Commute Time Regression Analysis: Feature Importance and Predictive Modeling with Census Data

Francesco Giuseppe Gillio
Department of Computer Science
Polytechnic University of Turin

***Abstract*—This research presents a comprehensive regression analysis to predict commute times via census data. The analysis investigates the main socio-economic variables that influence commuting patterns to develop and refine a predictive model for practical applications. The study evaluates a dataset with over 130,000 observations, enforcing rigorous preprocessing techniques to drop irrelevant or redundant entries and normalize feature vectors to match the requirements of some regression models. Several regression models, including Random Forest, Decision Tree, and Support Vector Regression (SVR), undergo evaluation to extract the most relevant features and assess the dataset's predictive power. Among these models, Random Forest shows superior performance, achieving the highest coefficient of determination (R^2). Core predictors, such as departure and arrival times, emerge as critical determinants of commute time. Furthermore, the introduction of the Commute Time to Work (CTTW) feature, i.e. the time difference between departure and arrival, enhances the performance of the models, with emphasis on the Random Forest model, which attains an R^2 value approaching 0.997 post-processing. Broader socio-economic variables, such as income and education, show minimal influence on commute times, reflecting their statistical independence from commuting patterns. By refining the dataset to prioritize temporal variables, the analysis improves model accuracy, offering a more efficient and reliable tool for predicting commute times.**

GitHub: https://github.com/305909/regana

I. Introduction

This research undertakes a rigorous analysis of comprehensive census data to detect the main determinants influencing commute time to work. The primary objective involves developing a statistical regression model [4] capable of predicting the commute time for a random subject in terms of minutes, as floating number between 0 and 200. The model prioritizes statistically significant variables, systematically excluding those with minimal relevance, thus enhancing predictive accuracy and minimizing extraneous bias [5].

The analysis evaluates a robust observational dataset, that branches into: a development set comprising 104,642 observations, each containing the target variable, and an evaluation set of 26,159 observations, which excludes the target variable. The primary difference between these sets lies in the presence of the target column, which corresponds to the commute time to work.

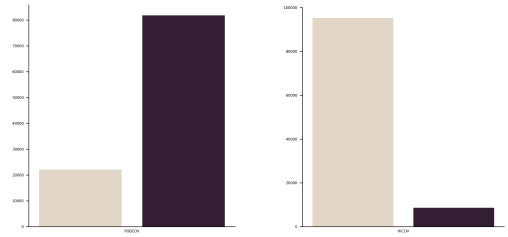

Fig. 1: The observations distributions from public health coverage recode (PUBCOV) and the health insurance coverage recode (HICOV).

Initial examination over the development set showcases that a unique code identifies each observation, ensuring structural integrity. Nevertheless, 711 observations within the set show identical values across columns, suggesting errors in data collection (duplicate data). This subset represents less than 1% of the development set. On the other hand, the development set shows no evidence of measurement errors, as no columns within the dataset lack values. By eliminating duplicate data, the remaining sample of observations offers a reliable foundation for conducting analyses to uncover statistical inferences and support evidence-based decision-making.

II. Methodology

A. Data Preprocessing

Data preprocessing constitutes a pivotal stage in the data mining workflow [5], entailing additional evaluations and statistical calculations to adapt the data into more suitable formats for subsequent analysis. This section outlines some of the processes undertaken to drop redundant census data that compromise the predictions of the regression model and reduce the dimensionality of the dataset. After the duplicate entries removal procedure, the development set comprises 103,931 observations (rows), each with 25 census characteristics (columns). These features categorize into: 22 categorical features, 3 numeric features. Note that, despite the categorical nature of some attributes, each attribute value in the development set appear expressible as an integer or floating number, a valuable property for subsequent phases.

The numeric features include information on total income, usual weekly working hours over the past 12 months, and commute time to work (the target variable). The categorical features encompass various socio-economic parameters. Among these features, some show a non-homogeneous trend in the observations distribution, compromising the statistical significance of the sample when taken into consideration by the regression model. In this context, Figure 1 shows the observations distribution relating the health insurance coverage recode and the public health coverage recode.

To underline the hypothesis above, take into consideration the census attribute relating to health insurance coverage. This attribute shows a non-homogeneity in the distribution of values (as shown in Figure 1 on the right), with nearly every observation in the development set indicating a lack of health insurance coverage. As consequence, omitting this attribute from the regression model risks leading to disadvantageous predictions for individuals with health insurance coverage. To validate the previous hypothesis, the study researches the correlation of the attribute with respect to the target variable, to understand its actual relevance in the forecasting process. To this end, and taking advantage of the numerical nature of each attribute value in the development set, the analysis trains several regression models, including Random Forest [1], Decision Tree [5], Linear Regression [5], Support Vector Regression (SVR) [2], and K-Neighbors [3]. This preliminary pre-processing step aims to assess the significance of each census data point in predicting the target variable. Since Linear Regression, SVR, and K-Neighbors models require data normalization to fit into a standard range, pre-processing involves normalization of the feature vectors.

To assign a score, the analysis implements a partitioning process over the development set by means of the train-test-split function, which partitions the set into random train and test subsets, assigning 75% of the observations to the train set and 25% to the test set (therefore respecting the same proportion shown between the development set and the evaluation set). Table I shows, for each model, the resulting coefficient of determination.

Model	R^2 Mean	R^2 Standard Deviation
Random Forest	0.98274	0.00543
Decision Tree	0.96979	0.00864
SVR	0.22344	0.00786
Linear Regression	0.20265	0.03214
K-Neighbors	0.06584	0.00512

TABLE I: Comparison of regression model performance via R^2 Mean and R^2 Standard Deviation.

As shown in Table I, the Random Forest model [1] already provides a significant R^2 value, an indicative measure of the proportion of the variation in the dependent variable resulting predictable from the independent variables.

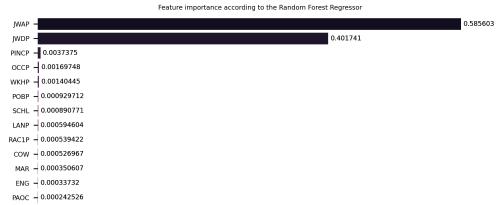

Fig. 2: Feature importance via Random Forest Regressor model [1] (attributes subset with importance value higher than 0.025%).

Thus, proceeding with the attribute relevance evaluation phase, the analysis attempts to answer the question: which census attributes contribute most to the prediction? Figure 2 showcases the answer. The attribute evaluation process reveals that, among the comprehensive census data, only the attributes relating to the time of arrival at work and the time of departure for work show statistical relevance in predicting the commute time to work, with an attribute importance value of 0.589 and 0.399, respectively. The above hypothesis, a natural logical deduction at first glance, thus also shows an observational character according to statistical evidence. Furthermore, the analysis evidences the statistical independence between the research object and the census data that most influence social stratification into separate classes, such as income, occupation, education, housing and various others. These features reveal importance values lower than 0.5%, the minimum threshold for classifying a cause-and-effect series as deterministic and non-random. Thus, reducing the comprehensive census data to only the features statistically significant in the prediction process of the target variable leads the development set to contain only two features per observation:

- Time of arrival at work (JWAP), with a range of values from 1 to 285.
- Time of departure for work(JWDP), with a range of values from 1 to 150.

The attribute values of both results representative of a time slot, mutually exclusive with the others in the same column. For example, as regards the JWDP column (time of departure for work), the value 1 represents the time slot from 12:00 a.m. to 12:29 a.m., while the value 2 represents the consecutive timeslot, i.e. from 12:30 a.m. to 12:59 a.m. The times slot continues until reaching the last time slot, i.e. the one from 11:30 p.m. to 11:59 p.m. Note that the time slots for each column appear not equally spaced in time, i.e. some time slots result of 30 minutes, others of 5 minutes, and so on.

B. Feature Transformation

From the pre-processing data evaluation, the analysis proposes a feature transformation process. Each observation in the development set shows specific values of departure time and arrival time. A simple correlation between these measures leads to a metric of commute time to work.

Considering the association between the attribute value and the relative time slot, the process transforms each attribute value in terms of minutes and then in terms of interval. Table II shows some examples to understand the transformation process.

(JWDP) Time Slot	Time Slot in Minutes	Mean Time
12:00 a.m. - 12:29 a.m.	0 min - 29 min	14.5 min
12:30 a.m. - 12:59 a.m.	30 min - 59 min	44.5 min
1:00 a.m. - 1:29 a.m.	60 min - 89 min	74.5 min
1:30 a.m. - 1:59 a.m.	90 min - 119 min	104.5 min
(JWAP) Time Slot	**Time Slot in Minutes**	**Mean Time**
12:00 a.m. - 12:04 a.m.	0 min - 4 min	2 min
12:05 a.m. - 12:09 a.m.	5 min - 9 min	7 min
12:10 a.m. - 12:14 a.m.	10 min - 14 min	12 min
12:15 a.m. - 12:19 a.m.	15 min - 19 min	17 min

TABLE II: Conversion process for time slots into minutes.

The first column in Table II represents the attribute value, as shown in the development set, the second column represents the relevant time slot. Instead, the third column shows the time slot into minutes conversion (counting from 12:00 a.m. as a value of 0 minutes and moving forward) and the fourth column shows the average value across the time slot thresholds. The analysis then implements a commute time metric (CTTW) in the development set, subtracting, for each observation, the JWAP column value from the JWDP column value, both with the above minute conversion. Table III shows an example of such implementation for a random observation within the development set.

JWDP	into min	JWAP	into min	CTTW	JWMNP
49	452 min	94	482 min	30	30

TABLE III: Feature transformation process.

The first column in Table III shows the actual value of the observation's JWDP attribute, i.e. the original value as shown in the development set, as does the third column for the JWAP attribute. The second and fourth columns show the conversion of the original values into minutes. The last columns represents the commute time to work (CTTW) value from the feature transformation process and the effective commute time to work (JWMNP) value, i.e. the target variable, of the development set, respectively.

For evaluation purposes, the analysis implements again the Random Forest model to re-evaluate the importance of census features in predicting the target variable, but at this stage with the addition of the novel parameter. Figure 3 showcases the result. The figure shows how the introduction of the CTTW feature, from the future transformation process, dominates over the others. With an importance value of 0.933, the CTTW attribute reduces the value of the other attribute, a consequence of the high predictive characteristic it covers in the regression model.

Fig. 3: Feature importance via Random Forest Regressor model [1] and commute time to work (CTTW) parameter (attributes subset with importance value higher than 0.025%).

To validate both the introduction of the CTTW feature and the elimination of statistically irrelevant census features in the prediction of the target variable, the analysis retests Random Forest [1], SVR [2] and K-Neighbors [3] regression models, with the same partitioning process over the development set by means of the train-test-split function. Table IV provides the results, each showing a clear increase in the coefficients of determination values with respect to the previous R^2 scores (Table I). The preprocessing outcomes appear quite promising, the analysis thus concludes the current phase, which results in the reduction of the development set to 103,931 observations (rows), each with four features (columns): the time of departure for work (JWDP), the time of arrival at work (JWAP), the commute time to work (CTTW), and the target variable, i.e. the original commute time to work (JWMNP).

Model	R^2 Mean	R^2 Standard Deviation
Random Forest	0.98274	0.00543
After preprocessing	0.99691	0.00023
SVR	0.22344	0.00786
After preprocessing	0.97321	0.00416
K-Neighbors	0.06584	0.00512
After preprocessing	0.99597	0.00041

TABLE IV: Coefficients of determination for regression models before and after preprocessing.

C. Model Selection and Hyperparameters Tuning

The model selection emerges as a constructive consequence of the pre-processing phase, which instead represents the most laborious phase of the research. Therefore, taking up the previous considerations, the analysis selects the Random Forest as the most promising regression model, with a discrimination coefficient of 0.997, i.e. the maximum value among the others, as shown in the following ranking. In order to research the most predicting hyperparameter configuration, the analysis partitions again the development set into random train and test subsets, assigning 75% of the observations to the train set and 25% to the test set. Then, implementing the grid search cross-validation module, emerges the optimal configuration of the Random Forest [1] hyperparameters, with 155 number of trees in the forest. Figure 4 shows the result of the model prediction, a promising result ending the analysis process.

Model	R^2 Mean	R^2 Standard Deviation
Random Forest	0.99691	0.00023
Decision Tree	0.99675	0.00032
K-Neighbors	0.99597	0.00041
SVR	0.97321	0.00416
Linear Regression	0.21589	0.04149

TABLE V: R^2 Mean and Standard Deviation for different regression models.

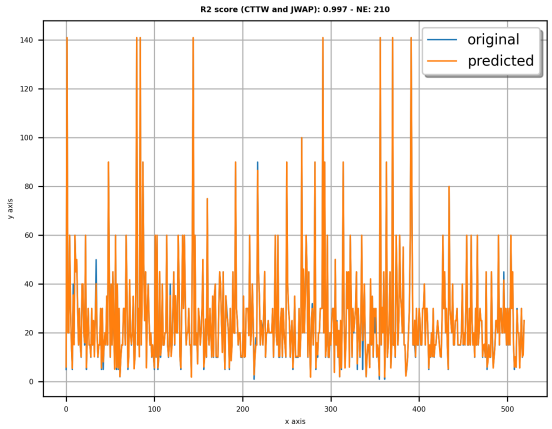

Fig. 4: R^2 score.

III. Conclusion

The report's analysis aims to detect census data showing statistical observational evidence in predicting the commute time to work. Since census data provides building blocks in the inequality investigation within a population, the analysis provides a practical detection into the fairness of algorithms implementing them, in order to understand the impact of socio-economic disparities and guide evidence-based policy decisions.

The analysis selects a Random Forest [1] regression model, which provides a prediction process combining the outputs of multiple decision trees, each structuring a logical chain of if-then conditional nodes. The model, by a global feature importance evaluation, i.e. an estimation of the features importance in predicting the target variable, shows the statistical irrelevance of some attributes in the conditional nodes optimization process, in terms of purity by coefficient of determination (Figure 2). These features, indeed, appear with importance values lower than 0.5%, the minimum threshold for classifying a cause-and-effect series as deterministic and non-random. Furthermore, following the dimensionality reduction process, which removes such irrelevant attributes from the development set, each regression model shows a clear increase in the level of predictiveness (Table V), underwriting also a redundancy property.

The analysis, then, proves the statistical independence between the research object and the census data that most influence social stratification into separate classes, such as income, occupation, education, housing and various others, since each shows a property of irrelevance and redundancy in the prediction.

References

[1] Leo Breiman. Random Forests. Machine Learning, 2001.

[2] Alex J. Smola, Bernhard Schölkopf. A Tutorial on Support Vector Regression, 2004.

[3] Thomas Cover, Peter Hart. Nearest Neighbor Pattern Classification, 1967.

[4] Douglas C. Montgomery, George C. Runger. Applied Statistics and Probability for Engineers. John Wiley & Sons, New York, 2013.

[5] Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques. 3rd Edition, Morgan Kaufmann Publishers, Burlington, 2011.