**DATA SCIENCE PROJECT**

**Final Report**

***Aim***

Our goal is to create a predictive model that will identify the most common means of transportation used by Americans living in different census tracts. We attempt to categorize people's preferences for modes of transportation—such as driving, carpooling, taking public transportation, walking, or choosing an alternative and their effects on pollution by examining economic, demographic, and commuting statistics.

***Introduction***

The application of machine learning is an excellent tool for prediction and analysis in the dynamically changing transportation sector, where many factors impact mode selection dynamics. This research uses a variety of statistical and analytical methods to take a rigorous approach to identifying and forecasting future transportation mode choices. Every stage of the procedure is methodical and planned, from the careful loading and preprocessing of the data to the precise application of ANOVA, Pearson's correlation, and feature value assessment. A strong feature selection procedure is ensured by using advanced methods like LASSO regression, mutual information, and PCA, and a thorough prediction framework is produced by experimenting with different classifiers. The ultimate goal of these project is not only to achieve academic excellence but also to offer practical perspectives that will influence the course of urban transportation planning and policy formation in the future.

***Steps Covered***

A. **Loading and Pre-Processing Data**

* After being imported, the data is cleaned up to remove irregularities and standardize formats before being ready for analysis.
* To ensure the integrity of the dataset, missing data handling techniques including imputation and exclusion are used.
* Data merging is conducted when multiple sources are involved, and the final dataset is structured for model ingestion.
* Files loaded and merged

B. **Pearson's Correlation Coefficient**

* The purpose of this statistical measure is to evaluate the direction and strength of a linear relationship between two continuous variables.
* A strong positive correlation is indicated by a coefficient near 1, while a strong negative correlation is indicated by a number near -1.
* Finding variables that have a strong correlation with the target variable is crucial for feature selection.

The below summary analyzes a script that explores the impact of feature correlation on the accuracy of a logistic regression model predicting a variable named 'TransportationMode'.

Here's a step-by-step breakdown:

1. Initial Model Accuracy: The script starts by establishing a baseline. It trains a logistic regression classifier using all features (except the first two columns and the target variable) from a dataset named merged data. The accuracy of this model on the test set is calculated and stored.
2. Correlation Matrix Calculation: Next, a correlation matrix is computed for the dataset. This matrix measures the linear relationship between pairs of features.
3. Dimensionality Reduction Through Correlation Thresholding: The script then iterates over various correlation thresholds from 0.1 to 0.9. For each threshold:
   * Features that are highly correlated (beyond the set threshold) with other features are identified and marked for removal.
   * The logistic regression model is retrained using this reduced set of features.
   * The accuracy of the model with the reduced feature set is calculated.
4. Accuracy vs. Threshold Visualization: After iterating through the thresholds, a bar graph is plotted, showing how the model's accuracy changes with different correlation thresholds.
5. Final Feature Selection: In the end, the script identifies and prints out the features that are highly correlated beyond a specified threshold (0.4 in this case) and are candidates for removal.

Key Observations:

* 1. The approach effectively demonstrates how reducing feature dimensionality (by removing highly correlated features) affects model accuracy.
  2. The visualization (bar graph) provides a clear view of the trade-off between model complexity (number of features) and performance (accuracy).
  3. The final output lists the features to be removed based on the chosen correlation threshold, aiding in model simplification without significantly compromising accuracy.

In step 3 of the analysis, a detailed examination was conducted to understand how varying correlation thresholds impact the accuracy of a logistic regression model. This was accomplished by creating a bar graph which visually represented the relationship between different correlation thresholds and the corresponding accuracy levels of the model. The thresholds ranged from 0.1 to 0.9, and for each threshold, the process involved identifying and excluding features from the dataset that were highly correlated (beyond the set threshold) with other features. This method of feature selection is crucial in reducing the complexity of the model by eliminating redundant or highly interdependent variables, which can lead to overfitting or multicollinearity issues.

Upon analyzing the graph, it was observed that a threshold of 0.4 yielded the highest accuracy. This implies that at this threshold, the balance between retaining informative features and removing redundant ones was optimal for the performance of the logistic regression model. Therefore, the decision was made to use this threshold of 0.4 as a reference point for future steps in the analysis. This threshold serves as a guiding criterion for feature selection, ensuring that the model remains robust and efficient without compromising on its predictive power.

C. **ANOVA (Analysis of Variance)**

-> ANOVA tests the hypothesis that the means of two or more groups are different, which can reveal important categorical variable effects.

-> It partitions variance into components attributable to various sources for further analysis.

-> This method is particularly useful when comparing the impact of different categorical independent variables on a continuous dependent variable.

-> Did this and carried out 4 most important columns out of it.

D. **Feature Importance (Machine Learning Models)**

-> Various machine learning models, such as decision trees and ensemble methods, provide scores indicating the importance of each feature in prediction.

-> This helps in prioritizing which variables have the most predictive power and should be included in the model.

-> Feature importance is crucial for simplifying models and improving interpretability without compromising on performance.

-> Performed this and taken away importance of every column for Random Forest.

E. **Mutual Information**

-> Mutual information quantifies the amount of shared information between variables, useful for detecting non-linear relationships.

-> It is a non-parametric method, making it versatile for various types of data, including continuous and categorical.

-> Unlike correlation, mutual information can capture any kind of dependency between variables, not just linear.

-> Calculated the importance and used that for dimensional reduction.

F. **Principal Component Analysis (PCA)**

-> PCA is a dimensionality-reduction technique that transforms a large set of variables into a smaller one that still contains most of the information.

-> It identifies the principal components that account for the maximum variance in the data, often revealing underlying patterns.

-> PCA can also improve model efficiency by reducing the number of features, which can alleviate the curse of dimensionality.

-> Did this and noticed the behaviours of n components of PCA over final accuracy.

G. **Running Different Classifiers**

-> Multiple classifier models are employed to identify the best predictive model for the dataset based on performance metrics.

-> Classifiers ranging from simple logistic regression to complex ensemble methods are compared.

-> This process includes hyperparameter tuning and cross-validation to ensure the robustness and reliability of the predictive models.

-> Used following models to compare results

1. Logistic Regression

2. Lasso Logistic Regression

3. Gaussian Naïve Bayes

4. Random Forest, out of all 4 random forests have the best accuracy.

***Conclusion***

In conclusion, the project successfully navigated the complexities of predicting transportation mode choices by employing a robust analytical framework. Through comprehensive data preprocessing, we ensured a clean and reliable dataset. Techniques like Pearson's correlation, ANOVA, and mutual information provided deep insights into the relationships between variables. LASSO regression and PCA were instrumental in feature selection, distilling the data to its most informative elements. The deployment of various classifiers demonstrated the nuanced differences in predictive capabilities. Out of all the classifiers Random Forest works best for the merged data. The culmination of these methodologies not only met the project's aim but also paved the way for future enhancements in the domain of transportation modelling, offering a substantial contribution to the predictive analytics field and potential applications in urban planning and policy making.