Website URL: https://sites.google.com/view/transit-ridership-insights/home

Github Repo: https://github.com/Yash-Yashwant/transit-system-analysis

Milestone 3 Summary - Model Implementation

Public transit is a fundamental pillar of urban planning, representing an efficient means of transport for large numbers of people within the city and surrounding areas. While broadly popular in Europe and parts of Asia, ridership is anecdotally much less common in North America. Why this discrepancy exists, even in relatively well developed metropolitan areas, is a subject of interest among researchers. This study is meant to use publicly available data to investigate what drives public transit ridership in the United States. We hope to identify factors which can inform US agencies' efforts to boost ridership while also suggesting additional data collection efforts that could help future studies regarding transit usage in the US.

General Project Introduction

We are by no means the first people to think about public transit or its relationship to good urban design. The advocacy group Strong Towns, which focuses on outreach to citizens and local governments to improve the quality of life and infrastructure in American cities, focuses on the necessity of a public transit system as a means to promote economic development and growth in the urban core (Strong Towns). It makes the case for redirecting spending on roads and suburban expansion to projects aimed at the urban core, and notes that roads cost money too, and can often be less cost effective for the people they serve than investing in public transit infrastructure. Strong Towns also alludes to one problem that public transit can potentially solve, namely, that the current roadway designs in many cities are resulting in traffic problems for their residents. As their article on the topic points out, "if every errand generates a 10 minute car trip, you're going to have a traffic problem" ("What Causes Traffic Problems?", Strong Towns). Public transit as a way to get people to drive less is one of the benefits we mentioned above, and this benefit is emphasized by the group as one benefit of robust transit systems.

As for what features are desirable for a public transit system, one of the nice things about this is that ordinary citizens can have intelligent opinions on what works and what doesn't. Because the goal of many improvements to public transit are to increase ridership, a good starting point is to consider what would incentivize you to take or not take public transport. Factors include the location of stops, frequency of service, and perceived safety of the system. Another common problem in North American transit is a seeming paradox: even when transit is available, people don't use it. One can find many pieces of evidence noting this paradox, some formal, some anecdotal. One interesting anecdotal account is of the problems regarding the transit system in Ottawa, Canada. In a video by the YouTube channel Answer in Progress, where they compare the efficiency of walking, driving, or taking transit. Incredibly, even in a city with as many documented transit issues as Ottawa, the person who took public transit got to all of the destinations in the group's experiment one hour and twenty minutes faster than the person driving. In fact, the driver was only thirteen minutes faster than the person that walked to every destination without any other mode of transportation. When transit works, it works.

Beyond videos by transit enthusiasts, advocates, and experts, there are also academic studies looking into the factors that drive public transit acceptance (or nonacceptance) in the population. Some of these papers focus on factors outside the control of agencies, like the increased prevalence of rideshare services or more people working from home (Erhardt et al.). Other studies do include factors that agencies can control, and while they also find that broad societal factors affect ridership significantly, they point to service frequency and reliability as key factors (Taylor and Fink). Some papers study very specific regions, like two papers from Yuxin He and collaborators examining transit trends in the Shenzhen Metro, or one study by Hyungun Sung and Ju-Taek Oh focusing on Seoul, South Korea. With regards to future work, the Taylor and Fink paper emphasizes the need for studies to take into account possible correlations between explanatory variables, try to build models with predictive power, and explain their findings in terms tailored to the agencies making the policy decisions surrounding transit infrastructure.

Data Collection and Preprocessing:

Data Sources:

When looking for data regarding public transit in the United States, the US Department of Transportation manages the National Transit Database (NTD) through the Federal Transit Administration. This is the primary repository of data about transit systems in the US, set up after reporting was first required by law in 1974. The NTD has data on a variety of different metrics spanning several years. For the purposes of our analysis, we were primarily interested in finding information about the annual expenditures and travel statistics for transit systems across the US. This includes information related to the safety records of the agencies as well as the amount they collect in fares.

Transit agencies in the United States are required by law to report to the NTD at least yearly, with reporting obligations differing depending on the size of the agency. In addition, all agencies are required to report safety and security incidents to the National Transit Administration. Our project is interested in researching factors that contribute to public transit usage in the United States, and so our sample of interest is the existing agencies in the United States. Most agencies are classified as full reporters, with the exception of the smallest or most remote agencies. Thus, even a consideration of only full reporters gives a decent cross-section of the country, since a full reporter could cover the Washington DC transit agency, RTD, or even the Southern Nevada transit agency. These are very different parts of the country with different needs, so any findings we make are at least reasonably likely to have general applicability. The legal requirements to report to the database, combined with the fact that this is a product released by the National Transit Administration, means that the data submitted must pass through initial scrutiny before being published. The NTA flags data that it considers to be questionable and takes steps to make sure the datasets are free of duplicates.

Data Cleaning and Transformation:

When extracting the data from the NTD, capital expenditures data are provided in different forms depending on the year. For 2022, the column format and data definition are spelled out on a dedicated webpage. For the years 2021 to 2016, the column formats are slightly different, usually defined in a dedicated Excel sheet in the XLSX or XLSM file provided by the database. The first step in compiling a single file containing data for each of these years is to ensure that all of the data sets share the same columns. This was done by removing the columns containing questionable data, which were largely empty, and then removing the columns that contained legacy identifiers or additional identifiers that were not included in older data sets (one can see the processing Jupyter notebook in the GitHub repository for details on this process). This left a common set of 21 columns detailing the expenditures of transit agencies from all over the United States for the years 2016-2022.

Data for collected fares is extracted similarly, with details for each agency separated into different Excel files by year in the NTD. As a published data product, the data has largely been vetted already, so data cleaning involved removing extraneous columns from the data and combining the sets into a single CSV file. Data cleaning for the safety incidents took on a slightly different form, as the NTD publishes a file that is updated periodically containing information on all incidents requiring reporting to the Transit Administration dating back to the start of data collection, which is farther back than our period of interest. Cleaning these datasets required aggregating information on the incidents reported in the Excel file by year, transit agency, and mode of operations. The results of these manipulations are then saved to CSV files for later use. Details on the specifics of the manipulations, as well as snapshots of the data frames, are found in the individual cleaning notebooks inside the Git repository.

Feature Selection:

In training the models used to explain the relative importance of different features in predicting the ridership of different transit systems, we primarily focus on features that broadly summarize the key aspects that we previously suggested might be important to the general public when weighing whether to take public transit or not. For example, total fare revenues divided by total unlinked passenger trips yields the cost of a single trip. The information regarding accidents, injuries, and fatalities on each system gives information about the general safety of the transit systems. Information about the capital expenditures on each system tells us how much investment each agency is putting into their services, and the same file gives us information about the population in the service area of the transit agency, which we can use to control for general population effects when evaluating the influence of each factor.

In creating a model used to predict a numeric quantity, we are primarily interested in two things. First, of course, is accuracy. The second is model simplicity. The stated goal of this project is to take the available data on transit ridership in the United States and turn it into a set of insights that can be used by agencies to help decide what to do to make their services more popular. In order to do that effectively, it would be best for the recommendations to be generally understandable. In the pursuit of simplicity, it is thus advisable to start with broad indicators and only increase the specificity if it proves necessary, thus keeping the number of free parameters low. In all cases, the feature selection focuses on identifying the features we think are likely to be the most impactful predictors of ridership and refining those selections by evaluating the resultant models.

Key Features:

- **Vehicle Revenue Hours**: A key driver of service coverage and ridership. Refers to the total hours vehicles are in service and available to passengers. It had a direct influence on ridership, as increased vehicle revenue hours improve service availability, making transit more convenient and accessible.
- **Passenger MIles:** It represents the cumulative miles traveled by all passengers on a transit system. A higher passenger miles indicates that riders are traveling for longer distances. This metric helped us understand the scale and the utilization of the system.
- **Operating Expense**: It represents the total cost associated with running the transit system, including labor, maintenance, fuel, and administrative costs. Higher expenses indicated that the system was larger and was much more complex. It was a good benchmark, systems with higher ridership and lower expenses per mile/trip are considered to be efficient.
- **Fares (Total, Paid by Organizations/Passengers):** Includes all fare revenue collected directly from passengers/organizations. Increased fare can directly reduce ridership, mainly amongst price-sensitive passengers. Agencies optimize fare by analyzing demand elasticity.
- **Primary UZA Population:** Represented the urban population served by a transit system, offering demographic insights into the coverage and demand for public transportation.

- **Mode VOMS (Vehicles Operated in Maximum Service):** Captured the total number of vehicles deployed at peak capacity, serving as a critical operational metric for measuring system capability.

- **Total Spending:** Summarized financial resources allocated to transit operations, including expenses for maintenance, labor, and infrastructure.

- **Safety Metrics:** Included data on minor injuries, major injuries, fatalities, and property damage, providing a quantitative view of the system’s safety record.

Model Implementation

Rail

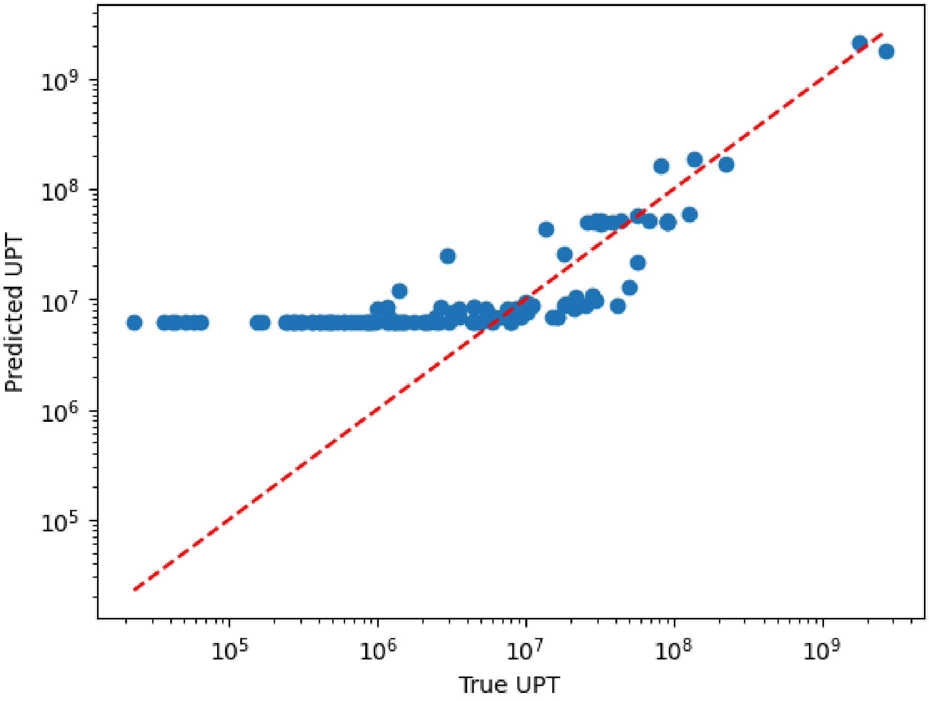
It is fairly intuitive to consider the fact that bus and rail ridership would be affected by different factors. As we mentioned in the data exploration, trains require track in order to run, and it is expensive to build guideway, much more than getting a bus to drive along existing streets. Building tracks and stations underground is even more costly. Given these differences, in this section we isolate rail systems from our data and try fitting models only on rail transit.

The NTD data contains information on the specific mode of transit, so it is easy enough to select only those modes that correspond to rail. In practice, this means systems like metro rail, light rail, trams, and commuter rail. Our metric of ridership for this section is unlinked passenger trips (UPT), which are a measurement of every instance of travel along the system. When estimating fares, we divide the total amount of fares collected by the transit agency for that mode by the number of passenger trips to get an estimate of the fare per trip. Then, we need to decide what features from the data to use as parameters to predict UPT. We want to predict a continuous variable, so we can isolate the numerical variables that we think capture information that would be relevant to users of transit systems. After isolating the data, we obtain the following table.

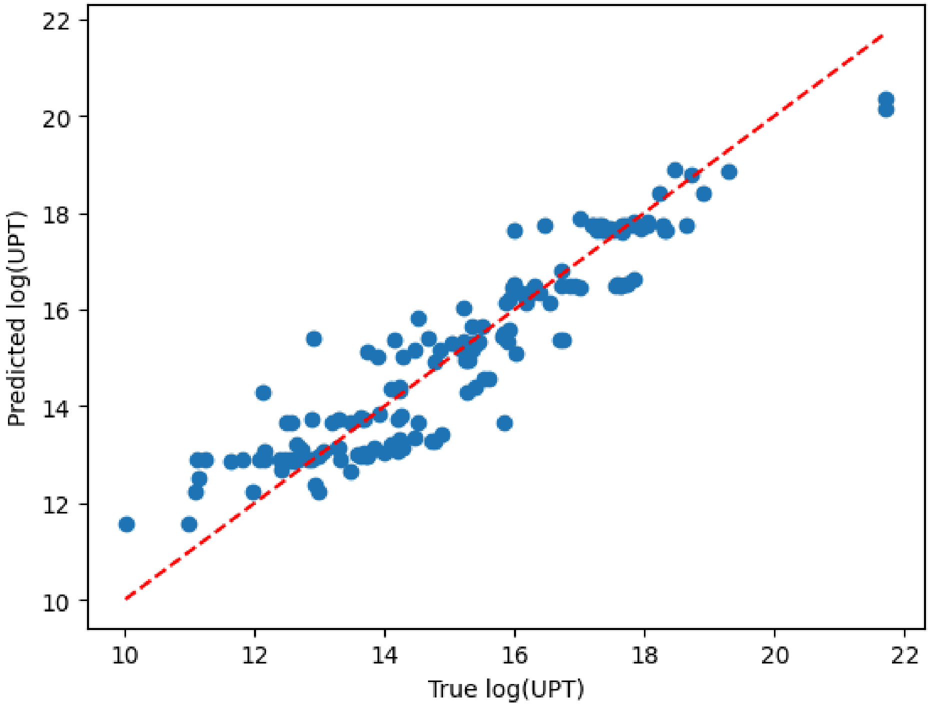
```
<class 'pandas.core.frame.DataFrame'>
Index: 632 entries, 18 to 6840
Data columns (total 10 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Primary UZA Population  632 non-null    float64
 1   Mode VOMS               632 non-null    float64
 2   Total Spending          632 non-null    float64
 3   Total Minor Injuries    632 non-null    float64
 4   Property Damage         632 non-null    float64
 5   Total Injuries          632 non-null    float64
 6   Total Fatalities        632 non-null    float64
 7   Total Serious Injuries  632 non-null    float64
 8   Fare                    632 non-null    float64
 9   UPT                     632 non-null    float64
dtypes: float64(10)
memory usage: 54.3 KB
```

UPT is our target variable, and the others capture information about the characteristics of the area served by the transit system (population), the number of vehicles the agency operates (Mode VOMS -- Vehicles Operated at Maximum Service), the safety record of the agency, and the cost to ride public transit. The first thing that stands out is the number of entries in the data table. Only 632. This is not to say that the data is incomplete, but to highlight how few rail systems exist in the US. As such, one recommendation could be to build more rail. This is largely due to efficiency considerations: trains can be longer than buses, and so can transport more people while only needing one driver. Additionally, trains can bypass traffic and are often more comfortable over long distances. While situations that require minimal disruption to an urban core are often better suited to buses, applications that have the space would be well advised to consider trains.

We first try a **random forest regressor**. The goal is to avoid overly complicated trees in favor of an ensemble of many simple ones, so we limit our maximum depth. After some iteration, we settle on trees with a maximum of three levels, and obtain a model with an R-squared parameter of 0.91. This is, by that metric, an excellent fit. However, a slightly closer look reveals an issue.



The predicted UPT values are artificially bounded from below, rendering the model almost entirely ineffective for predicting the ridership of agencies that see less than 10 million UPT. This strange prediction floor is certainly not desirable behavior, and a look at the data suggests a possible reason why. The ranges on the values used as predictor variables are wildly mismatched; some of them are in the millions and some only go up to 100. Thus, one possible avenue for improvement is to transform the data in order to bring the values closer to each other. An obvious transformation is to take the logarithm of the values, an instinct slightly complicated by the fact that some of the values go to zero, which makes the logarithm undefined. However, for the columns that contain only very large values, 0 can be replaced with 1 without much effect at all, which then allows the logarithmic transformation. Applying it to the data results in a new range, one which is much better matched across all the predictors. Training a new tree produces a new plot, one which avoids the artificial bounding problem.



This new tree produces an R-squared value of 0.863 and appears to do quite a good job predicting the overall ridership given the information provided. Even with a relatively simple model, there is good agreement, although there is a suggestion that the model may be systematically underestimating the ridership for very large systems and underestimating the ridership for very small systems. Taking a look at the features most influential to the random forest model, we obtain the following table.

```
Mode VOMS                 0.884647
Total Minor Injuries      0.036011
Total Injuries            0.030670
Primary UZA Population    0.022279
Fare                      0.016620
Total Spending            0.004799
Total Fatalities          0.002813
Total Serious Injuries    0.002161
Property Damage           0.000000
```

We see that the most significant contribution by far is the number of vehicles in operation, trailed significantly by the other variables. However, this output is limited in that it does not provide information on the statistical significance of the predictors. For that, we turn to regression.

The scikit-learn package does not provide statistical significance information as part of its features, which is why we turn to statsmodels instead. When doing the fits, we stuck with simple linear models with intercepts, and fit using **Generalized Linear Regression**. When fitting to the data, we get quite a bit of information out.

```
                            GLS Regression Results
==============================================================================
Dep. Variable:                    UPT   R-squared:                       0.858
Model:                            GLS   Adj. R-squared:                  0.856
Method:                 Least Squares   F-statistic:                     408.6
Date:                Mon, 18 Nov 2024   Prob (F-statistic):          2.19e-251
Time:                        17:48:43   Log-Likelihood:                -756.24
No. Observations:                 619   AIC:                             1532.
Df Residuals:                     609   BIC:                             1577.
Df Model:                           9
Covariance Type:            nonrobust
======================================================================================
====
                             coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------------
----
const                      7.7530      0.489     15.855      0.000       6.793
8.713
Primary UZA Population     0.2540      0.035      7.161      0.000       0.184
0.324
Mode VOMS                  0.9420      0.030     31.661      0.000       0.884
1.000
Total Spending             0.0260      0.009      2.989      0.003       0.009
0.043
Total Minor Injuries       0.0004      0.000      1.104      0.270      -0.000
0.001
Property Damage            0.0600      0.011      5.328      0.000       0.038
0.082
Total Injuries             0.0030      0.002      1.732      0.084      -0.000
0.006
```

```
Total Fatalities          -0.0260      0.011     -2.326      0.020      -0.048
-0.004
Total Serious Injuries     0.0107      0.006      1.790      0.074      -0.001
0.022
Fare                      -0.1197      0.012    -10.079      0.000      -0.143
-0.096
==============================================================================
Omnibus:                       40.241   Durbin-Watson:                   0.765
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               49.258
Skew:                          -0.584   Prob(JB):                     2.01e-11
Kurtosis:                       3.738   Cond. No.                     3.47e+03
==============================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 3.47e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

This is, admittedly, a lot of output, but it does what the random forest does not. It takes all of the predictors that we thought might be important and tells us the statistical significance of the effect suggested by the data. With that information, we see that the effects for all of the injury and fatality metrics are not statistically different from zero. This suggests a new model to fit that only keeps the significant data. Refitting obtains

```
                            GLS Regression Results
==============================================================================
Dep. Variable:                    UPT   R-squared:                       0.855
Model:                            GLS   Adj. R-squared:                  0.853
Method:                 Least Squares   F-statistic:                     600.7
Date:                Mon, 18 Nov 2024   Prob (F-statistic):         1.15e-252
Time:                        18:04:22   Log-Likelihood:                -762.89
No. Observations:                 619   AIC:                             1540.
Df Residuals:                     612   BIC:                             1571.
Df Model:                           6
Covariance Type:            nonrobust
==================================================================================
====
                             coef    std err          t      P>|t|      [0.025
0.975]
----------------------------------------------------------------------------------
----
const                      7.7696      0.493     15.761      0.000       6.802
8.738
Primary UZA Population     0.2514      0.036      7.032      0.000       0.181
0.322
Mode VOMS                  0.9588      0.029     32.654      0.000       0.901
1.016
Total Spending             0.0262      0.009      2.985      0.003       0.009
0.043
Property Damage            0.0566      0.011      5.124      0.000       0.035
0.078
Total Fatalities           0.0099      0.005      2.175      0.030       0.001
0.019
Fare                      -0.1252      0.012    -10.633      0.000      -0.148
-0.102
```

```
==============================================================================
Omnibus:                       36.566   Durbin-Watson:                   0.754
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               43.706
Skew:                          -0.556   Prob(JB):                     3.23e-10
Kurtosis:                       3.677   Cond. No.                         332.
==============================================================================


Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Now, each of the effects are statistically significant, although some have rather puzzling signs. Total Fatalities and Property Damage both have positive coefficients, albeit small ones, even though these are metrics that capture the number of accidents on transit systems. Total Spending is also significant, though not all that important compared to the other fit coefficients. Given this, it is natural to wonder if we can make a "minimal model" consisting of only the values that have coefficients larger than 0.1. This minimal model produces the below results.

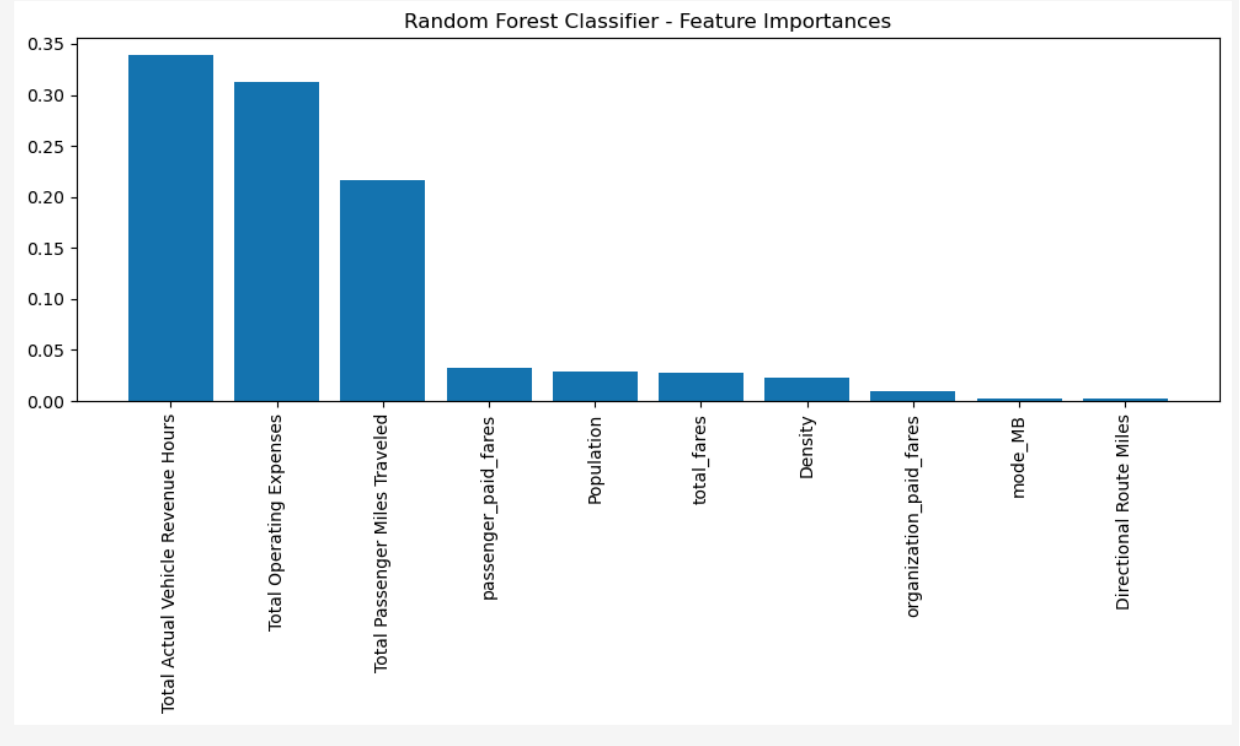
```
                            GLS Regression Results
==============================================================================
Dep. Variable:                    UPT   R-squared:                       0.845
Model:                            GLS   Adj. R-squared:                  0.845
Method:                 Least Squares   F-statistic:                     1121.
Date:                Mon, 18 Nov 2024   Prob (F-statistic):          9.56e-249
Time:                        18:11:27   Log-Likelihood:                -782.49
No. Observations:                 619   AIC:                             1573.
Df Residuals:                     615   BIC:                             1591.
Df Model:                           3
Covariance Type:            nonrobust
======================================================================================
====
                             coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------------
----
const                      7.9791      0.505     15.792      0.000       6.987
8.971
Primary UZA Population     0.2549      0.037      6.952      0.000       0.183
0.327
Mode VOMS                  1.0300      0.024     43.030      0.000       0.983
1.077
Fare                      -0.1403      0.012    -11.981      0.000      -0.163
-0.117
==============================================================================
Omnibus:                       38.770   Durbin-Watson:                   0.795
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               45.521
Skew:                          -0.595   Prob(JB):                     1.30e-10
Kurtosis:                       3.590   Cond. No.                         228.
==============================================================================


Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

This model has almost the same explanatory power as the model that is fit to all of the variables, but it is much, much simpler, and all of the fitted coefficients are highly statistically significant. Other metrics which inform how good the model is, like AIC, BIC, and the Condition Number, are also much better than the model that fits to all of the variables.

Road

Road-based transportation systems, unlike rail, are not dependent on fixed guideway and other similarly expensive infrastructure, instead using existing roadways to move people around. Their success is highly dependent on how efficiently they operate. The NTD sorts buses by subtype, but all road transit can be extracted from the data with relative ease by filtering for specific modes, including motor bus (MB) and paratransit bus (PB). As mentioned in the above section, the key ridership metric is Unlinked Passenger Trips (UPT). We expect bus ridership is influenced by several factors, including operational capacity, population served, fares, and safety metrics. To begin to test this, we trained a random forest on the data, which did quite well. Assessing the relative feature importance, "Total Actual Vehicle Revenue Hours" stands out as the most significant, which makes intuitive sense; the more time the buses are running, the more people are going to take them. Total operating expenses also contributed heavily. Factors such as the area population and the length of the routes were not as significant. The suggestion, based on the features, suggests that focusing on making transit more available, spending wisely, and ensuring it is accessible to people are important contributors to ridership.

Random Forest Classifier



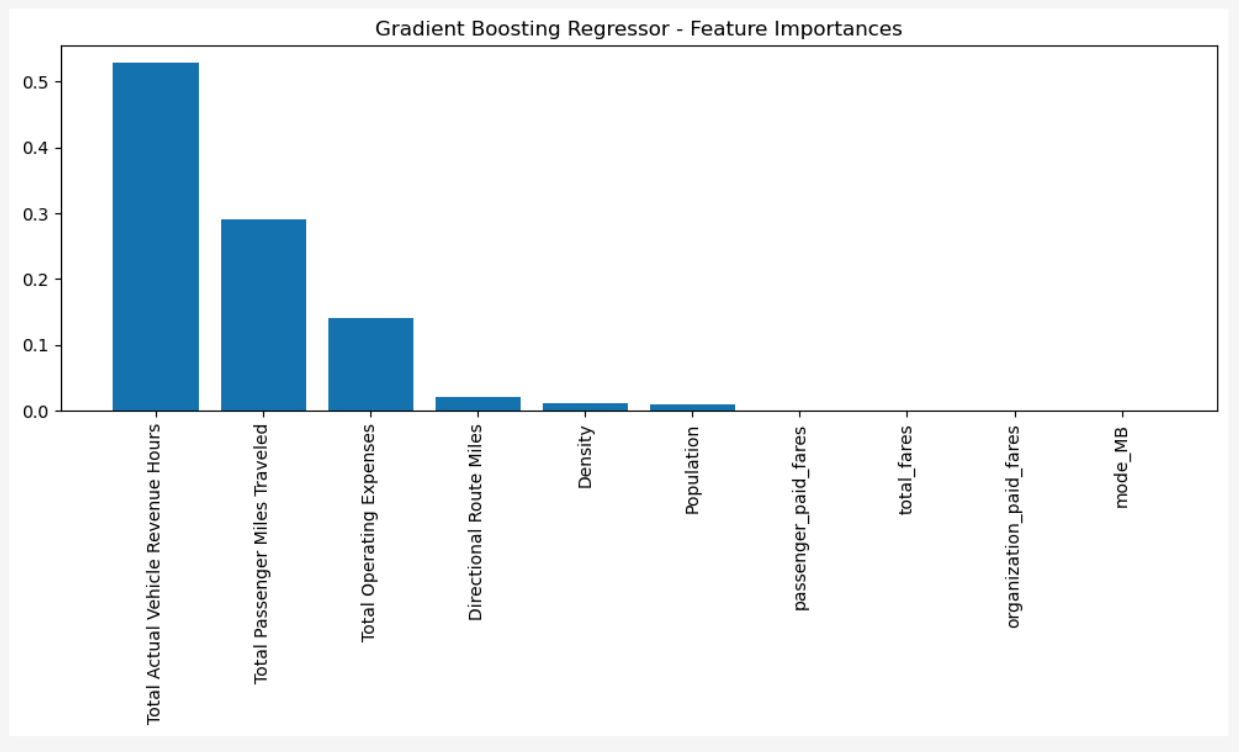
Passenger Miles Traveled (PMT) was used as one of the features to classify ridership levels into categories: **Low, Medium, and High.** These categories were created from **Unlinked Passenger Trips(UPT)** using a method called quantile-based discretization. While **UPT** measures how often passengers board a vehicle, **PMT** shows how far passengers travel.

For Example:

1. Low UPT and Low PMT: A bus route with few passengers and short trips.
2. Low UPT and High PMT: A bus route with low passengers but those passengers are traveling long distances.

The random forest classifier stood out in our analysis, as it had accuracy of 94.86 accuracy and f1 score of 94.8%. The "Total Actual Vehicle Revenue Hours", basically the more hours buses and the transit systems in operation can drive ridership. "Operating Expenses" happened to be an important factor, as spending more on operations, the more people happened to use them. It makes sense hypothetically, because the developed a transit system is, it motivates people to use it as their choice of commute. Features like population and Route miles didn't matter as much, it could be because having a bigger population and route miles doesn't intend to increase ridership. Focusing more on making transit available, spending wisely, and ensuring it's accessible to people is important and directly proportional to ridership.

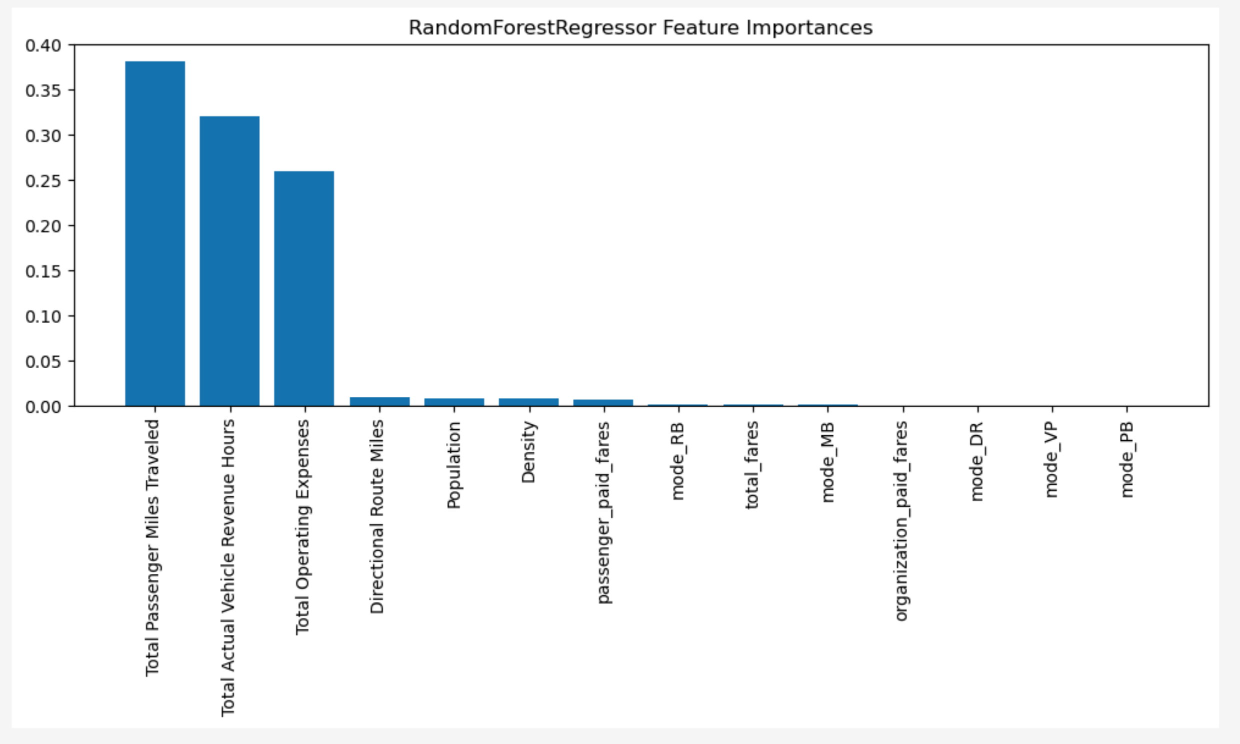
Gradient Boosting Regressor



Gradient Boosting was chosen for this project, because of its ability to capture nonlinear patterns, which are common in real-world (eg., population, operating hours, and miles traveled).

The Gradient Boosting regressor was pretty good when it comes to precision, the R2 value was 0.99 which is nearly 100% of the variance. MAE is around 148, 321 trips might sound a lot, but considering the scale of the data , it's actually quite small. The "Total Actual Vehicle Revenue Hours" seems to be a powerful predictor. We could clearly say that the more time buses are on the road, the more people use them. Intuitively it makes sense, but it's always good to see data back it up. As seen above features like Directional Route Miles and Population had very little effect.

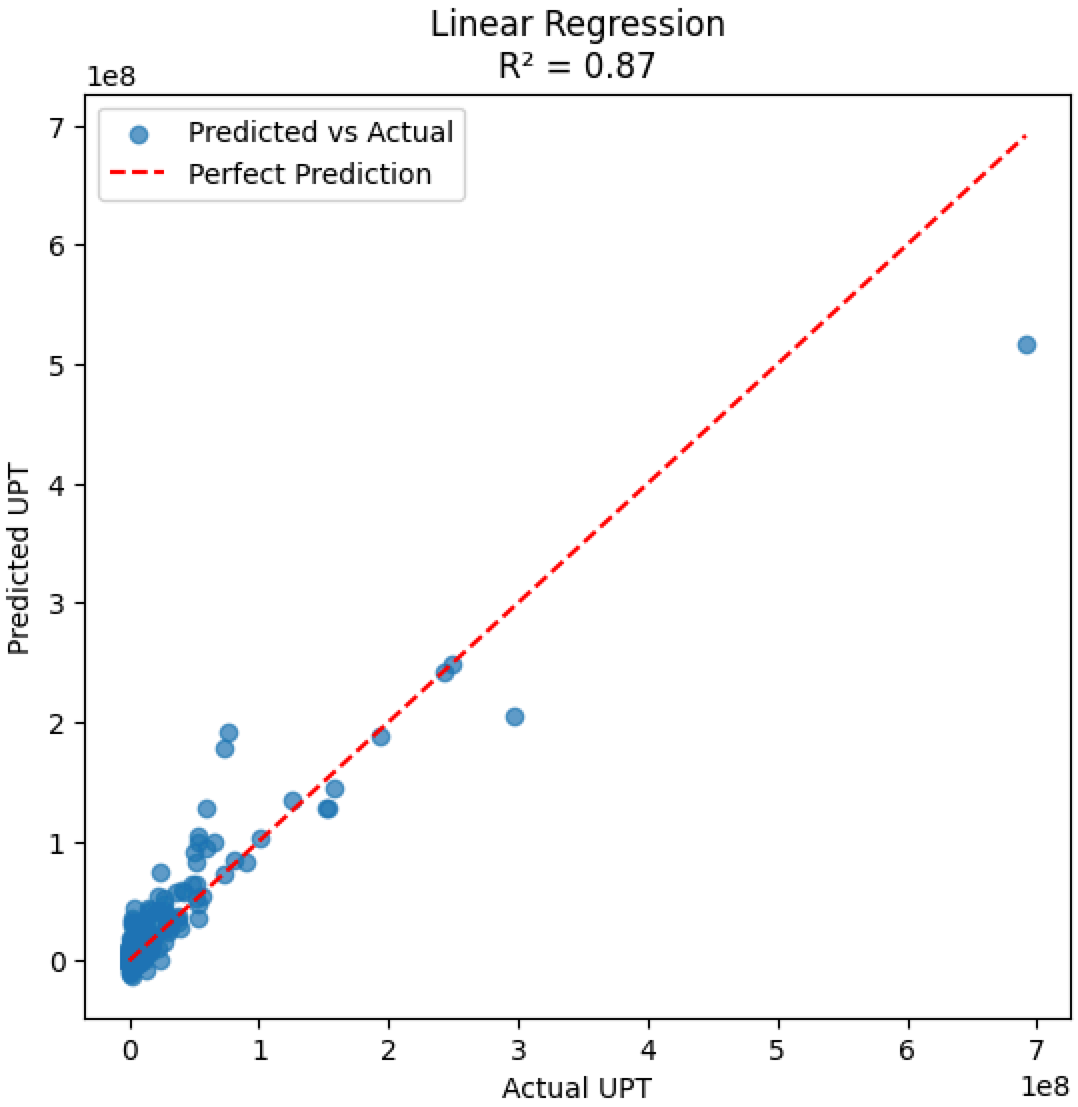
Random Forest Regressor



Random forest regressor was used to understand the factors driving ridership. Using features like passenger miles traveled, vehicle revenue hours, operating expenses. The results showed that **Passenger Miles Traveled** and **Vehicle Revenue Hours** had the highest importance, meaning the increased service availability strongly influenced ridership.

Linear Regression Insights

Linear regression was applied to predict **Unlinked Passenger Trips (UPT)**, a measure of ridership, based on operational, demographic, and safety-related features.



Feature Importance:

```
Mode VOMS                 0.421195
Total Minor Injuries      0.409121
Total Injuries            0.135111
Total Spending            0.019043
Property Damage           0.007934
Primary UZA Population    0.005259
Total Fatalities          0.002337
Total Serious Injuries    0.000000
dtype: float64
```

Key Findings:

Mode VOMS Dominance:

Mode VOMS emerged as the strongest predictor, explaining **42%** of the variance in UPT.

This metric reflects a system's operational capacity—higher numbers of vehicles in service correlate with better coverage, convenience, and accessibility, directly boosting ridership.

For example, a transit system with more buses running during peak hours offers shorter wait times and greater convenience, making it a more attractive option for commuters.

Significance of Safety Metrics:

Total Minor Injuries and **Total Injuries** collectively accounted for **54.4%** of feature importance.

These metrics are strongly tied to public perception of safety. A transit system perceived as unsafe is less likely to attract riders, regardless of its operational efficiency or affordability.

Safety issues, such as frequent minor accidents or injuries, can deter passengers and lead to long-term reputational damage.

Low Impact of Spending:

Total Spending had a relatively small impact on ridership, suggesting that merely increasing financial resources does not guarantee better outcomes.

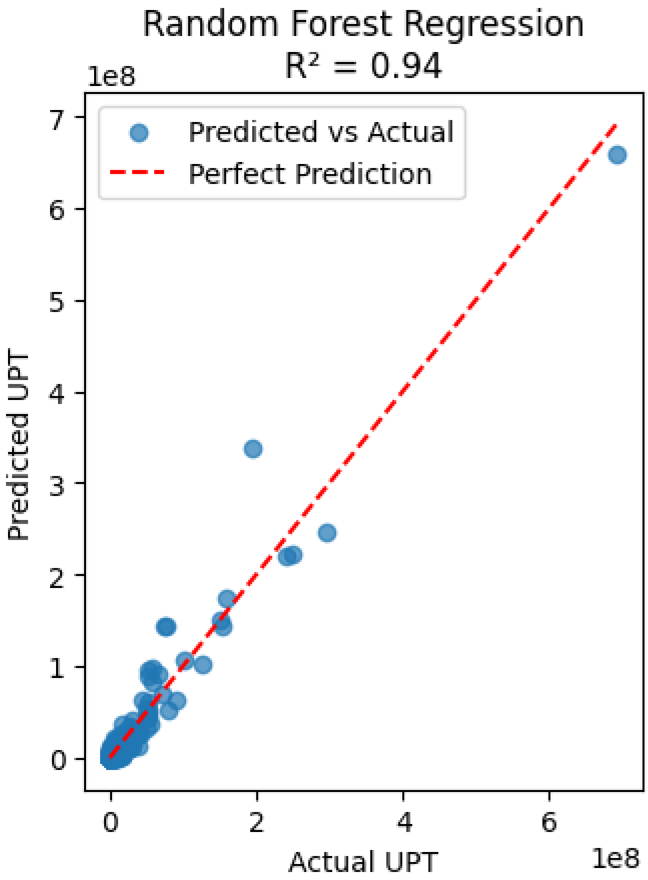
The findings indicate that investments need to be strategically targeted—such as improving service frequency, safety, and reliability—to directly influence ridership.

Model Limitations:

Linear regression struggled to fully capture the non-linear relationships inherent in transit systems, where the interactions between variables like safety, operations, and demographics are more complex than a straightforward linear correlation.

Random Forest Regressor Insights

To overcome the limitations of linear regression, a **Random Forest Regressor** was employed. This machine learning model excelled in handling complex, non-linear relationships between features and ridership.



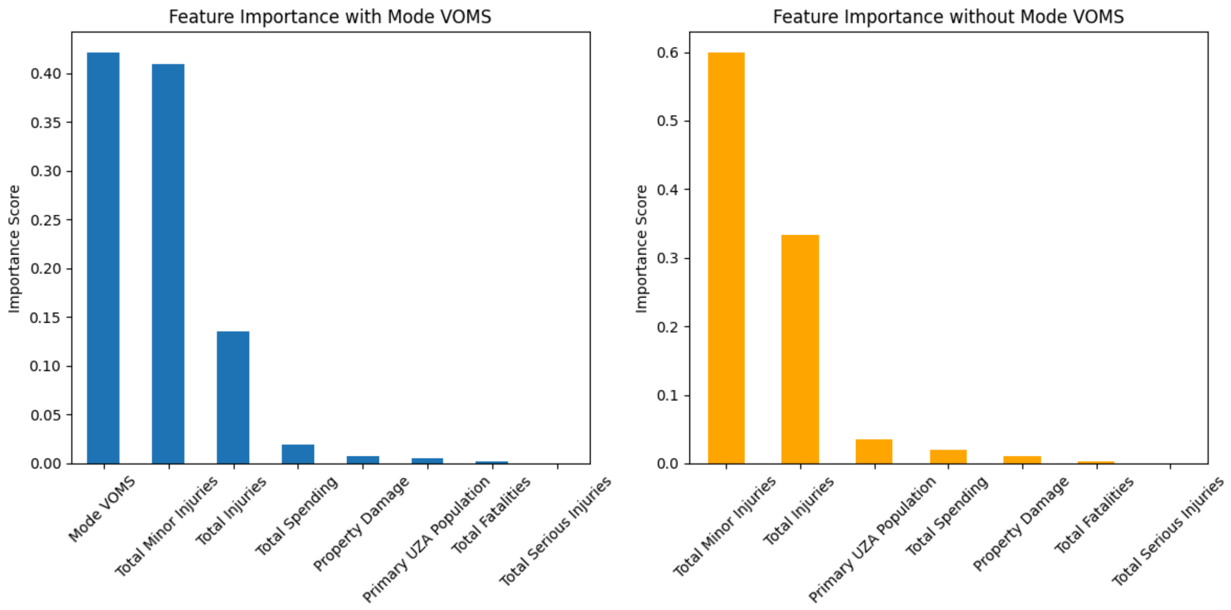
Model R-squared without Mode VOMS: 0.921928076433969

```
Total Minor Injuries      0.599493
Total Injuries            0.333020
Primary UZA Population    0.034406
Total Spending            0.020015
Property Damage           0.010524
Total Fatalities          0.002541
Total Serious Injuries    0.000000
dtype: float64
```

Key Findings:

Performance Metrics:

The Random Forest Regressor achieved an **R² of 0.94**, indicating excellent predictive power and a significant improvement over linear regression.



Feature Importance:

Mode VOMS remained the most critical predictor, reaffirming its direct link to ridership. It highlighted the importance of operational efficiency in attracting passengers.

When **Mode VOMS** was removed from the model, safety-related metrics—**Total Minor Injuries** and **Total Injuries**—became dominant, accounting for **93.2%** of feature importance. This suggests that in the absence of operational data, safety perceptions strongly influence ridership decisions.

Minimal Role of Demographics and Spending:

Features like **Primary UZA Population** (urban population served) and **Total Spending** had minor roles, emphasizing that operational metrics (e.g., vehicle availability and safety) are more immediate and impactful in driving ridership.

Model Robustness:

Even without Mode VOMS, the model maintained a strong R² of **0.92**, demonstrating its adaptability and the importance of other features like safety in explaining ridership patterns.

Key Recommendations

Based on these findings, several actionable recommendations emerge for transit agencies seeking to boost ridership:

Prioritize Operational Metrics:

Increase Mode VOMS**:** Agencies should expand their fleet capacity and optimize vehicle deployment during peak hours to improve service coverage, reduce wait times, and enhance convenience.

This is particularly important in densely populated areas where high-frequency service can significantly reduce reliance on private vehicles.

Enhance Safety Measures:

Focus on Reducing Injuries and Incidents: Agencies should invest in improving driver training, maintaining vehicles, and implementing safety protocols to minimize accidents.

Public campaigns promoting safety improvements can help rebuild trust and encourage ridership.

Evaluate Spending Efficiency:

Agencies should revise how financial investments are allocated. Instead of blanket spending increases, funds should be strategically directed toward initiatives that demonstrably improve operational efficiency and safety.

Leverage Data-Driven Decisions:

Agencies should collect and analyze high-quality, detailed data on operational and safety metrics. Insights derived from this data can guide policy decisions, ensuring that interventions align with rider priorities.

Adopt a Balanced Strategy:

While operational capacity is vital, agencies should also focus on complementary factors like fare adjustments and targeted marketing to create a well-rounded approach.

For example, offering discounted fares during off-peak hours can attract cost-sensitive riders, while emphasizing safety improvements can appeal to hesitant commuters.

Results

Rail

After fitting the models, we can make some general assertions about rail ridership. Accounting for population, which understandably correlates with higher ridership, the two most important factors within the data collected are the fare and the number of vehicles the agency operates. More vehicles means more riders, and higher fare means fewer, which follows general intuition. Metrics like the number of accidents on the transit system are largely not significant, which suggests that the real safety record of a transit system does not affect ridership. That does not necessarily mean that people's perceptions of safety do not affect ridership, only that the hard data is not predictive. Finally, the large value of the constant fitting parameter suggests two possible conclusions. First, that there are other effects not captured by the data in the NTD that could help explain more ridership trends, and second, that base ridership for rail systems is generally high independent of other factors. Some combination of these two could also be true as well.

Bus

This analysis underscores the importance of operational and safety metrics in predicting and boosting bus transit ridership. Investments in operational capacity (e.g., increasing Mode VOMS) and safety improvements should take precedence, while spending strategies need to be reevaluated for effectiveness. By adopting data-driven, targeted interventions, transit agencies can create systems that not only meet the needs of current users but also attract new riders, ultimately enhancing the sustainability of public transportation.