01,0.1,0.25,0.5,0.75, and 1. The optimal ridge regression. from our procedure found a penalty of 1.

B.3.4 Elastic Net Regression

The elastic net regression combines the benefits of ridge regression and lasso regression by combining both a L-1 and L-2 penalty. The grid search methodology was done for L-1 ratios 0.1,0.25,0.5, and 0.75, and penalties of 0.01,0.1,0.25,0.5,0.75, and 1. The optimal Elastic Net Regression had a penalty of 0.5 and a L-1 ratio of 0.5.

B.3.5 Random Forest Regression

The random forest regression estimates a series of trees using different features and different subsets of the data using a bagging procedure. The Random Forest Regressor was not optimized and was run by estimating 100 trees.

B.3.6 Multi Layer Perceptron

The multi layer perceptron implements a basic neural network, and these type of neural network excel at extracting unseen nonlinear model features. The neural network was solved using the lbfsg protocol provided in sci-kit learn. The activation functions were set to be relu. The number of layers was optimized using the gird search parameters selecting between 8 in the first layer, 4 in the second layer and 2 in the final layer. The model selected the three hidden layer option. Similarly, the learning rate was also optimized with alpha values between 0.01 and 1 x $10^{-}7$, and the grid search procedure found a alpha value of 0.001.

B.3.7 Support Vector Machine

The support vector regressor was used because it provides a way to estimate more nonlinear effects with a simpler model. The support vector machine was optimized using a grid search which had the following kernels: polynomial and restricted basis function. The other component of the support vector machine was the C values ranging from 0.1 to 1.5. After optimizing the model, the C value was 1, and the kernel was polynomial.

B.4 Model Selection

Our model selection was a 10 cross fold validation procedure. We selected the root mean square error because of its ease of use, communication value, and effectiveness in solving regression problems. Our 10 fold cross validation was used to ensure that the model did not have significant training bias.

C Appendix: Results

The Random Forest model has a root mean squared error of 11.09 indicating adequate performance. The neural network had a root mean squared error of 30.44 indicating extremely poor performance on this data. The lasso regression had a root mean squared error of 9.19 indicating superior performance to all other models. The Elastic regression model had the same solution as the Ridge regression model with a root mean squared error of 9.36. The Support Vector Regressor had a performance of 10.95. The k nearest neighbors had a root mean squared error of 9.41.

D Appendix: Code

```
din = data_df.merge(census_df, left_on=['COUNTY','TRACT','BLOCK'], right_on=['County','Census Tract','Block'])
din.head()
```

Listing 1: Merging Code

```
pipe_df = din.drop(columns=['COUNTY','TRACT','BLOCK'],axis=1)
pipe_df =pipe_df.sample(frac=1)
#Spliting into X, and y
X_df = pipe_df.drop(columns=['TRIPS','TRIP_RATE'])
y = pipe_df.TRIPS
scaler = preprocessing.StandardScaler()
nn_scaler = preprocessing.MaxAbsScaler()
X = scaler.fit_transform(X_df)
X_nn = nn_scaler = nn_scaler.fit_transform(X_df)
X1 = scaler.fit_transform(pipe_df)
poly = PolynomialFeatures(2)
X_poly = poly.fit_transform(X_df)
X_poly = scaler.fit_transform(X_poly)
```

Listing 2: Scaling code

```
alphas = [0.25, 0.5, 0.75, 1]
coefs = []
for alpha in alphas:
    las_reg = Lasso(alpha=alpha)
    las_reg.fit(X,y)
    coefs.append(las_reg.coef_)
#fig, ax = plt.subplots()
coefs = pd.DataFrame(data=coefs,index=alphas,columns=X_df.columns)
coef_bool = coefs.applymap(lambda x: x != 0)
temp = coef_bool.sum()
temp[temp != 0].plot.barh()
```

Listing 3: Sparse Model Estimation Code

```
rf = RandomForestRegressor(n_estimators=100)
rf.fit(X_2,y)
```

```
    vif = pd.DataFrame(rf.feature_importances_,index=X_df_2.columns
    )
    vif.sort_values(by=0,ascending=False,inplace=True)
    fig, ax = plt.subplots(figsize=(15,15))
    ax = plt.barh(vif.index,vif[0])
```

Listing 4: Variable Importance Analysis

```
    from sklearn.ensemble import RandomForestRegressor
    from sklearn.model_selection import cross_val_score
    rf = RandomForestRegressor(n_estimators=100)
    temp = cross_val_score(rf,X,y,cv=10,scoring='
    neg_mean_squared_error')
    print(sum(temp * -1)/10)
    cv_scores.append(("RF",sum(temp * -1)/10))
```

Listing 5: Model Fitting and Testing Code

```
    corrs = []
    for i in range(0,X.shape[1]):
        corrs.append(np.corrcoef(y.to_numpy(),X[:,i])[0,1])
```

Listing 6: Correlation Analysis Code

D.1 Github

Documented annotated code for this project can be found at `https://github.com/jemorgan1000/ML_Final_Project`. In this repository the file preprocessing.py describes the creation of the different data sets.

E Appendix: Figures

	stop_id	stop_code	stop_name	stop_lat	stop_lon	RAIL_STOP
0	100148	100148	MARTIN L KING JR DR NW @ CHICAMAUGA AVE NW	33.753823	-84.431838	0
1	210829	99971061	DEERFIELD PKWY @ TWO VERIZON PL	34.092878	-84.271101	0
2	100150	100150	MARTIN L KING JR DR NW @ CHAPPELL RD	33.753841	-84.433273	0
3	100152	100152	FAIR ST SW @ 1ST ST SW	33.748905	-84.425545	0
4	100154	100154	FAIR ST@FIRST ST	33.749077	-84.425298	0

Figure 1: First 4 rows of the stops table used in the analysis.

trip_id	breeze_id	start_stop	start_time	end_stop	end_time	stops	routes	num_legs	used_bus	used_train	error_bool	error_details
1000000	53225	908911	1/7/18 6:24	908845.0	1/7/18 7:04	[908911, 908845]	['0']	1	0	1	0	[]
1000001	53225	908845	1/7/18 14:40	908911.0	1/7/18 15:17	[908845, 908911]	['0']	1	0	1	0	[]
1000002	121242	902725	1/7/18 8:45	NaN	NaN	[902725]	['85']	1	1	0	0	[]
1000003	423253	905782	1/7/18 8:11	908435.0	1/7/18 9:03	[905782, 908845, 908435]	['124', '0']	2	1	1	0	[]
1000004	531875	906647	1/7/18 11:16	906369.0	1/7/18 11:45	[906647, 906369]	['0']	1	0	1	0	[]

Figure 2: First 4 rows of the stops table used in the analysis.

MODEL	RMSE
RF	11.095866
Neural Network	30.445530
Lasso	9.193099
Ridge	9.365712
ELAST	9.365712
SVR	10.954346
KNR	9.413995

Figure 3: The results from the optimized models.

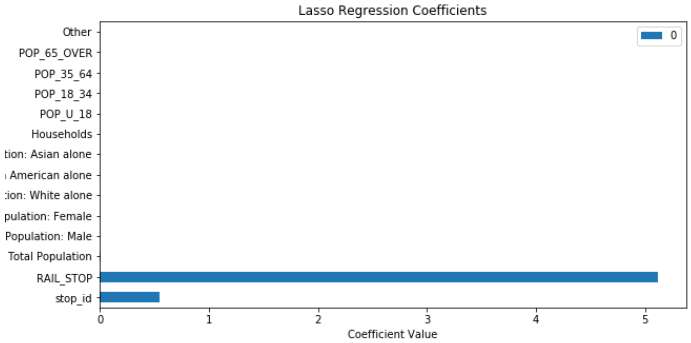

Figure 4: The coefficients from the final Lasso Regression Model.

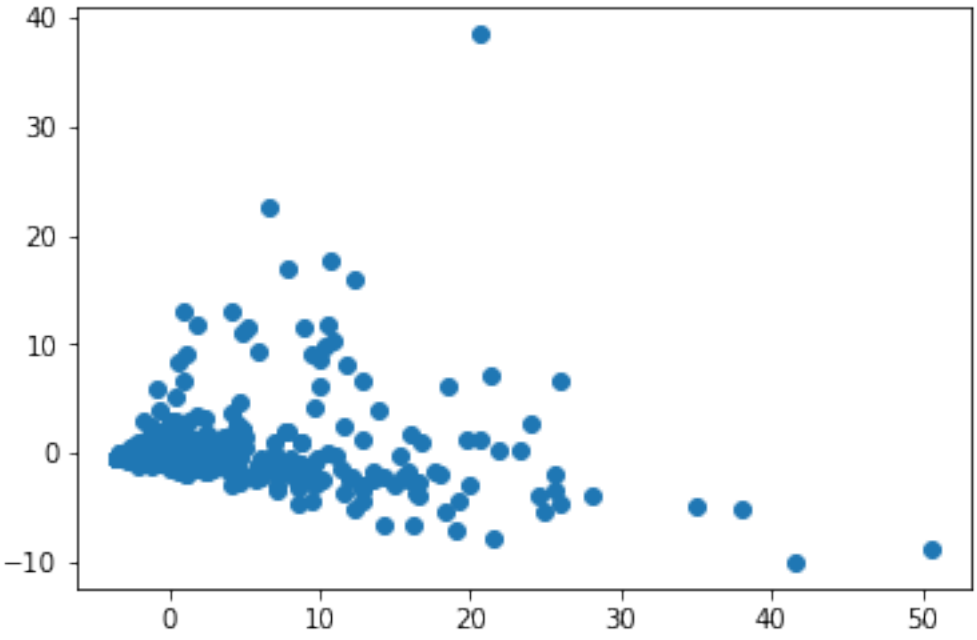

Figure 5: Plot of the last two principle components transformed.

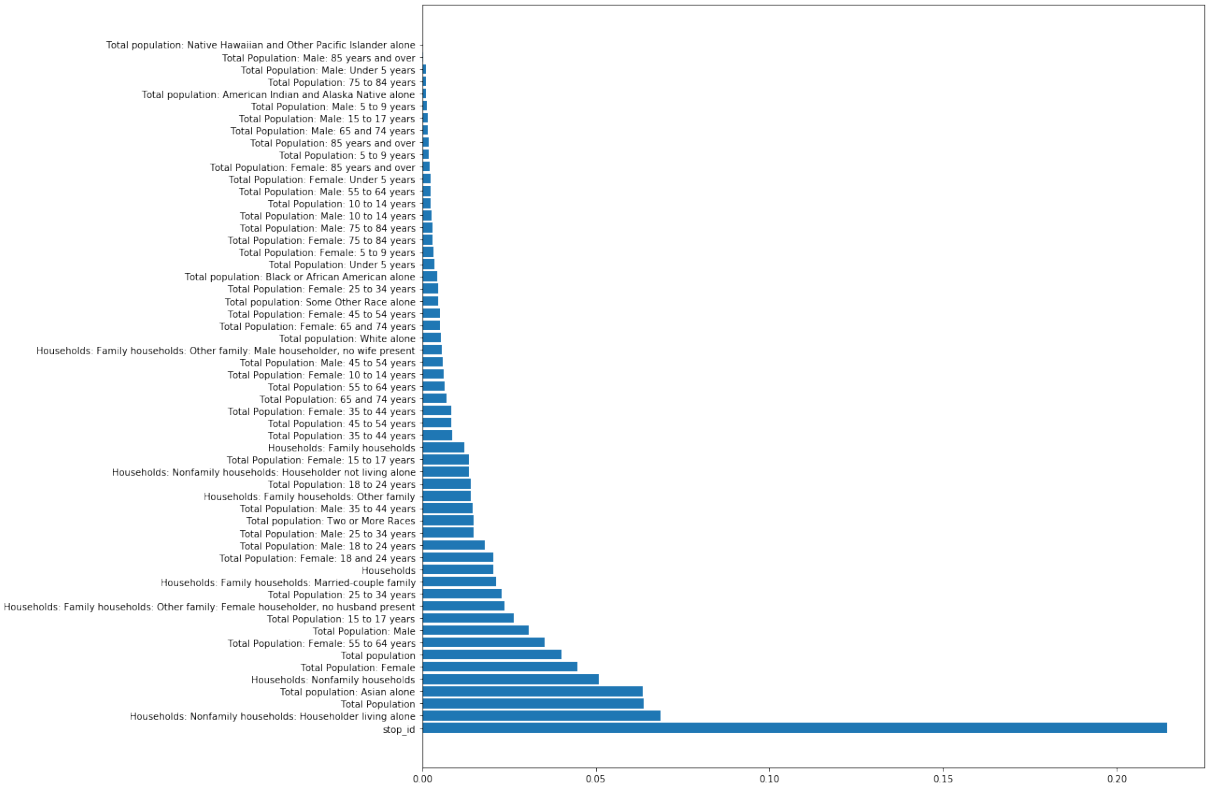

Figure 6: The variable importance factors from the original model.

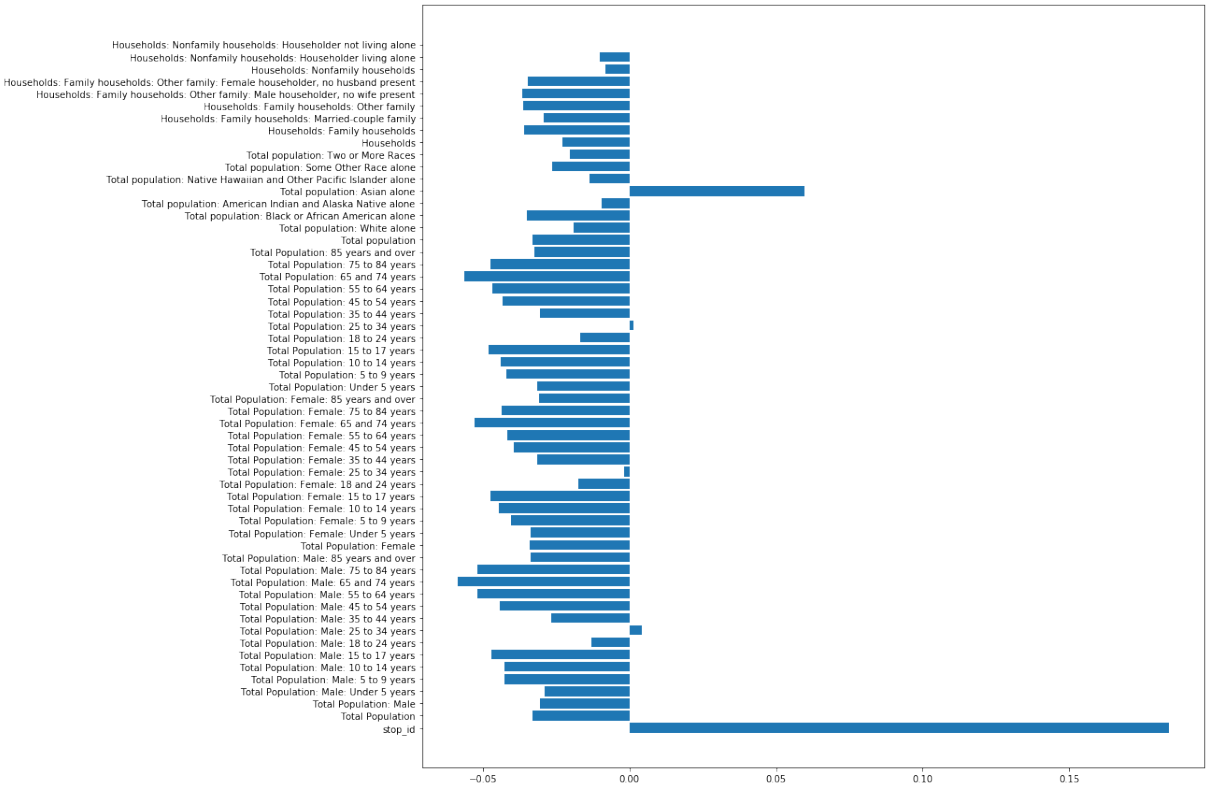

Figure 7: The correlations from the original model.

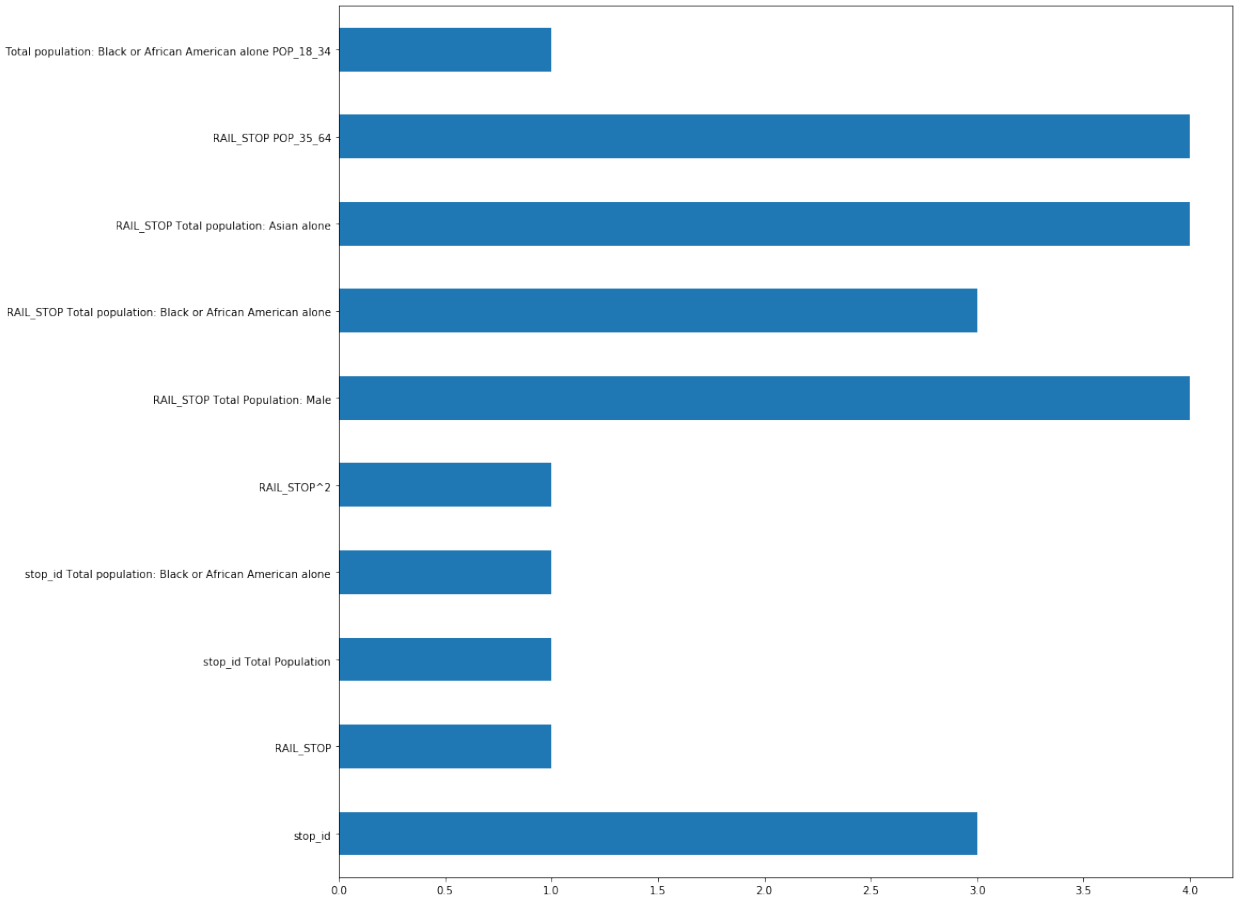

Figure 8: Sparsity analysis results from final variable set.

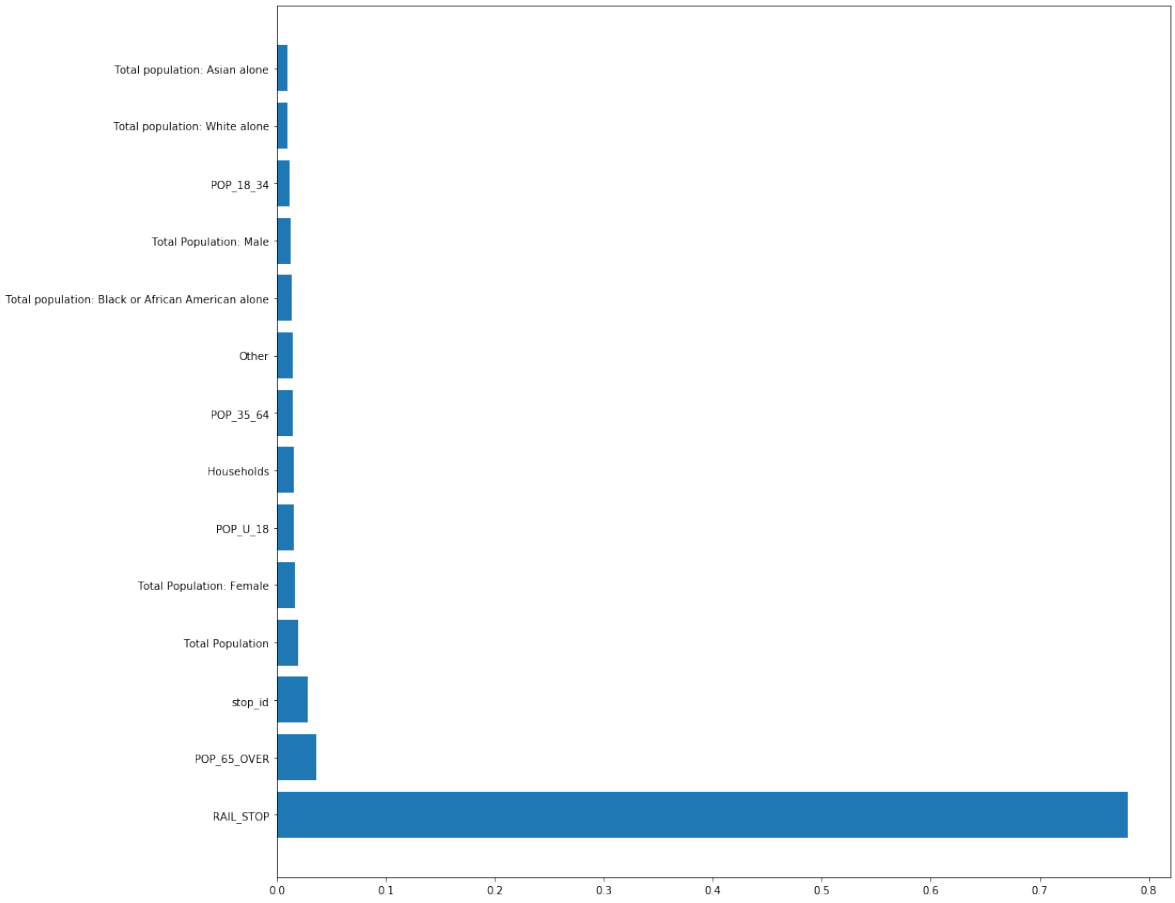

Figure 9: Variable Influence Analysis results with final feature set.

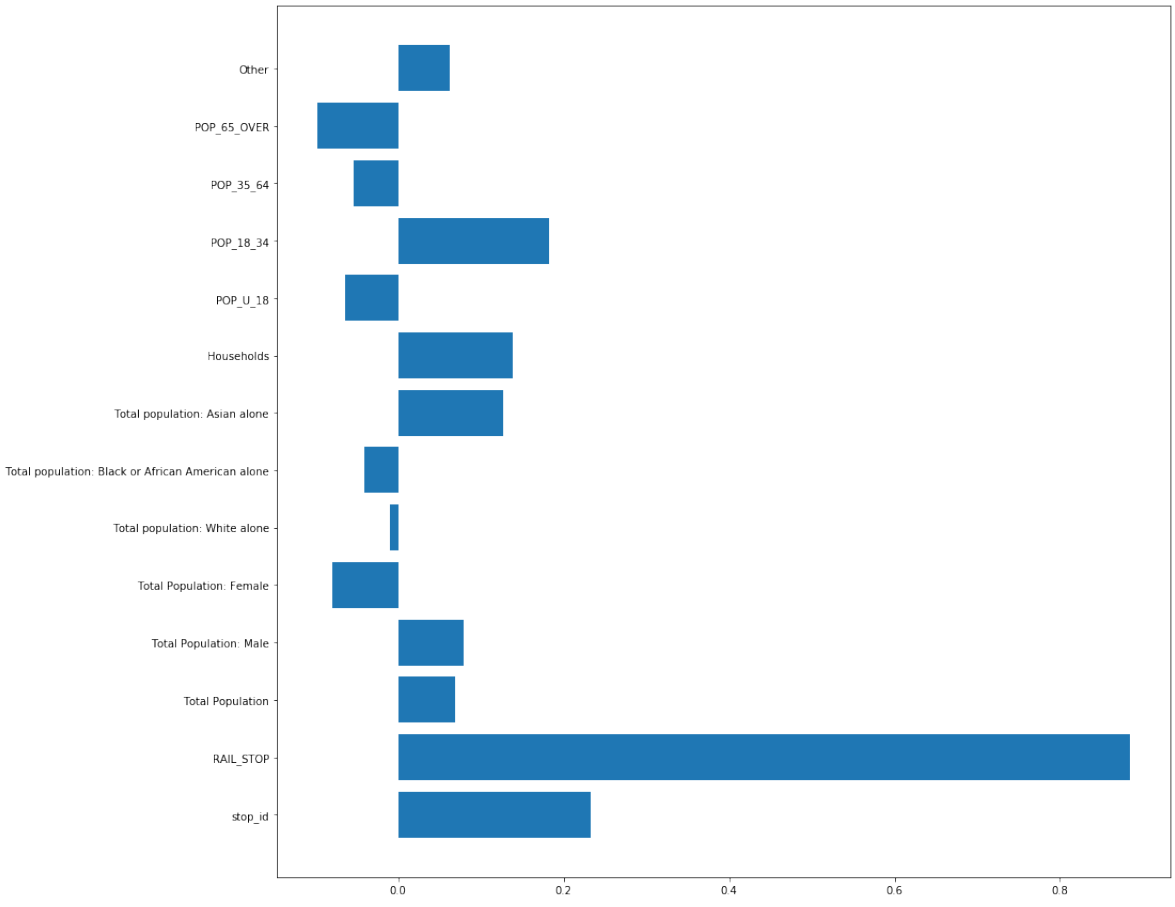

Figure 10: Results of correlation analysis on the final variable set.

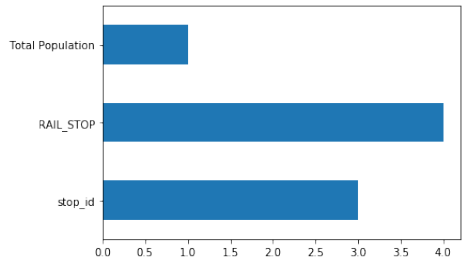

Figure 11: Results of correlation analysis on the final variable set.

References

[1] National Research Council et al. *Using the American Community Survey: benefits and challenges*. National Academies Press, 2007.

[2] *GTFS Static Overview — Static Transit — Google Developers*. URL: https://developers.google.com/transit/gtfs.

[3] Kelsey Jordahl et al. *geopandas/geopandas: v0.6.1*. Version v0.6.1. Oct. 2019. DOI: 10.5281/zenodo.3483425. URL: https://doi.org/10.5281/zenodo.3483425.

[4] "MARTA". In: *MARTA* (June 2019). URL: https://www.itsmarta.com/board-approves-expansion-sequencing.aspx.

[5] John Robert Meyer and Jose A Gomez-Ibanez. *Autos transit and cities*. Tech. rep. 1981.

[6] "Public Transportation Fact Book". In: *Public Transportation Fact Book* (). URL: https://www.apta.com/research-technical-resources/transit-statistics/public-transportation-fact-book/.

[7] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: https://doi.org/10.1038/s41592-019-0686-2.