A Machine Learning Approach for Assessing Labor Supply to the Online Gig Economy

Esabella Fung[1]
[1]Northwestern University, Evanston, IL 60208 USA
Corresponding author: Esabella Fung (e-mail: esabellafung1@gmail.com).

Abstract

The online labor market, comprised of companies such as Upwork, Amazon Mechanical Turk, and their freelancer workforce, has expanded worldwide over the past 15 years and has changed the labor market landscape. Although qualitative studies have been done to identify factors related to the global supply of the online labor market, few data modeling studies have been conducted to quantify the importance of these factors in this area. This study applied tree-based supervised learning techniques, decision tree regression, random forest, and gradient boosting, to systematically evaluate the online labor supply with 70 features related to climate, population, economics, education, health, language, and technology adoption. To provide machine learning explainability, SHAP, based on the Shapley values, was introduced to identify features with high marginal contributions. The top 5 contributing features indicate the tight integration of technology adoption, language, and human migration patterns with the online labor market supply.

Keywords: business, boosting, commerce and trade, digital divide, economics, ensemble learning, globalization, machine learning, random forest, social factors, statistical learning, sharing economy, trade

1. Introduction

The gig economy is a system where people provide short-term goods and services. An example of the gig economy is freelance work, which can be completed independently or under a more prominent company acting as an intermediary platform. The online labor market is part of this gig economy, and it facilitates exchanges of all virtual services: software development, multimedia content, translation, and marketing support.

From 2016 to 2021, the online labor market grew by over 10% annually.[1] This corroborates the idea that globalization increases the exchange of products or services across countries.[2] The online labor market has gained greater visibility with the growth of internet accessibility and the COVID-19 pandemic from 2020, when there was a rise in the need for virtual services with limited face-to-face contact worldwide.[3] Migration rates, influenced by economic opportunity and labor policies, are critical to understanding workforce mobility in digital economies. Understanding the factors influencing labor supply can help policymakers design targeted programs that improve job accessibility, particularly for migrants and low-income populations.

Current labor market models often ignore technological and migratory factors, which are critical in digital economies. To better understand this growth, machine learning techniques were used to evaluate features associated with labor supply in the online labor market. Machine learning models were created using data on the online labor market activities, climate, population, economics, education, health, language, and technology adoption over 5 years. With these data points, 6 models, multiple linear regression, ridge, LASSO, decision tree regression, random forest, and gradient boosting, were trained, validated, and evaluated for factors related to the online labor market. We apply SHAP values to interpret machine learning predictions of labor supply and identify technology adoption (mobile subscriptions) as a key predictor of online labor participation.

2. Data

Measurement of online labor market activity is based on the Online Labour Index collected from the Online Labour Observatory created by the International Labour Organization and the Oxford Internet Institute. These data on the online labor market supply were collected by examining the application programming interfaces from digital platforms or downloading the web user interface of 5 online labor platforms: Amazon Mechanical Turk, Upwork.com, Freelancer.com, Guru.com, and Peopleperhour.com. These were the top 5 platforms representing at least 70% of the total online labor platform traffic, according to Alexa.com, when the index was created in 2016. As of 2023, Freelancer.com is the only one of these 5 platforms that supports multiple languages, while the

[1] Stephany, F., Kässi, O., Rani, U., and Lehdonvirta, V. (2021). Online Labour Index 2020: New ways to measure the world's remote freelancing market. Big Data & Society, 8(2), 205395172110432. https://doi.org/10.1177/20539517211043240

[2] Friedman, T. L. (2005). The world is flat: A brief history of the twenty-first century. Macmillan.

[3] Tan, Z. M., Aggarwal, N., Cowls, J., Morley, J., Taddeo, M., and Floridi, L. (2021). The ethical debate about the gig economy: A review and critical analysis. Technology in Society, 65(101594), 101594. https://doi.org/10.1016/j.techsoc.2021.101594

other 4 platforms only support English. There are 992170 rows of daily online activities from June 16, 2017, to December 31, 2022.[4]

The data gathered by the Online Labour Observatory was determined by looking at the number of open vacancies or projects for clients who wanted to hire workers. The occupation types of these vacancies were classified using machine learning based on specific keywords in the vacancy title or description. Employer country was estimated by taking a sample of vacancies from the two platforms, Guru.com and Upwork.com, that show country information, then weighing the samples to reflect all platforms' occupation distribution. The observations found in 203 countries were separated based on country, number of workers, number of projects, and occupation. Occupation was divided into six categories: clerical and data entry, creative and multimedia, professional services, sales and marketing support, software development and technology, and writing and translation. The online economy data were combined with data from the International Monetary Fund,[5] United Nations Development Programme,[6] Oxford COVID-19 Government Response Tracker,[7] United Nations Department of Economic and Social Affairs,[8] World Factbook,[9] Ethnologue,[10] Area Database of the Global Data Lab,[11, 12] World Bank,[13] and United Nations High Commissioner for Refugees[14] to create a dataset for modeling.

[4] Stephany, F., Kässi, O., Rani, U., and Lehdonvirta, V. (2021). Online Labour Index 2020: New ways to measure the world's remote freelancing market. Big Data & Society, 8(2), 205395172110432. https://doi.org/10.1177/20539517211043240

[5] International Monetary Fund. (n.d.). Climate Change Indicators Dashboard [Data set]. Retrieved August 30, 2023, from https://climatedata.imf.org/pages/access-data

[6] UNDP. (2022). Human development report 2021-22: Uncertain times, unsettled lives: Shaping our future in a transforming world. https://policycommons.net/artifacts/3533799/humandevelopment-report-2021-22-human-development-reports/4335012/

[7] Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron Blake, E., Hallas, L., Majumdar, S., and Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker).

[8] World Population Prospects 2022. (2023). UN DESA Publications. https://desapublications.un.org/publications/world-population-prospects-2022-summaryresults

[9] The World Factbook. (2023). Cia.gov. https://www.cia.gov/the-world-factbook/

[10] Eberhard, D., Simons, G., and Fennig, C. (2019). Ethnologue: Languages of the World. Ethnologue: Languages of the World.

[11] Jeroen, Permanyer, and Smits, I. (2019). The Subnational Human Development Database. Scientific Data, 6(1). https://doi.org/10.1038/sdata.2019.38

[12] Smits, J. (2016). GDL Area Database. Sub-national development indicators for research and policy making. GDL Working Paper, 16–101.

[13] Education statistics - all indicators. (n.d.). Worldbank.org. Retrieved August 30, 2023, from https://databank.worldbank.org/indicator/SE.SCH.LIFE?id=c755d342&report_name=EdStats_Indicators_Report&populartype=series

[14] UNHCR. (n.d.). Unhcr.org. Retrieved August 30, 2023, from https://www.unhcr.org/refugeestatistics/download/?url=HF39gS

2.1 Features

Seventy features related to schooling, gross national income, labor force, technology subscriptions, tax rates, unemployment rates, inflation rates, migration, country expenditures, life expectancy, population, natural disasters, number of speakers of top languages, refugees, and COVID-19 were categorized for correlation analysis in this study (**Appendix A**).

Data were collected in Arabic, Chinese, English, French, Hindi, Russian, and Spanish since they were the official languages designated by the United Nations. The number of first language (L1) and second language (L2) speakers and the Expanded Graded Intergenerational Disruption Scale (EGIDS)[15] of these languages across all countries was collected from Ethnologue to test the impact of languages on the growth of the online labor market.[16] The categorical EGIDS data were converted to numerical values in this study, similar to an approach taken in modeling language adoption in the digital space.[17] Data on COVID-19 and internet users were also collected to test the potential increase in online labor market supply with increased isolation and internet users. For some attributes, data were missing in the latter years of 2021 or 2022. Because the data was collected from public sources, potential biases or missing regions in the dataset may result from underreporting in low-income countries, limiting the generalizability of results to regions with weaker infrastructure. In addition, countries with limited recognition may have missing or double-counted data. In these cases, forward fill was used to fill gaps in data since trends in attributes with missing data remained relatively stable.

2.2 Correlation

The combined dataset was evaluated for high correlation among features. For linear models, if multiple features have high correlations with each other, the model may become inaccurate and generate a larger error of sum squares.[18] The existence of correlated features could also increase the error of tree-based models by reducing the effectiveness of features that balance the other features with heavy correlation.[19] The dataset was analyzed using Pearson correlation and Spearman's ranking correlation to identify heavily correlated features. The Pearson correlation coefficient was used to

[15] Lewis, M. P., and Simons, G. F. (2022, September 15). Assessing endangerment: Expanding fishman's GIDS. SIL International. https://www.sil.org/resources/archives/44183

[16] Eberhard, D., Simons, G., and Fennig, C. (2019). Ethnologue: Languages of the World. Ethnologue: Languages of the World.

[17] Kornai, A. (2013). Digital language death. PLOS One, 8(10), e77056. https://doi.org/10.1371/journal.pone.0077056

[18] Bühlmann, P., Rütimann, P., van de Geer, S., and Zhang, C.-H. (2013). Correlated variables in regression: Clustering and sparse estimation. Journal of Statistical Planning and Inference, 143(11), 1835–1858. https://doi.org/10.1016/j.jspi.2013.05.019

[19] Tolosi, L., and Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics (Oxford, England), 27(14), 1986–1994. https://doi.org/10.1093/bioinformatics/btr300

evaluate collinearity between 2 features. A negative Pearson correlation coefficient indicates an inverse relationship between two features, whereas a positive Pearson coefficient represents a positive relationship between two features. This Pearson correlation is based on the assumption of continuous features following a linear relationship and assumes each observation has a pair of values, without accounting for outliers. The formula for the Pearson correlation coefficient r is computed using the equation:

$$r = \frac{n(\Sigma x_i y_i) - (\Sigma x_i)(\Sigma y_i)}{\sqrt{[n\Sigma x_i^2 - (\Sigma x_i)^2][n\Sigma y_i^2 - (\Sigma y_i)^2]}} \tag{1}$$

where n is the number of pairs and x, y are individual points.

For strongly correlated features paired with coefficients with absolute values above 0.7, clusters of features were identified using hierarchical clustering. Based on hierarchical clustering with Ward's linkage, a dendrogram that displays the high collinearity based on the Pearson correlation coefficient was created (**Figure 1**). In each cluster, one feature was chosen to represent the cluster while the rest of the features were removed from the dataset in this feature selection process.

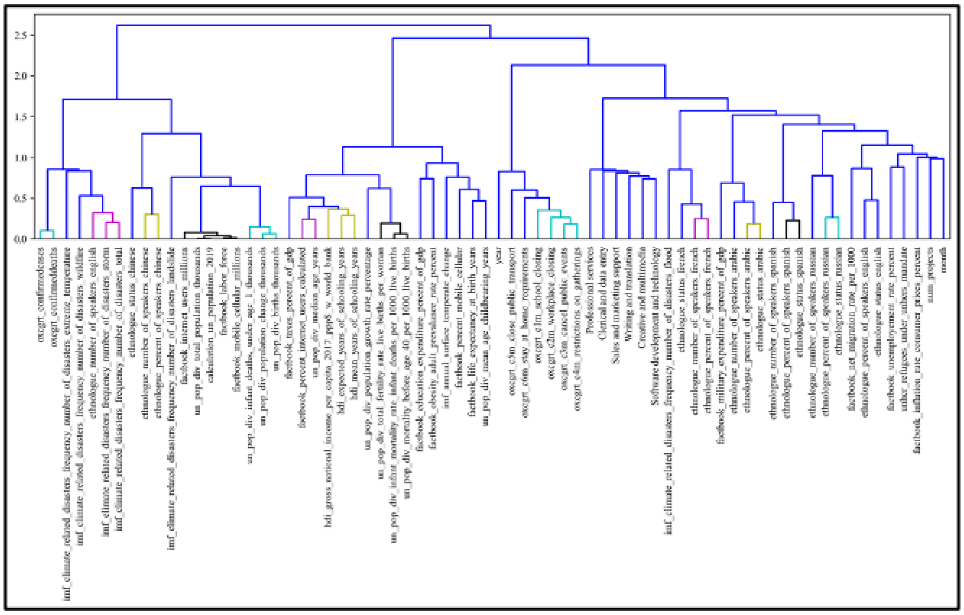


FIGURE 1. Dendrogram based on clustering with Pearson coefficient.

Unlike the Pearson correlation coefficient, which measures the linear relationship between 2 features, the Spearman correlation coefficient measures the monotonic relationship between 2 features. Spearman correlation does not assume features are continuous but assumes that data is ordinal. The Spearman correlation coefficient is calculated with the formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \qquad (2)$$

where n is the number of observations and di is the difference between ranks of each observation. A dendrogram was created based on hierarchical clustering of highly correlated features with Spearman correlation coefficients with an absolute value above 0.7 (**Figure 2**).

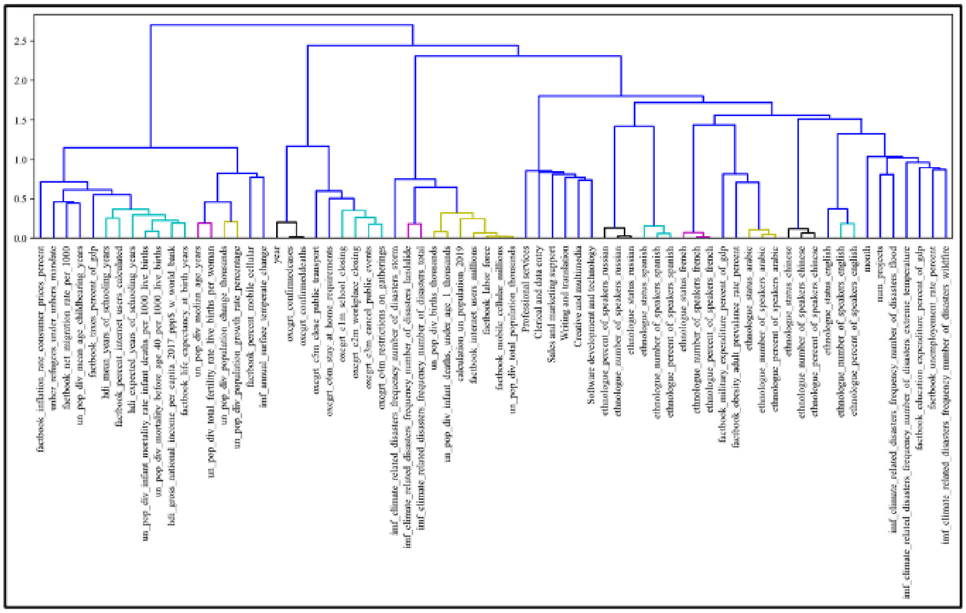


FIGURE 2. Dendrogram based on clustering with Spearman coefficient.

Representative features were chosen with the intent of maximizing the number of different features in the dataset. For example, the dendrogram based on clustering with Pearson correlation coefficients shows a cluster with two features: the percentage of Arabic speakers and the Ethnologue status of Arabic. Since the dataset already contained a feature with the number of Arabic speakers that do not have a high Pearson coefficient with other features, the percentage of Arabic speakers was eliminated from the dataset due to its similarity with existing variables. In addition, we aimed to choose the same representative features between the datasets with feature selection based on Pearson and Spearman coefficients. The dendrogram with clustering based on Spearman coefficients shows a cluster with the number of Arabic speakers, the percent of Arabic speakers, and the Ethnologue status of Arabic. Since the Ethnologue status of Arabic was chosen as the representative feature in the dataset with feature selection based on Pearson coefficients, this feature was selected as the representative feature for the dataset with feature selection based on Spearman coefficients. These two methods were used to choose representative features for the rest of the clusters in the dataset with feature selection based on Pearson coefficients (**Appendix B**) and the dataset with feature selection based on Spearman coefficients (**Appendix C**).

3. Models

Six models were used to identify the best match for the data. For all models, data were split into a training set and a testing set. The train set of data is used to create the model and train it to be more accurate, while the test set validates the model. A third of the randomized data were used for testing the data, while the remaining two-thirds of the data were used for training.

3.1 Accuracy Indicators

The accuracy of the models is measured by Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2).

3.2. Linear Models (Multiple Linear Regression, Ridge Regression, LASSO Regression)

Three linear models were used in this study: multiple linear regression, ridge regression, and LASSO regression.[20] Multiple linear regression is a model that finds the linear relationship between an independent variable and dependent variables. The shrinkage penalties in the ridge and LASSO regressions are adjusted to tune the models to address multicollinearity. While ridge only has a shrinkage penalty, LASSO can also perform variable selection by zeroing out variable coefficients.

3.3. Tree-based Models (Decision Tree, Random Forest, Gradient Boosting)

Three tree-based models, decision tree regression, random forest, and gradient boosting, were tested in this study.[21] To tune these models, maximum tree depth, number of trees, and number of iterations were used on decision tree regression, random forest, and gradient boosting models respectively. Decision tree regression is a nonlinear model that splits data from the root into multiple nodes to classify the data. Random forest and gradient boosting models build upon decision tree regression by containing multiple decision trees. While random forest builds each decision tree independently, gradient boosting builds decision trees consecutively so each subsequent decision tree is formed to reduce errors

[20] James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). An introduction to statistical learning: With applications in python (1st ed.). Springer International Publishing.

[21] Breiman, L. (2001). Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

from the previous tree. As a result, random forest models handle feature interactions well, while gradient boosting models often provide higher accuracy but require fine-tuning.

3.4. SHAP Values

SHAP, an abbreviation of SHapley Additive exPlanations,[22] values measure the feature importance of models based on the concept of Shapley values in game theory, helping to explain why certain predictors (e.g., mobile subscriptions) are important. Shapley values are calculated based on the marginal contributions of each component. For each component, a series of iterations was conducted by incrementally adding features to the model besides the component itself. The Shapley value is calculated by averaging the differences among all iterations without the component and the full model result. When features are heavily correlated, SHAP values stray further from their true Shapley value.[23]

4. Empirical Findings

In this study, all models were created using 3 datasets: the full dataset, the dataset with clusters based on Pearson coefficient analysis, and the dataset with clusters based on Spearman coefficient analysis.

4.1. Correlation Analysis with Pearson Coefficient

Correlation among features can create inaccurate modeling results in linear models. By analyzing the dataset based on the Pearson coefficient and creating a dendrogram based on Ward's linkage, 13 clusters of features were identified. Eleven out of the 13 new clusters had features formed within feature categories. Two clusters were identified to have features from multiple categories. A new dataset was created based on this analysis, and the number of features used was reduced from 70 features to 45 features.

4.2. Correlation Analysis with Spearman Coefficient

Spearman coefficient analysis identified monotonic relationships among features. In the analysis, 13 clusters were identified. Two out of 13 clusters had features that spanned across multiple categories. One cluster was formed based on mean years of schooling, percent of internet users in the country, expected years of schooling years, infant mortality rate rates and gross national income, and life expectancy, while another cluster was formed based on births, infant deaths, population, internet users and mobile subscriptions, and number of people in the labor force. With

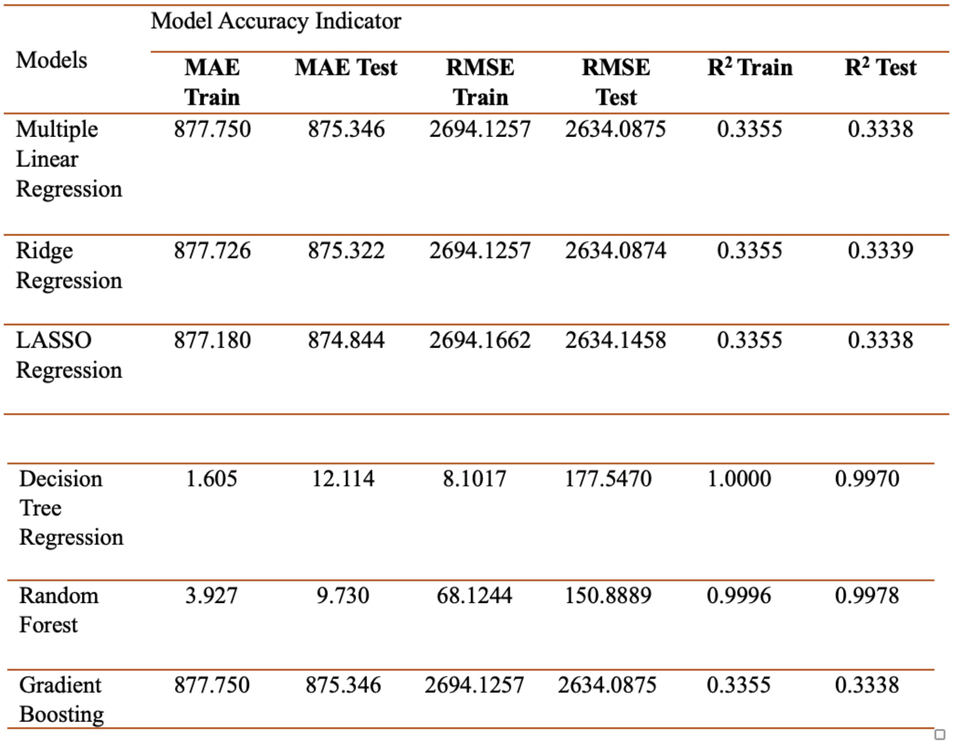
[22] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[23] Molnar, C. (2020). Interpretable machine learning. Lulu.com.

hierarchical clustering, the features used in this new dataset were reduced from 70 to 39 features.

4.3. Evaluation of Linear Models

In **Table 1**, which displays each model's training and testing accuracy indicators, linear models had a higher error than the tree-based models. All linear models had low accuracy, and none of the linear models reached convergence after 1000 iterations during the training. Even though there may be a linear relationship between online labor market activity and features on an individual basis, online labor market activity does not have a linear relationship with features as an aggregate in this study.



Models	Model Accuracy Indicator					
	MAE Train	**MAE Test**	**RMSE Train**	**RMSE Test**	**R^2 Train**	**R^2 Test**
Multiple Linear Regression	877.750	875.346	2694.1257	2634.0875	0.3355	0.3338
Ridge Regression	877.726	875.322	2694.1257	2634.0874	0.3355	0.3339
LASSO Regression	877.180	874.844	2694.1662	2634.1458	0.3355	0.3338
Decision Tree Regression	1.605	12.114	8.1017	177.5470	1.0000	0.9970
Random Forest	3.927	9.730	68.1244	150.8889	0.9996	0.9978
Gradient Boosting	877.750	875.346	2694.1257	2634.0875	0.3355	0.3338

TABLE 1. Comparison of MAE, RMSE, and R^2 for all models.

4.4. Evaluation of Tree-Based Models

All 3 tree-based models, decision tree regression, random forest, and gradient boosting, were tuned by testing for the highest R^2 with combinations of model-specific parameters. The decision tree regression model's accuracy improved with increasing R^2 value and decreasing RMSE value when the maximum number of tree depths increased (**Figure 3**). Its value reached its highest when the depth was 21. For random forest models, the R^2 improved with the number of trees added to the model (**Figure 4**). R^2 was the highest when there were 34 trees, and the maximum tree depth was 55.

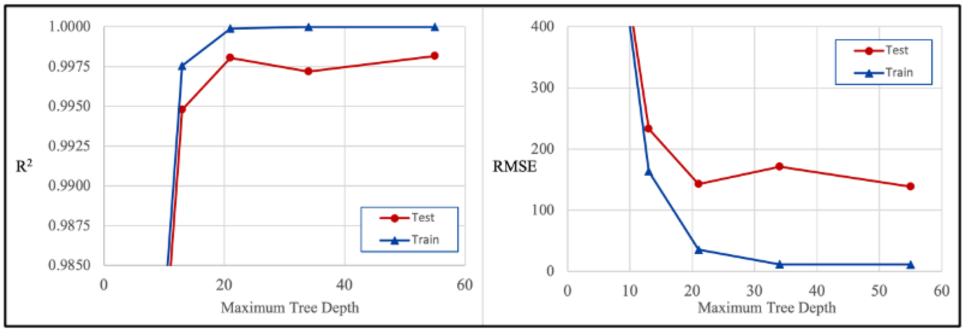


FIGURE 3. R^2 and RMSE as a function of maximum tree depth in decision tree regression model with feature selection using Spearman ranking coefficient.

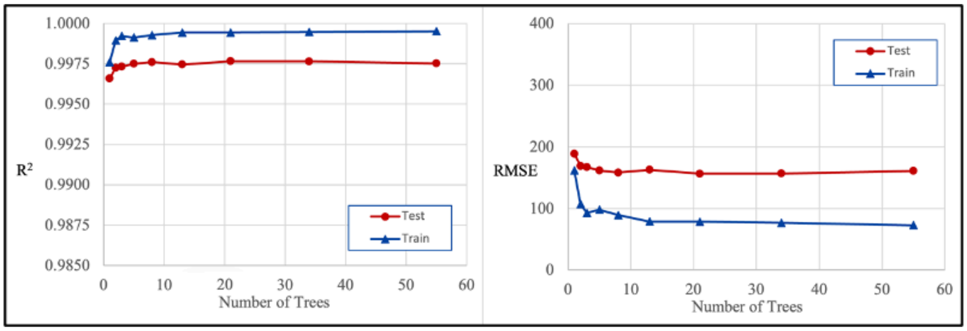


FIGURE 4. R^2 and RMSE as a function of the number of trees in the random forest model with feature selection using Spearman ranking coefficient.

In the case of gradient boosting modeling, improvements were found in R^2 with the number of iterative corrections being made to the model (**Figure 5**). Despite the iterative improvements, the RMSE of the gradient boosting models was still the highest among all tree-based models.

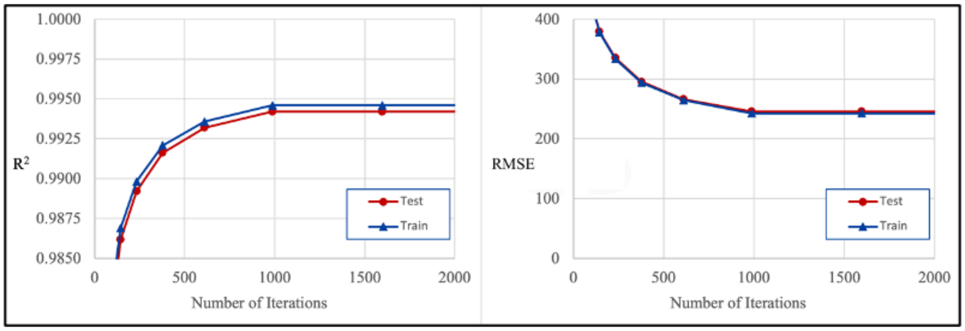


FIGURE 5. R^2 and RMSE as a function of number of iterations in gradient boosting model with feature selection using Spearman ranking coefficient.

Using these 3 parameters to establish the optimized models, tree-based models had higher accuracy and lower errors than the linear models (**Table 1**). The average testing MAE values for all tree-based models were 96.38% smaller than the testing MAE values for linear models. The average testing RMSE values for tree-based models was 93.21% smaller than the average testing RMSE values for linear models. For model fit, the average testing R^2 values for tree-based models were 242.15% larger than that of linear models. The random forest model had higher accuracy than the LASSO model, as shown in the residual plot comparison (**Figure 6**).

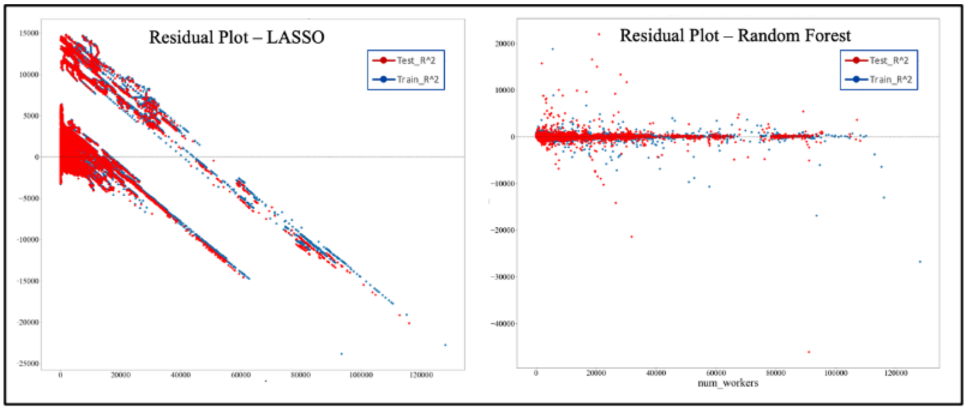


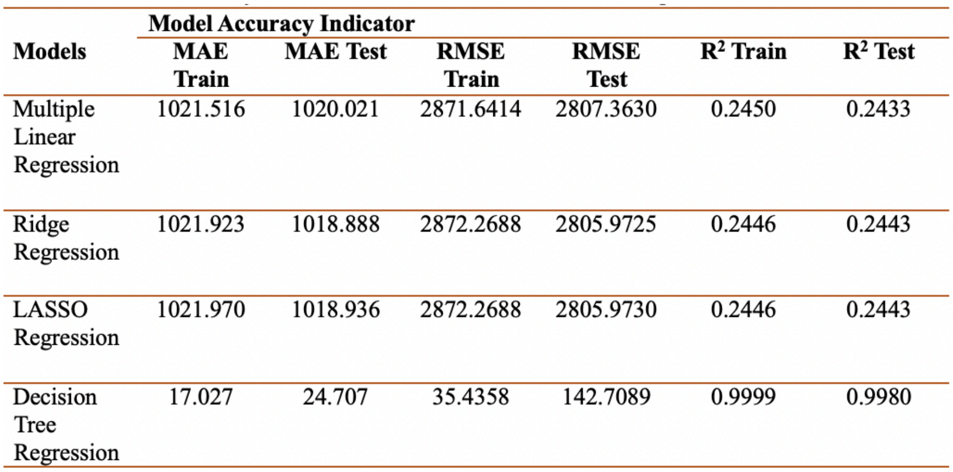
FIGURE 6. Residual comparison between LASSO and random forest models.

For tree-based models, highly correlated features co-existing in the dataset can affect the evaluation of the feature importance.[24] Feature selection based on Pearson and Spearman coefficients did not affect the R^2 of the tree-based models (**Table 1, Table 2, Table 3**). In comparison, for all 3 tree-based models, R^2 values of training data among all 3 datasets varied by less than 0.25%. Random forest models with the feature selections had the least variation in RMSE and MAE test values from the model with the original dataset. By applying feature selection, RMSE and MAE values of the random forest models varied by less than 5%.

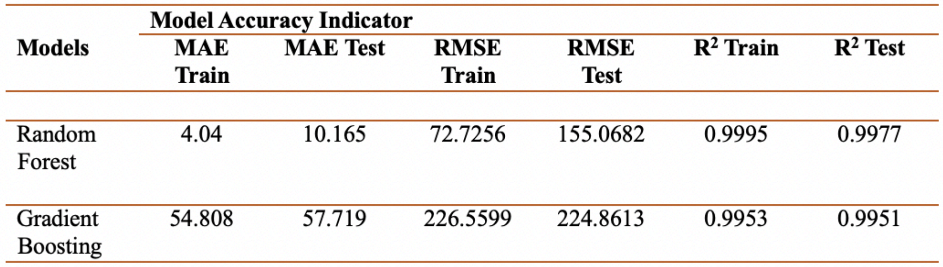
	Model Accuracy Indicator					
Models	**MAE Train**	**MAE Test**	**RMSE Train**	**RMSE Test**	**R^2 Train**	**R^2 Test**
Multiple Linear Regression	927.830	925.250	2772.7066	2706.4776	0.2962	0.2967
Ridge Regression	927.633	925.622	2771.4472	2709.0626	0.2967	0.2956
LASSO Regression	927.625	925.614	2771.4471	2709.0619	0.2967	0.2956
Decision Tree Regression	0.990	11.048	6.9186	156.0806	1.0000	0.9977
Random Forest	3.873	9.494	68.9942	148.4736	0.9996	0.9979
Gradient Boosting	98.849	101.128	285.3918	271.5192	0.9925	0.9929

[24] Molnar, C. (2020). Interpretable machine learning. Lulu.com.

Table 2. Model Accuracy: Dataset with feature selection based on Pearson coefficient ≥ 0.7



Models	Model Accuracy Indicator					
	MAE Train	MAE Test	RMSE Train	RMSE Test	R^2 Train	R^2 Test
Multiple Linear Regression	1021.516	1020.021	2871.6414	2807.3630	0.2450	0.2433
Ridge Regression	1021.923	1018.888	2872.2688	2805.9725	0.2446	0.2443
LASSO Regression	1021.970	1018.936	2872.2688	2805.9730	0.2446	0.2443
Decision Tree Regression	17.027	24.707	35.4358	142.7089	0.9999	0.9980



Models	Model Accuracy Indicator					
	MAE Train	MAE Test	RMSE Train	RMSE Test	R^2 Train	R^2 Test
Random Forest	4.04	10.165	72.7256	155.0682	0.9995	0.9977
Gradient Boosting	54.808	57.719	226.5599	224.8613	0.9953	0.9951

Table 3. Model Accuracy: Dataset with feature selection based on Spearman coefficient ≥ 0.7

Comparing the accuracy indicators between testing and training data can identify the level of underfitting or overfitting among all the models. Among the 3 tree-based models, decision tree regression models have overfitting with large differences in RMSE and MAE values between test and train data and low RMSE and MAE training scores.

Gradient boosting models have the least overfitting among all tree-based models. It was the only tree-based model where the RMSE of the training data was higher than that of the testing data. While gradient boosting has the least overfitting among all tree-based models, it has the highest MAE and RMSE values and the lowest R^2 values.

Overall, among the 3 tree-based models, the random forest model was the most optimal model, which had lower error than the gradient boosting model and less overfitting than the decision tree regression model. The random forest model also had the least variation in R^2, MAE, and RMSE after the model was simplified with correlation analysis.

4.5. SHAP Values

Ensemble models, like random forest and gradient boosting, have had challenges in their explainability despite their high accuracy.[25] SHAP was used to address the challenges of explainability. Of these ensemble models, SHAP values of features in the random forest model with feature selection based on Spearman correlation coefficients were computed (**Figure 7**). In order of importance, the top five features identified in the random forest model were the number of mobile subscriptions, the number of English speakers, the software development occupation category, the creative and multimedia occupation category, and the net migration rate. Mobile subscriptions emerged as the most influential feature, which may be due to higher mobile subscriptions facilitating access to online labor platforms, especially in low-income or rural regions.

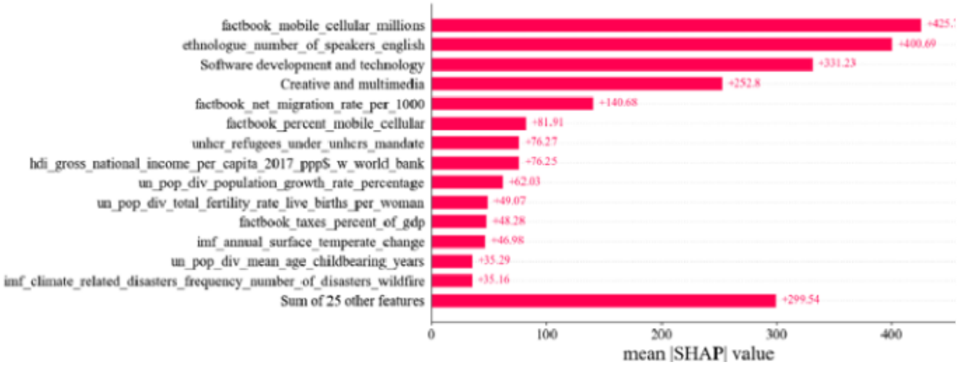


Figure 7. SHAP Values of random forest model with feature selection using Spearman coefficient.

5. Discussion

In the tree-based models with feature selection using Spearman correlation, none of the features was identified as the single most important contributing factor with a majority of all SHAP values. Although the number of English speakers, cellular mobile subscriptions, and software development were the top 3 features related to the online labor market supply, these factors represent 49% of the marginal contribution to the random forest model. While the models were adjusted with feature selection, the models may have overfit certain features or failed to capture dynamic changes in the labor supply.

5.1. Mobile Cellular Subscriptions

When considering Pearson and Spearman correlation coefficients, clusters were formed based on the number of mobile subscriptions, internet users, labor force, and the total population, representing the strong adoption of the technology in everyday life where the mobile subscriptions rise in

[25] Peet, E., Vegetabile, B., Cefalu, M., Pane, J., and Damberg, C. (2022). Machine learning in public policy: the perils and the promise of interpretability. RAND Corporation. https://apo.org.au/node/320715

conjunction with the total population and internet users. Mobile subscriptions represent technological infrastructure, a foundational driver of human capital formation in the digital economy.[26] The increased adoption of new technologies has a significant role in the growth of the online labor market. The supply of online labor activities in the top 10 countries, which represents 49.09% of all activities, has grown 26.56% during the period of this study from June 2017 to December 2022. These top 10 countries had 35.64% of all mobile subscriptions worldwide in 2022. The result of this study reinforces the importance of information and communications technology (ICT) in creating access to economic opportunities for workers.[27]

5.2. English Speakers

The online labor market spans across borders with participants from a large range of countries and territories. A common language is required for trade in this economy. Although English is not necessarily the required language for international trade, the ability for parties to communicate in a common language is more critical than trust and ethnicity.[28] English's importance in all 3 tree-based models is a result of the need for a common language used in the demand and supply sides of this online labor market. From the demand side, the top 7 buying countries, representing 65.87% of all demand, have English as a national language. At the same time, from the supply perspective, the top 10 supplying countries, representing 49.09% of the 2022 online labor supply, had 61.03% of all first-language (L1) and second-language (L2) English speakers, as defined by Ethnologue.[29, 30] Although the number of English speakers is one of the top features based on SHAP values, it only represents 17% of the total contribution in the random forest model.

5.3. Software Development Category and Creative and Multimedia Category

The importance of the creative and multimedia occupation category is a result of an existing adoption of freelancing in the profession. Artists, who work in this occupation category, have been practicing freelancing for

[26] Becker, G. S. (1993). Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education (3rd ed.). Chicago: University of Chicago Press. http://doi.org/10.7208/chicago/9780226041223.001.0001

[27] Atasoy, H. (2013). The effects of broadband Internet expansion on labor market outcomes. Industrial & Labor Relations Review, 66(2), 315–345. https://doi.org/10.1177/001979391306600202

[28] Melitz, J., and Toubal, F. (2014). Native language, spoken language, translation and trade. Journal of International Economics, 93(2), 351–363. https://doi.org/10.1016/j.jinteco.2014.04.004

[29] Stephany, F., Kässi, O., Rani, U., and Lehdonvirta, V. (2021). Online Labour Index 2020: New ways to measure the world's remote freelancing market. Big Data & Society, 8(2), 205395172110432. https://doi.org/10.1177/20539517211043240

[30] Eberhard, D., Simons, G., and Fennig, C. (2019). Ethnologue: Languages of the World. Ethnologue: Languages of the World.

more than 20 years. From 2003 to 2015, in the United States, more than 30% of all artists were self-employed.[31] Eighteen percent of all graphic designers, part of the artist group in this creative and multimedia occupation category, are self-employed workers as of 2023 in the United States.[32] The online labor platforms and adoption of cloud applications supported freelancing in these creative and multimedia projects.[33] Although the COVID-19 pandemic negatively affected the creative industry,[34] this category remained the 2nd highest in demand in the online labor market.[35]

The importance of software development features reflects the demand for specific technical skills in the global market. Since 2016, software development has been highlighted as an occupation with high demand across multiple countries.[36, 37, 38, 39] In 2022, software development was one of the occupations with the most common labor shortage in Europe.[40] While the online gig economy provides online opportunities to workers worldwide, the skills required for these opportunities need to be widely taught. National programs, such as the National Freelance Training Program by the Ministry of Information Technology and Telecommunications in Pakistan, have been teaching digital skills, such as software development and creative media, to meet the demand provided by freelance online gig work.[41]

5.4. Migration Rate

[31] Woronkowicz, J., and Noonan, D. S. (2019). Who Goes Freelance? The Determinants of Self-Employment for Artists. Entrepreneurship Theory and Practice, 43(4), 651–672. https://doi.org/10.1177/1042258717728067

[32] Bureau of Labor Statistics, U.S. Department of Labor. (n.d.). National employment matrix 27-1024 Graphic designers. Bls.gov. Retrieved September 8, 2023, from https://data.bls.gov/projections/nationalMatrix?queryParams=27-1024&ioType=o

[33] Sutherland, W., and Jarrahi, M. H. (2017). The gig economy and information infrastructure: The case of the digital nomad community. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1–24. https://doi.org/10.1145/3134732

[34] Khlystova, O., Kalyuzhnova, Y., and Belitski, M. (2022). The impact of the COVID-19 pandemic on the creative industries: A literature review and future research agenda. Journal of Business Research, 139, 1192–1210. https://doi.org/10.1016/j.jbusres.2021.09.062

[35] Stephany, F., Kässi, O., Rani, U., and Lehdonvirta, V. (2021). Online Labour Index 2020: New ways to measure the world's remote freelancing market. Big Data & Society, 8(2), 205395172110432. https://doi.org/10.1177/20539517211043240

[36] The Future of Jobs 2016. (2016, January 1). World Economic Forum. https://www.weforum.org/reports/the-future-of-jobs-2016

[37] The Future of Jobs Report 2018. (2018, September 17). World Economic Forum. https://www.weforum.org/reports/the-future-of-jobs-report-2018

[38] The Future of Jobs Report 2020. (2020, October 20). World Economic Forum. https://www.weforum.org/reports/the-future-of-jobs-report-2020

[39] The Future of Jobs Report 2023. (2023, April 30). World Economic Forum. https://www.weforum.org/reports/the-future-of-jobs-report-2023

[40] Ibid.

[41] National Freelance Training Program. (n.d.). Nftp.pitb.gov.Pk. Retrieved August 30, 2023, from https://nftp.pitb.gov.pk/

In all 3 tree-based models, migration and the number of refugees had high SHAP values and were important for creating the models on the online labor market. Many migrants suffer in adapting to new countries where barriers limit job opportunities and render them uncertain.[42] Migrants are more likely to participate in online labor markets due to barriers to traditional employment in host countries, highlighting the need for policies that integrate migrant workers into digital economies. The online labor market often relies on migrants and their lack of alternative job options with low barriers to entry.[43] In addition, wages offered on platforms like Upwork may disproportionately attract workers from low-income countries, where the cost of living is lower, creating regional inequalities in labor participation. While the online labor market provides jobs, the income is often insufficient for a standalone job.[44] Although the number of refugees was one of the top 9 most contributing features, the lack of affordable mobile cellular access for refugees is a barrier to connecting with online economic activity.[45]

5.5. Implications

These findings from the models about key features in the online labor market can inform policy or labor economics. Limited access to mobile technology in low-income regions exacerbates labor market inequalities, restricting participation in online platforms. Governments can invest in affordable mobile broadband and digital literacy programs to increase online labor market participation. Increased accessibility to education in software development or related fields would allow more workers to fulfill the great demand for software developer roles in the online labor market. In addition, more resources dedicated to learning English and first languages in respective countries would address the language barriers for migrant workers in digital platforms. While promoting inclusive labor policies can reduce unemployment in vulnerable populations by helping migrants access employment opportunities through online platforms, there should be further research on minimum wage standards or fair labor practices in online platforms can support additional policies.

6. Conclusion

[42] Altenried, M. (2021). Mobile workers, contingent labour: Migration, the gig economy and the multiplication of labour. Environment & Planning A, 0308518X2110548. https://doi.org/10.1177/0308518x211054846

[43] Hackl, A., and International Labor Organization. (2021). Digital refugee livelihoods and decent work: towards inclusion in a fairer digital economy. https://policycommons.net/artifacts/1528335/digital-refugee-livelihoods-and-decentwork/2218020/

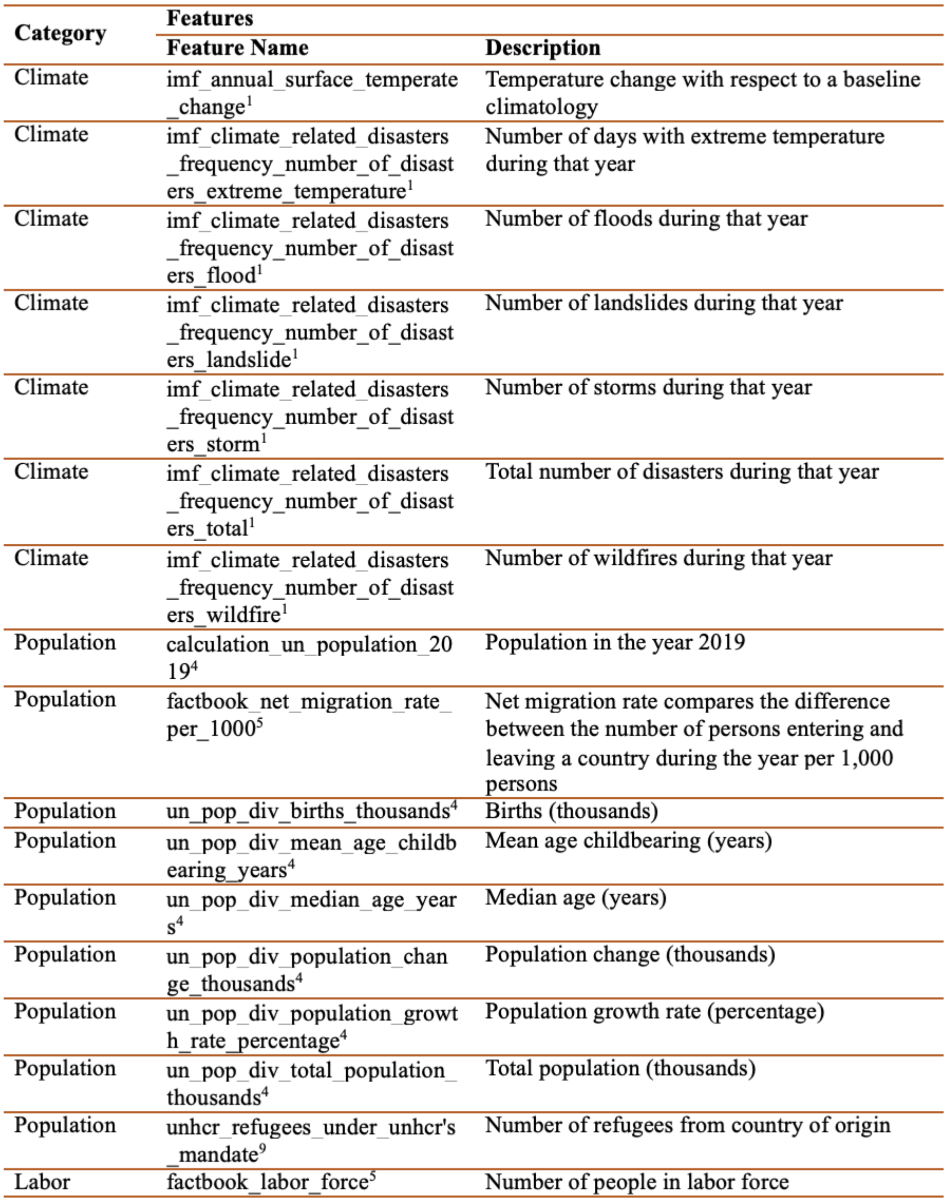
[44] Vernon, A., Deriche, K., and Eisenhauer, S. (2016). Connecting refugees-how Internet and mobile connectivity can improve refugee well-being and transform humanitarian action. UNHCR.

[45] Ibid.

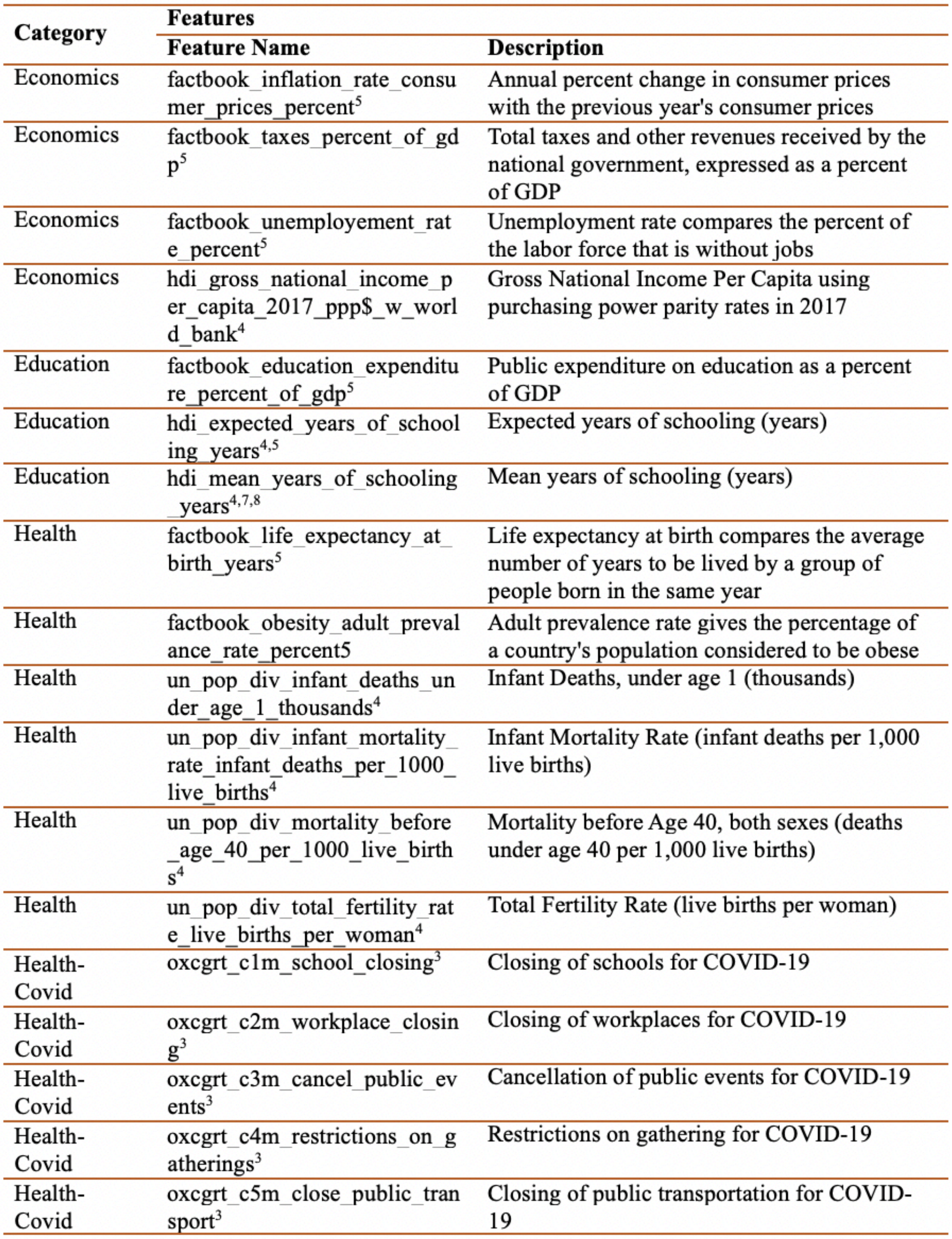
The random forest machine learning model is the optimal model to assess the factors that correlate with the labor supply in the online labor market. From the study, technology adoption, linguistics, and social factors are uncovered to be important to the overall supply at a global level. Our findings highlight the importance of technological infrastructure and migration policies in shaping online labor markets. By identifying key drivers of online labor participation, this research provides a foundation for policies that promote inclusive economic growth in the digital economy.

Appendix A

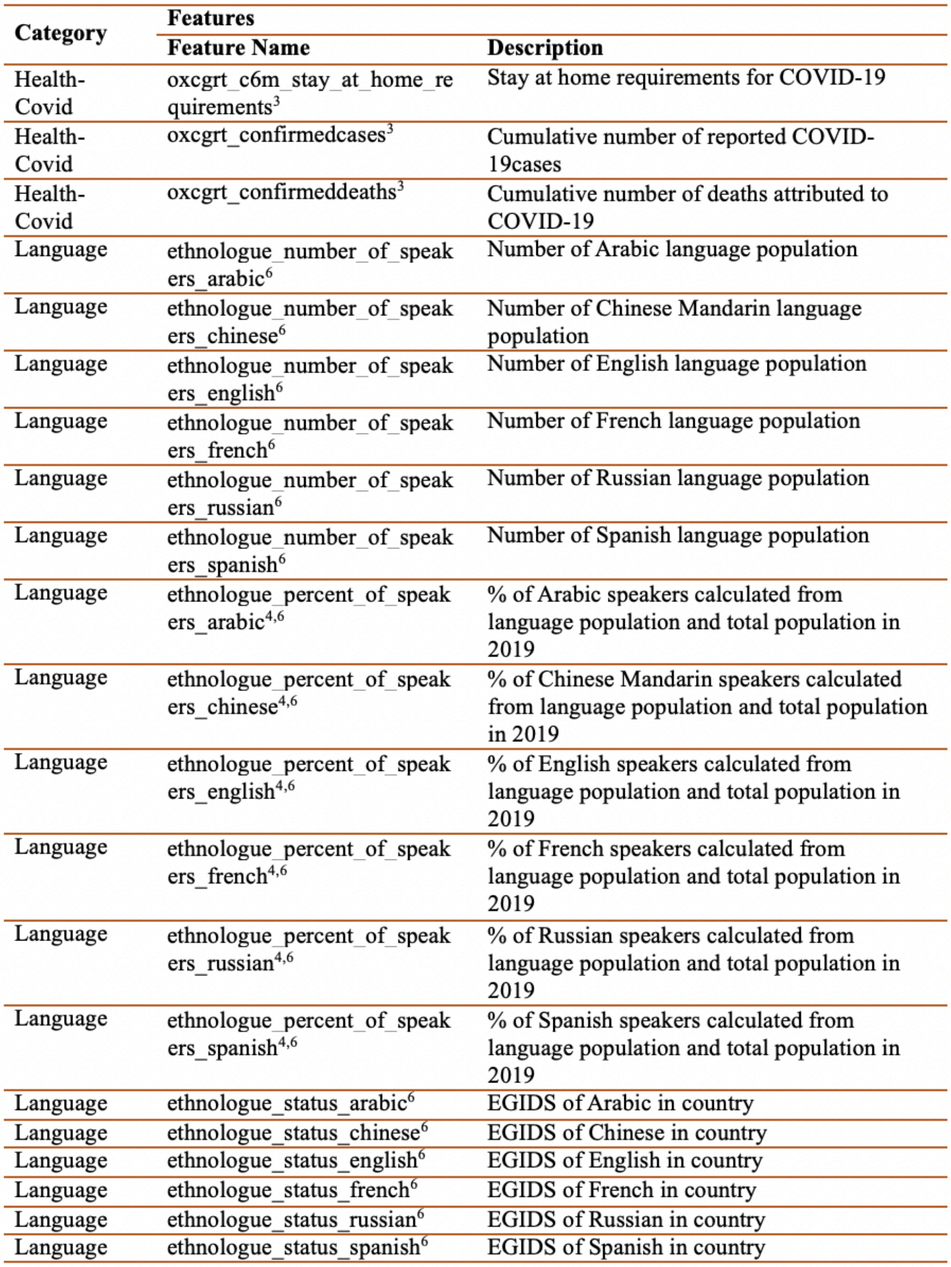
List of all feature names used to create the models and their source.



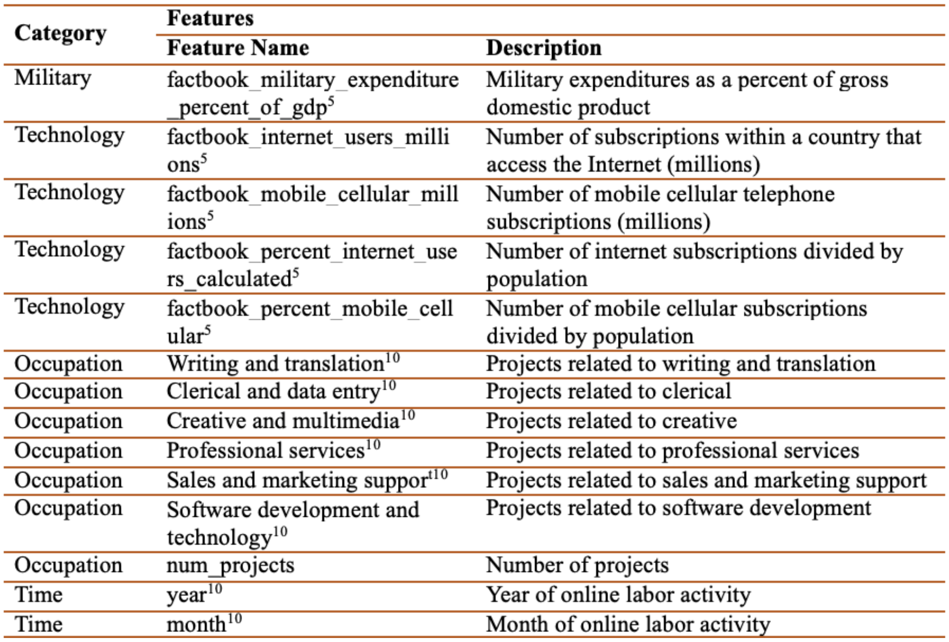
Category	Features	
	Feature Name	**Description**
Climate	imf_annual_surface_temperate_change[1]	Temperature change with respect to a baseline climatology
Climate	imf_climate_related_disasters_frequency_number_of_disasters_extreme_temperature[1]	Number of days with extreme temperature during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_flood[1]	Number of floods during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_landslide[1]	Number of landslides during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_storm[1]	Number of storms during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_total[1]	Total number of disasters during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_wildfire[1]	Number of wildfires during that year
Population	calculation_un_population_2019[4]	Population in the year 2019
Population	factbook_net_migration_rate_per_1000[5]	Net migration rate compares the difference between the number of persons entering and leaving a country during the year per 1,000 persons
Population	un_pop_div_births_thousands[4]	Births (thousands)
Population	un_pop_div_mean_age_childbearing_years[4]	Mean age childbearing (years)
Population	un_pop_div_median_age_years[4]	Median age (years)
Population	un_pop_div_population_change_thousands[4]	Population change (thousands)
Population	un_pop_div_population_growth_rate_percentage[4]	Population growth rate (percentage)
Population	un_pop_div_total_population_thousands[4]	Total population (thousands)
Population	unhcr_refugees_under_unhcr's_mandate[9]	Number of refugees from country of origin
Labor	factbook_labor_force[5]	Number of people in labor force



Category	Features	
	Feature Name	**Description**
Economics	factbook_inflation_rate_consumer_prices_percent[5]	Annual percent change in consumer prices with the previous year's consumer prices
Economics	factbook_taxes_percent_of_gdp[5]	Total taxes and other revenues received by the national government, expressed as a percent of GDP
Economics	factbook_unemployement_rate_percent[5]	Unemployment rate compares the percent of the labor force that is without jobs
Economics	hdi_gross_national_income_per_capita_2017_ppp$_w_world_bank[4]	Gross National Income Per Capita using purchasing power parity rates in 2017
Education	factbook_education_expenditure_percent_of_gdp[5]	Public expenditure on education as a percent of GDP
Education	hdi_expected_years_of_schooling_years[4,5]	Expected years of schooling (years)
Education	hdi_mean_years_of_schooling_years[4,7,8]	Mean years of schooling (years)
Health	factbook_life_expectancy_at_birth_years[5]	Life expectancy at birth compares the average number of years to be lived by a group of people born in the same year
Health	factbook_obesity_adult_prevalance_rate_percent5	Adult prevalence rate gives the percentage of a country's population considered to be obese
Health	un_pop_div_infant_deaths_under_age_1_thousands[4]	Infant Deaths, under age 1 (thousands)
Health	un_pop_div_infant_mortality_rate_infant_deaths_per_1000_live_births[4]	Infant Mortality Rate (infant deaths per 1,000 live births)
Health	un_pop_div_mortality_before_age_40_per_1000_live_births[4]	Mortality before Age 40, both sexes (deaths under age 40 per 1,000 live births)
Health	un_pop_div_total_fertility_rate_live_births_per_woman[4]	Total Fertility Rate (live births per woman)
Health-Covid	oxcgrt_c1m_school_closing[3]	Closing of schools for COVID-19
Health-Covid	oxcgrt_c2m_workplace_closing[3]	Closing of workplaces for COVID-19
Health-Covid	oxcgrt_c3m_cancel_public_events[3]	Cancellation of public events for COVID-19
Health-Covid	oxcgrt_c4m_restrictions_on_gatherings[3]	Restrictions on gathering for COVID-19
Health-Covid	oxcgrt_c5m_close_public_transport[3]	Closing of public transportation for COVID-19



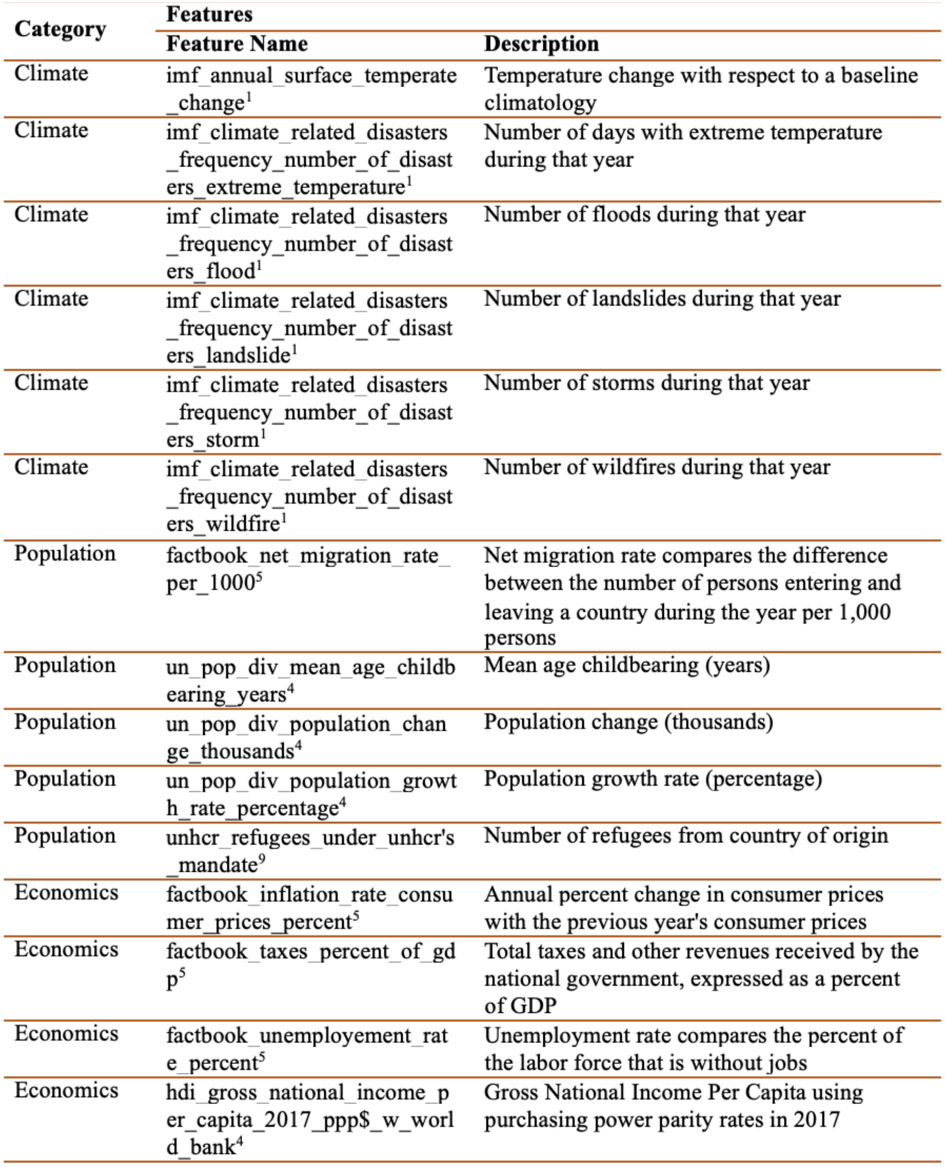
Category	Features	
	Feature Name	Description
Health-Covid	oxcgrt_c6m_stay_at_home_requirements[3]	Stay at home requirements for COVID-19
Health-Covid	oxcgrt_confirmedcases[3]	Cumulative number of reported COVID-19cases
Health-Covid	oxcgrt_confirmeddeaths[3]	Cumulative number of deaths attributed to COVID-19
Language	ethnologue_number_of_speakers_arabic[6]	Number of Arabic language population
Language	ethnologue_number_of_speakers_chinese[6]	Number of Chinese Mandarin language population
Language	ethnologue_number_of_speakers_english[6]	Number of English language population
Language	ethnologue_number_of_speakers_french[6]	Number of French language population
Language	ethnologue_number_of_speakers_russian[6]	Number of Russian language population
Language	ethnologue_number_of_speakers_spanish[6]	Number of Spanish language population
Language	ethnologue_percent_of_speakers_arabic[4,6]	% of Arabic speakers calculated from language population and total population in 2019
Language	ethnologue_percent_of_speakers_chinese[4,6]	% of Chinese Mandarin speakers calculated from language population and total population in 2019
Language	ethnologue_percent_of_speakers_english[4,6]	% of English speakers calculated from language population and total population in 2019
Language	ethnologue_percent_of_speakers_french[4,6]	% of French speakers calculated from language population and total population in 2019
Language	ethnologue_percent_of_speakers_russian[4,6]	% of Russian speakers calculated from language population and total population in 2019
Language	ethnologue_percent_of_speakers_spanish[4,6]	% of Spanish speakers calculated from language population and total population in 2019
Language	ethnologue_status_arabic[6]	EGIDS of Arabic in country
Language	ethnologue_status_chinese[6]	EGIDS of Chinese in country
Language	ethnologue_status_english[6]	EGIDS of English in country
Language	ethnologue_status_french[6]	EGIDS of French in country
Language	ethnologue_status_russian[6]	EGIDS of Russian in country
Language	ethnologue_status_spanish[6]	EGIDS of Spanish in country



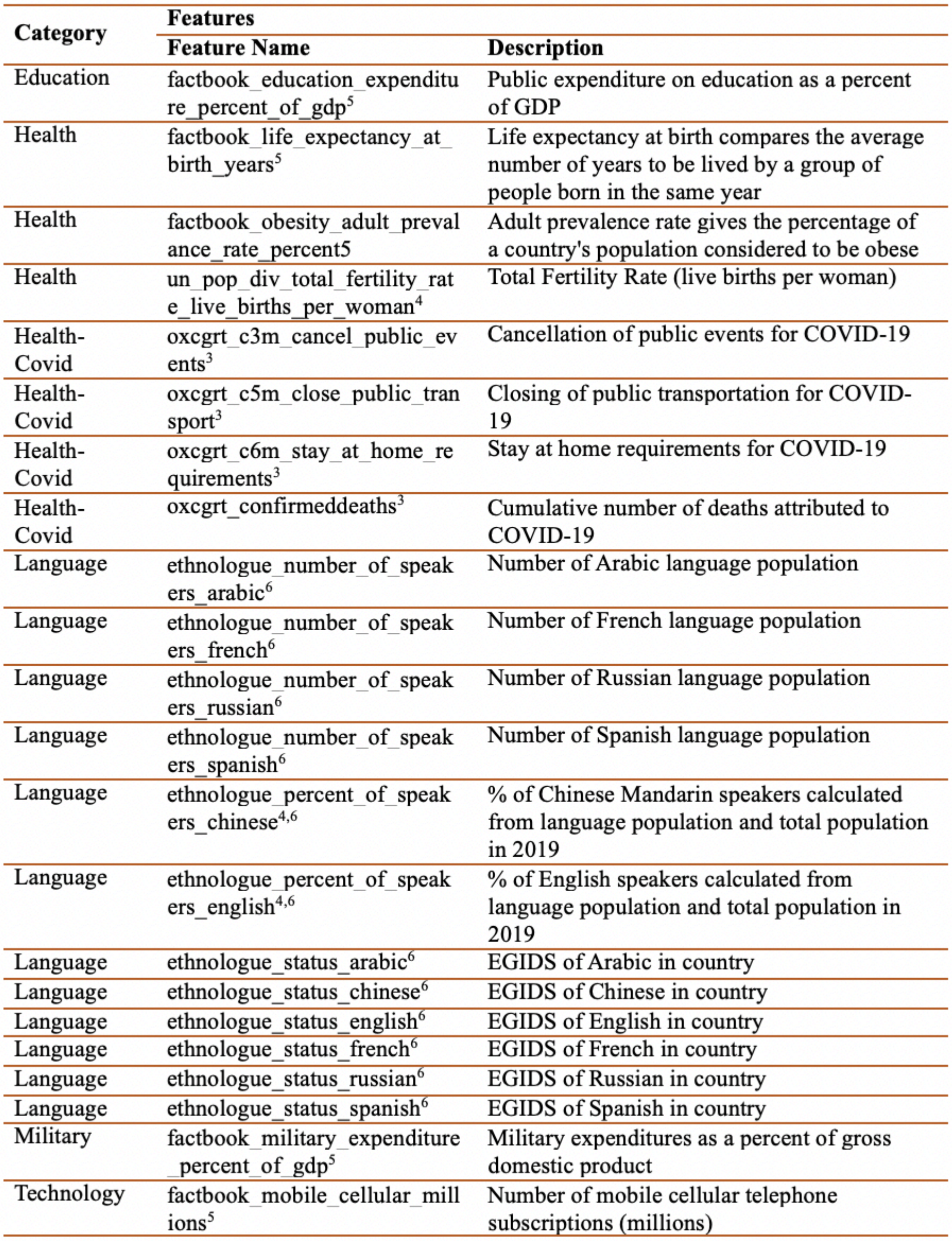
Category	Features	
	Feature Name	**Description**
Military	factbook_military_expenditure_percent_of_gdp[5]	Military expenditures as a percent of gross domestic product
Technology	factbook_internet_users_millions[5]	Number of subscriptions within a country that access the Internet (millions)
Technology	factbook_mobile_cellular_millions[5]	Number of mobile cellular telephone subscriptions (millions)
Technology	factbook_percent_internet_users_calculated[5]	Number of internet subscriptions divided by population
Technology	factbook_percent_mobile_cellular[5]	Number of mobile cellular subscriptions divided by population
Occupation	Writing and translation[10]	Projects related to writing and translation
Occupation	Clerical and data entry[10]	Projects related to clerical
Occupation	Creative and multimedia[10]	Projects related to creative
Occupation	Professional services[10]	Projects related to professional services
Occupation	Sales and marketing suppor[t10]	Projects related to sales and marketing support
Occupation	Software development and technology[10]	Projects related to software development
Occupation	num_projects	Number of projects
Time	year[10]	Year of online labor activity
Time	month[10]	Month of online labor activity

Appendix B

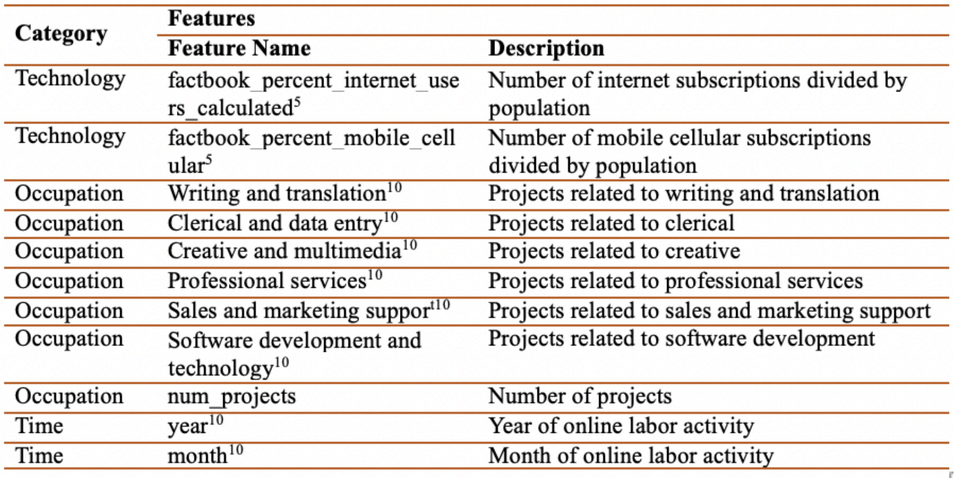
List of all feature names used to create the models and their source with feature selection based on Pearson correlation coefficient.



Category	Features	
	Feature Name	Description
Climate	imf_annual_surface_temperate_change[1]	Temperature change with respect to a baseline climatology
Climate	imf_climate_related_disasters_frequency_number_of_disasters_extreme_temperature[1]	Number of days with extreme temperature during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_flood[1]	Number of floods during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_landslide[1]	Number of landslides during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_storm[1]	Number of storms during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_wildfire[1]	Number of wildfires during that year
Population	factbook_net_migration_rate_per_1000[5]	Net migration rate compares the difference between the number of persons entering and leaving a country during the year per 1,000 persons
Population	un_pop_div_mean_age_childbearing_years[4]	Mean age childbearing (years)
Population	un_pop_div_population_change_thousands[4]	Population change (thousands)
Population	un_pop_div_population_growth_rate_percentage[4]	Population growth rate (percentage)
Population	unhcr_refugees_under_unhcr's_mandate[9]	Number of refugees from country of origin
Economics	factbook_inflation_rate_consumer_prices_percent[5]	Annual percent change in consumer prices with the previous year's consumer prices
Economics	factbook_taxes_percent_of_gdp[5]	Total taxes and other revenues received by the national government, expressed as a percent of GDP
Economics	factbook_unemployement_rate_percent[5]	Unemployment rate compares the percent of the labor force that is without jobs
Economics	hdi_gross_national_income_per_capita_2017_ppp$_w_world_bank[4]	Gross National Income Per Capita using purchasing power parity rates in 2017



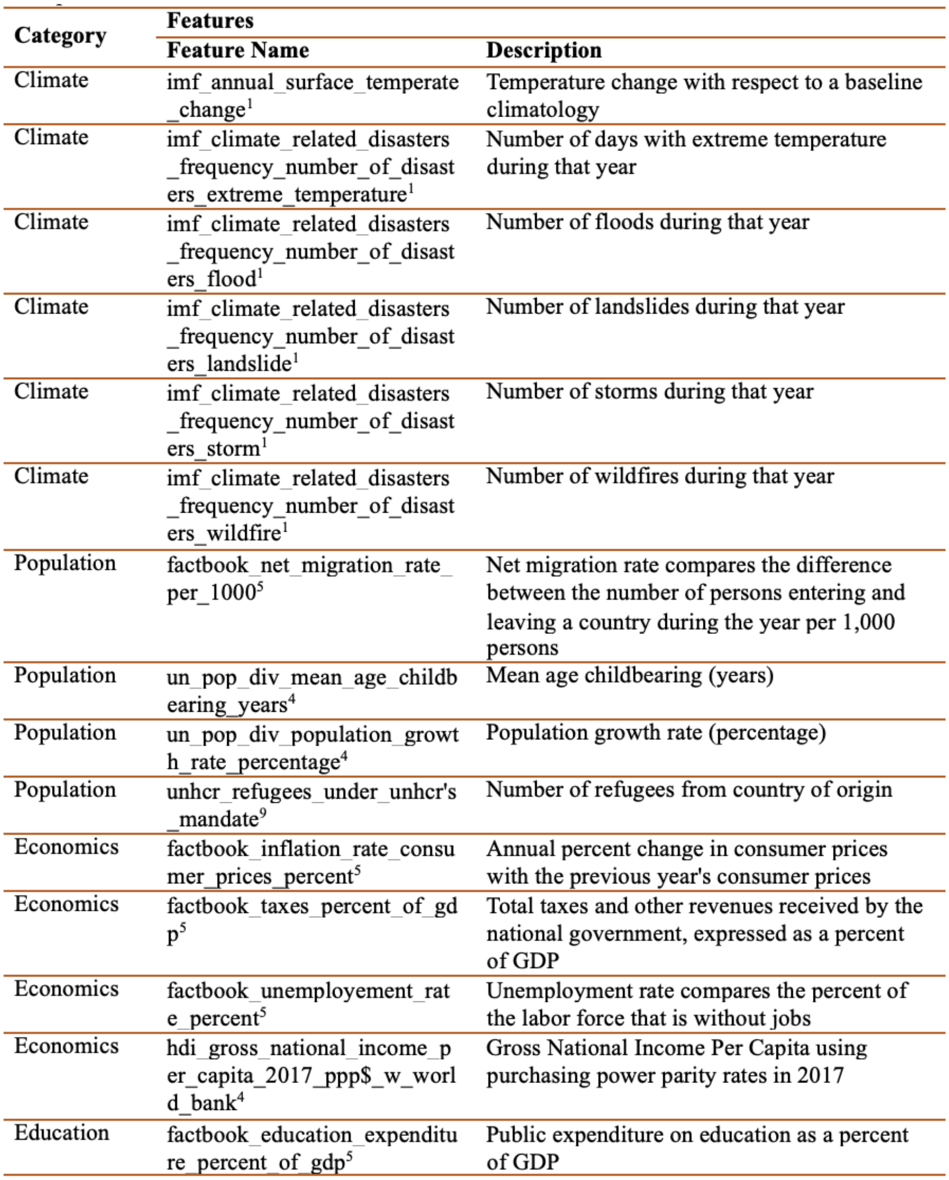
Category	Features	
	Feature Name	Description
Education	factbook_education_expenditure_percent_of_gdp[5]	Public expenditure on education as a percent of GDP
Health	factbook_life_expectancy_at_birth_years[5]	Life expectancy at birth compares the average number of years to be lived by a group of people born in the same year
Health	factbook_obesity_adult_prevalance_rate_percent5	Adult prevalence rate gives the percentage of a country's population considered to be obese
Health	un_pop_div_total_fertility_rate_live_births_per_woman[4]	Total Fertility Rate (live births per woman)
Health-Covid	oxcgrt_c3m_cancel_public_events[3]	Cancellation of public events for COVID-19
Health-Covid	oxcgrt_c5m_close_public_transport[3]	Closing of public transportation for COVID-19
Health-Covid	oxcgrt_c6m_stay_at_home_requirements[3]	Stay at home requirements for COVID-19
Health-Covid	oxcgrt_confirmeddeaths[3]	Cumulative number of deaths attributed to COVID-19
Language	ethnologue_number_of_speakers_arabic[6]	Number of Arabic language population
Language	ethnologue_number_of_speakers_french[6]	Number of French language population
Language	ethnologue_number_of_speakers_russian[6]	Number of Russian language population
Language	ethnologue_number_of_speakers_spanish[6]	Number of Spanish language population
Language	ethnologue_percent_of_speakers_chinese[4,6]	% of Chinese Mandarin speakers calculated from language population and total population in 2019
Language	ethnologue_percent_of_speakers_english[4,6]	% of English speakers calculated from language population and total population in 2019
Language	ethnologue_status_arabic[6]	EGIDS of Arabic in country
Language	ethnologue_status_chinese[6]	EGIDS of Chinese in country
Language	ethnologue_status_english[6]	EGIDS of English in country
Language	ethnologue_status_french[6]	EGIDS of French in country
Language	ethnologue_status_russian[6]	EGIDS of Russian in country
Language	ethnologue_status_spanish[6]	EGIDS of Spanish in country
Military	factbook_military_expenditure_percent_of_gdp[5]	Military expenditures as a percent of gross domestic product
Technology	factbook_mobile_cellular_millions[5]	Number of mobile cellular telephone subscriptions (millions)



Category	Features	
	Feature Name	Description
Technology	factbook_percent_internet_users_calculated[5]	Number of internet subscriptions divided by population
Technology	factbook_percent_mobile_cellular[5]	Number of mobile cellular subscriptions divided by population
Occupation	Writing and translation[10]	Projects related to writing and translation
Occupation	Clerical and data entry[10]	Projects related to clerical
Occupation	Creative and multimedia[10]	Projects related to creative
Occupation	Professional services[10]	Projects related to professional services
Occupation	Sales and marketing suppor[t10]	Projects related to sales and marketing support
Occupation	Software development and technology[10]	Projects related to software development
Occupation	num_projects	Number of projects
Time	year[10]	Year of online labor activity
Time	month[10]	Month of online labor activity

Appendix C

List of all feature names used to create the models and their source with feature selection based on Spearman correlation coefficient.



Category	Features	
	Feature Name	Description
Climate	imf_annual_surface_temperate_change[1]	Temperature change with respect to a baseline climatology
Climate	imf_climate_related_disasters_frequency_number_of_disasters_extreme_temperature[1]	Number of days with extreme temperature during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_flood[1]	Number of floods during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_landslide[1]	Number of landslides during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_storm[1]	Number of storms during that year
Climate	imf_climate_related_disasters_frequency_number_of_disasters_wildfire[1]	Number of wildfires during that year
Population	factbook_net_migration_rate_per_1000[5]	Net migration rate compares the difference between the number of persons entering and leaving a country during the year per 1,000 persons
Population	un_pop_div_mean_age_childbearing_years[4]	Mean age childbearing (years)
Population	un_pop_div_population_growth_rate_percentage[4]	Population growth rate (percentage)
Population	unhcr_refugees_under_unhcr's_mandate[9]	Number of refugees from country of origin
Economics	factbook_inflation_rate_consumer_prices_percent[5]	Annual percent change in consumer prices with the previous year's consumer prices
Economics	factbook_taxes_percent_of_gdp[5]	Total taxes and other revenues received by the national government, expressed as a percent of GDP
Economics	factbook_unemployement_rate_percent[5]	Unemployment rate compares the percent of the labor force that is without jobs
Economics	hdi_gross_national_income_per_capita_2017_ppp$_w_world_bank[4]	Gross National Income Per Capita using purchasing power parity rates in 2017
Education	factbook_education_expenditure_percent_of_gdp[5]	Public expenditure on education as a percent of GDP

Category	Features	
	Feature Name	Description
Health	factbook_obesity_adult_prevalance_rate_percent5	Adult prevalence rate gives the percentage of a country's population considered to be obese
Health	un_pop_div_total_fertility_rate_live_births_per_woman[4]	Total Fertility Rate (live births per woman)
Health-Covid	oxcgrt_c3m_cancel_public_events[3]	Cancellation of public events for COVID-19
Health-Covid	oxcgrt_c5m_close_public_transport[3]	Closing of public transportation for COVID-19
Health-Covid	oxcgrt_c6m_stay_at_home_requirements[3]	Stay at home requirements for COVID-19
Health-Covid	oxcgrt_confirmeddeaths[3]	Cumulative number of deaths attributed to COVID-19
Language	ethnologue_number_of_speakers_english[6]	Number of English language population
Language	ethnologue_status_arabic[6]	EGIDS of Arabic in country
Language	ethnologue_status_chinese[6]	EGIDS of Chinese in country
Language	ethnologue_status_english[6]	EGIDS of English in country
Language	ethnologue_status_french[6]	EGIDS of French in country
Language	ethnologue_status_russian[6]	EGIDS of Russian in country
Language	ethnologue_status_spanish[6]	EGIDS of Spanish in country
Military	factbook_military_expenditure_percent_of_gdp[5]	Military expenditures as a percent of gross domestic product
Technology	factbook_mobile_cellular_millions[5]	Number of mobile cellular telephone subscriptions (millions)
Technology	factbook_percent_mobile_cellular[5]	Number of mobile cellular subscriptions divided by population
Occupation	Writing and translation[10]	Projects related to writing and translation
Occupation	Clerical and data entry[10]	Projects related to clerical
Occupation	Creative and multimedia[10]	Projects related to creative
Occupation	Professional services[10]	Projects related to professional services
Occupation	Sales and marketing suppor[t10]	Projects related to sales and marketing support
Occupation	Software development and technology[10]	Projects related to software development
Occupation	num_projects	Number of projects
Time	month[10]	Month of online labor activity

Notes

[1] See International Monetary Fund (n.d.).

[2] See UNDP (2022).

[3] See Hale et al. (2021).

[4] See World Population Prospects 2022 (2023).

[5] See The World Factbook (2023).

[6] See Eberhard et al. (2019).

[7] See Jeroen, Smits, and Permanyer, 2019; J. Smits (2016).

[8] See Education statistics - all indicators (n.d.).

[9] See UNHCR (n.d.).

[10] See Stephany et al (2021).