AI-Powered Lifecycle Financial Planning – Predictive Modeling in Consumption

August 2023

Abstract

Consumption plays an essential role in both human life and the economics of a country. It is important to understand the relationship between people's background information and their annual consumption. For this project, our goal is to build machine learning based regression models that are useful for predicting the annual expenditure of people who reside in the United States. Given the dataset from the Consumer Expenditure Surveys, we tried to figure out the relationships between tables, build models for predicting individual consumption, and also find out important features for predicting consumption. Starting from simple linear regression, we work towards more advanced ensemble trees such as random forests and boosted trees. The project generates relatively good results but still has some limitations, and further steps are needed to attain the final goal.

1 Introduction

1.1 Background

Consumption is an indispensable part of people's life. Also, consumption is one of the most significant concepts in economics and is extremely important. Throughout the year, people might purchase a great variety of products to maintain basic living standards and promote happiness in daily life. As is known to all, people of different backgrounds or lifestyles might have different consumption habits. Given defined background information and features, how do humans' annual consumption vary? What are the estimated annual expenditures for that person? Are there any rules behind people's purchasing habits? As a subgroup of the AI-Powered Lifecycle Financial Planning project, we are interested in finding those "rules" and building models to understand better the purchasing trends for some people in future years and propose some advice for planning their future lives.

Our goal is to build predictive models useful for predicting the annual expenditures given some information related to the referenced person. To be specific, we divide the total expenditure into several subcategories, including subsistence

and discretionary expenditures, and predict them separately. What we have done so far is successfully building predictive models for one-person families on total annual expenditures with reasonable prediction errors and estimating the total individual expenditure.

1.2 Dataset Description

The dataset is from the Consumer Expenditure Surveys (CES) conducted by the Bureau of Labor Statistics (BLS). [1] The Consumer Expenditure Surveys program provides data on consumer expenditures, income, and demographic characteristics. The CES collects data through two surveys: the Quarterly Interview Survey (interview) and the Diary Survey(Diary). The Interview Survey is a rotating panel survey in which approximately 10,000 addresses are contacted each calendar quarter, yielding approximately 6,000 usable interviews. It generally tracks consumer units' large expenditures, such as major appliances and cars, and is conducted quarterly with each consumer unit. Meanwhile, the Diary Survey is a panel survey in which approximately 5,000 addresses are contacted each calendar quarter, yielding approximately 3,000 usable interviews. The Diary dataset tracks small, everyday expenditures and is conducted over two consecutive one-week periods with each respondent. Our project focuses on major expenditures, so we have chosen the Interview Survey dataset. Specifically,we use FMLI and MEMI files in the Interview Survey dataset.

The datasets FMLI 212/213/214/221 (located in the "interview21" and "interview22" folders) contain data on household consumption for each quarter. Each row represents an individual family identified by the NEWID. The data includes information on consumer characteristics, income, demographics, limited geography, weighting, and a summary of household expenditures.

The datasets MEMI 212/213/214/221 (located in the "interview21"and "interview22" folders) contain information about household members. Each row represents a single member of a household, and there are multiple records for each family. Members of the same family share the same NEWID. Unique records are defined by the combination of NEWID and MEMBNO. The dataset includes variables such as member income, employment information, demographics, and their relationship to the reference person.

[1] https://www.bls.gov/cex/

1.3 Dataset Preprocessing

The missing percentage for each column in MEMI dataset.

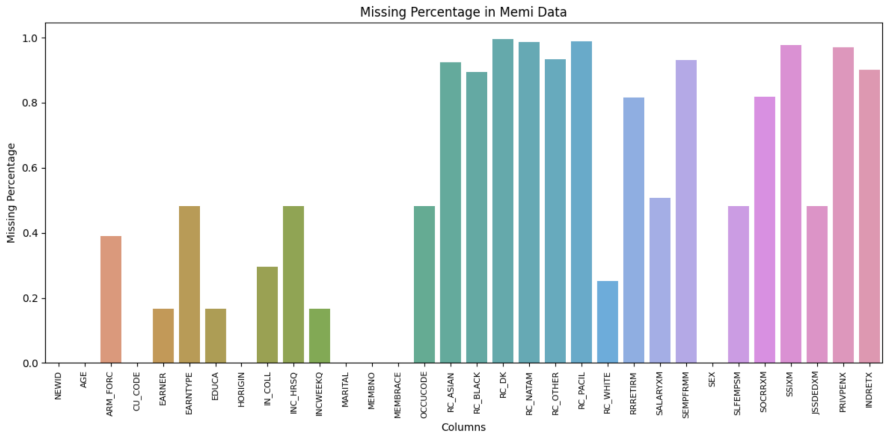


Figure 1: Plot missing percentage for MEMI data

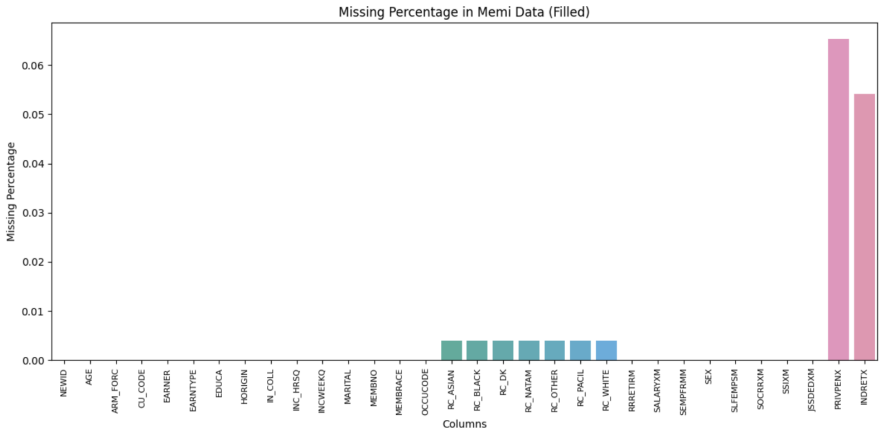


Figure 2: Comparison of missing percentage for MEMI data

The missing percentage for each column in FMLI dataset.

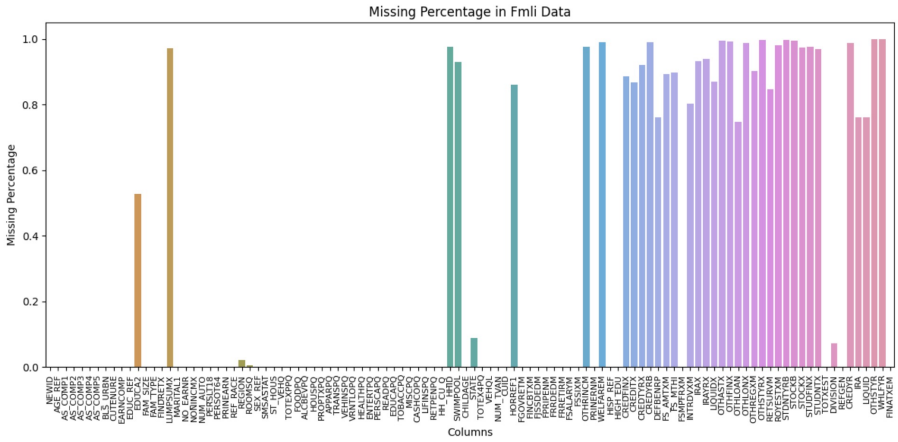


Figure 3: Plot missing percentage for FMLI data

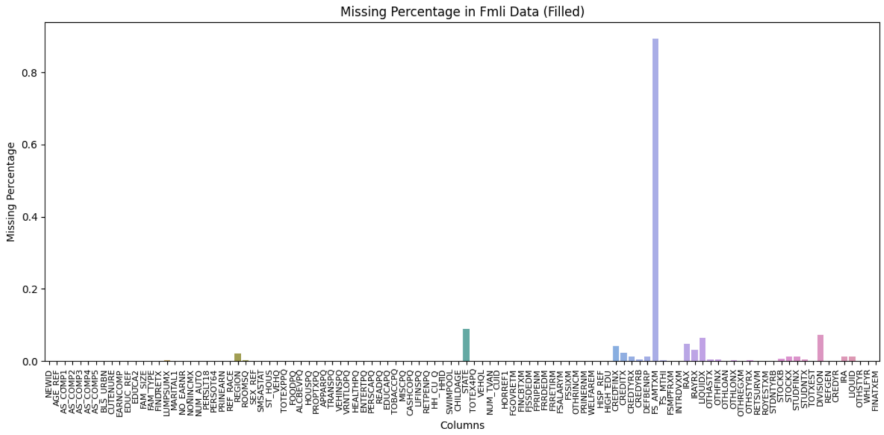


Figure 4: Comparison of missing percentage for FMLI data

We use the dataset FMLI (2021), which is a quarterly dataset. Each row represents a family with NEWID as a unique identifier. Each row is a household consumption for each quarter, including CU characteristics, income, demographics, geography (limited), weighting, a summary of expenditures at the household level.

MEMI is a quarterly dataset. Based on the NEWID in the FMLI dataset, it splits each family into members in that family with member characteristics. Each row is one member of a household. Members belonging to the same family have the same NEWID. The dataset contains variables like member income and employment information, member demographics, relationship to reference person.

The "Imputation Variables" section deals with filling in missing information in a dataset called "MEMI." Imagine this dataset as a collection of details about people in different families. Sometimes, some information is missing because people didn't provide it. To fix this, we use a process called "imputation." It is more like guessing what the missing info could be, so we can still understand the data well. This is important because we want to make sure our conclusions from the data are accurate.

The Consumer Expenditure Surveys (CE) include income data that have been produced using multiple imputation. The purpose of this procedure is to fill in blanks due to nonresponse (i.e., the respondent does not know or refuses to provide a value for a source of income received by the consumer unit or a member therein) in such a way that statistical inferences can be validly drawn from the data.

The following two tables show the overall missing rate of imputation variables before and after imputation for MEMI dataset.

Missing rate of MEMI dataset						
SALARYX	JSSDEDX	RRRETIRX	SEMPFRMX	SLFEMPSS	SOCRRX	SSIX
0.690566	0.488917	0.862367	0.968152	0.488917	0.170361	0.983516
0.683271	0.485488	0.854965	0.966248	0.485488	0.169167	0.984102
0.675789	0.480546	0.844365	0.963277	0.480546	0.162018	0.979103
0.665870	0.470700	0.850902	0.962252	0.470700	0.162367	0.979477

Table 1: Missing rate before imputation of MEMI dataset

The comparison shows that the overall missing rate of each imputation variables slightly decrease except for the SOCRRX variable.

The following table is the comparisons of missing rate of flag A and the overall missing rate of the imputation variables for the MEMI dataset.

The result shows that: Salary(SALARYX), Social Security Income(JSSDEDX),

Missing rate of MEMI dataset						
SALARYX	JSSDEDX	RRRETIRX	SEMPFRMX	SLFEMPSS	SOCRRX	SSIX
0.514604	0.488917	0.820997	0.931584	0.488917	0.824438	0.979995
0.512310	0.485488	0.816892	0.930540	0.485488	0.820642	0.980026
0.505465	0.480546	0.807030	0.930576	0.480546	0.810964	0.975431
0.497981	0.470700	0.814308	0.928789	0.470700	0.818676	0.975604

Table 2: Missing rate after imputation of MEMI dataset

"Missing rate of flag A" VS "Overall missing rate of the imputation variables"							
Rate	SALARYX	JSSDEDX	RRRETIRX	SEMPFRMX	SLFEMPSS	SOCRRX	SSIX
Flag A	0.514604	0.488917	0.824438	0.931584	0.488917	0.170361	0.980795
Original	0.690566	0.488917	0.862367	0.968152	0.488917	0.170361	0.983516
Imputed	0.514604	0.488917	0.820997	0.931584	0.488917	0.824438	0.979995

Table 3: MEMI Dataset Comparision

Self Employment Income or Loss(SEMPFRMX) and Self-Employment Social Security (SLFEMPSS) have the same missing rate of both flag A and imputed mean, indicating that after imputation, all the blanks with other flags are filled. However, there are slight difference between the flag A missing and imputed mean missing of the other variables. To be specific, the imputation procedure regards some original flag A as other flags and performed imputation, while we could not know the exact reason or steps of doing these.

We also have a table of comparisons of missing rate of flag A and the overall missing rate of the imputation variables for FMLI dataset.

"Missing rate of flag A" VS "Overall missing rate of the imputation variables"							
Rate	RETSURVX	INTRDVX	ROYESTX	NETRENTX	OTHREGX	WELFAREX	OTHRINCX
Flag A	0.852964	0.805427	0.981909	0.955543	0.889338	0.992109	0.978060
Original	0.881640	0.880100	0.984796	0.963818	0.908006	0.992687	0.979985
Imputed	0.848730	0.802733	0.980947	0.955158	0.886644	0.991724	0.976328

Table 4: FMLI Dataset Comparision

2 EDA

2.1 Target Variable

For One-person Family Data, our target value is **TOTEXPPQ** (Total expenditures last quarter.)

Here is a distribution of the target variable.

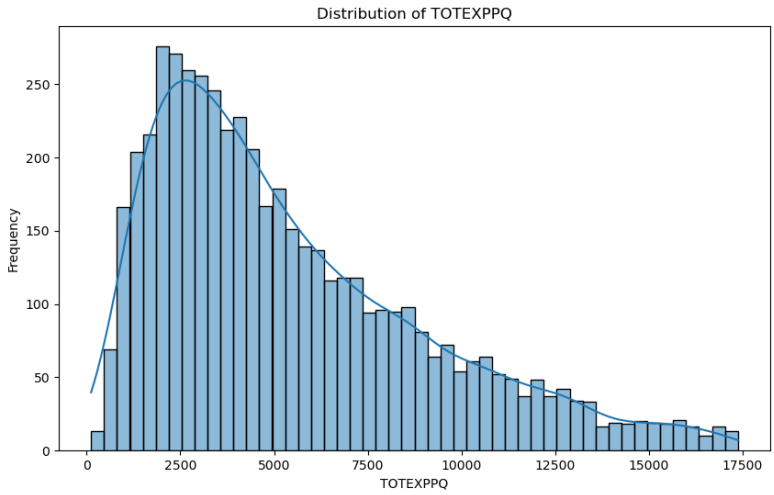


Figure 5: Distribution of TOTEXPPQ

Upon looking at the plot, we notice that the distribution is right-skewed. The tail on the right side of the distribution is longer and heavier, indicating positive kurtosis. This suggests that there is a higher concentration of data in the center of the distribution with a few extreme values on the right. This means that there are relatively few people with very high expenditures, while the majority of individuals have lower to moderate expenditures.

The spread of expenditure levels is relatively large, indicating substantial expenditure variability within the population. The mean expenditure is significantly higher than the median expenditure. This is a typical characteristic of right-skewed distributions, as the few high-expenditure individuals pull the mean to the right. The median expenditure is a more appropriate measure of central tendency in this case, as it is less affected by extreme values. It represents the expenditure level at which half of the population spends less, and half spends more.

2.2 Feature Variable

The clean data we used for the model has **32** features out of the target variable.

In order to facilitate the analysis, we group these variables by their meanings in Table 5.

subcategory	variable name	variable description
identity	AGE	What is the member's age?
	ARM_FORC	Is member now in the Armed Forces?
	EDUCA	What is the highest level of school the member has completed or the highest degree the member has received?
	IN_COLL	Is the member currently enrolled in a college or university either . . .?
	MEMBRACE	Race of member
	SEX	Sex of Member
relationship	CU_CODE	What is this person's relation to reference person?
	MARITAL	Marital status of member
location	BLS_URBN	Is this CU located in an urban or rural area
	DIVISION	Census Division
	FAM_SIZE	Number of Members in CU
	SMSASTAT	Does CU reside inside a Metropolitan Statistical Area (MSA)?
income	EARNER	Indicates whether the member earned income or not
	EARNTYPE	Type of earner
	INC_HRSQ	Number of hours worked per week
	INCWEEKQ	Number of weeks worked full time or part time (last 12 months)
	INTRDVXM	Amount of income received from interest and dividends, mean of the iterations
	OCCUCODE	The job in which the member received the most earnings during the past 12 months fits best in the following category.
	RETSURVM	Amount of income received from retirement, survivor, or disability pensions, mean of the iterations
	SEMPFRMM	Amount of income received from self-employment, mean of the iterations
	SOCRRXM	Annual amount received from Social Security benefits and Railroad payments, mean of imputation iterations.
	SSIXM	Amount received in supplemental security income checks combined, mean of imputation iterations.
retirement	FS_AMTXM	What was the dollar value of the last food stamps or EBT received (based on mean of imputation iterations)?
	IRAX	As of today, what is the total value of all retirement accounts, such as 401(k)s, IRAs, and Thrift Savings Plans that you own?
	JSSDEDXM	Social Security payment during the past 12 months, mean of imputation iterations.
	PRIVPENX	Amount of private pension deducted from last pay
	RRRETIRM	Amount of last Social Security or Railroad Retirement check, mean of imputation iterations.
asset	INTRDVXM	Amount of income received from interest and dividends, mean of the iterations
	LIQUIDX	As of today, what is the total value of all checking, savings, money market accounts, and certificated of deposit or CDs you have?
	NUM_AUTO	Total number of owned cars
	NUM_TVAN	Total number of owned trucks and vans
	STOCKX	As of today, what is the total value of all directly-held stocks, bonds, and mutual funds?

Table 5: Group of variables.

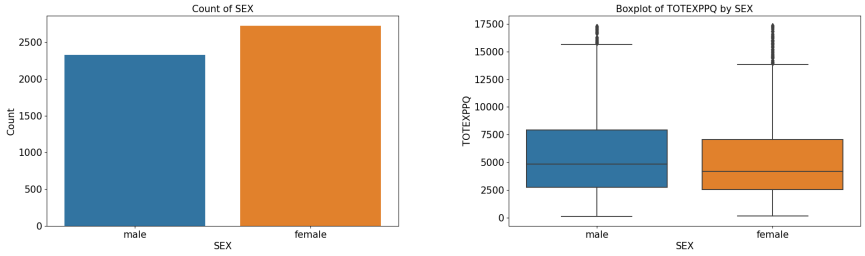
In order to have a clear look about the relationship between target variable and feature variables, we pick several examples from each group and split them between categorical and numerical variables.

2.2.1 Categorical Variable

Sex

We get the distribution of "SEX" variable and the boxplot of "TOTEXPPQ" by "SEX". In this data, we have more females than males.

The median of male is higher than the median of female, implies that, on average, male have a higher expenditure. The height of box for female is narrower than the box for male, showing that the expenditures in female are more tightly clustered around the median. The whiskers for female are shorter than those for male, it indicates that female has fewer extreme expenditures. Female has more outliers than male, it suggests that female may have more people with exceptionally high expenditures.



(a) Count of SEX

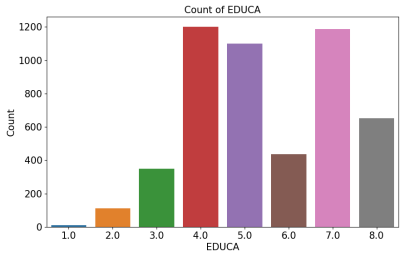
(b) Boxplot of TOTEXPPQ by SEX

Figure 6: SEX plots

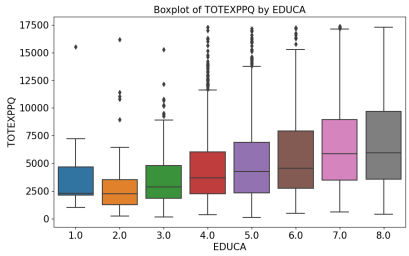
EDUCA

We get the distribution of "EDUCA" variable and the boxplot of "TOTEXPPQ" by "EDUCA". In this data, most people's education level is in the range of 4.0, 5.0, and 7.0.

Group 7.0 and 8.0 have a higher median, implies that, on average, they have a higher expenditure. The height of box for group 2.0 is narrower than boxes for other groups, showing that the expenditures in group 2.0 are more tightly clustered around the median. The whiskers for group 1.0 and 2.0 are shorter than those for other groups, it indicates that these groups have fewer extreme expenditures. Group 4.0 has more outliers than other groups, it suggests that group 4.0 may have more people with exceptionally high expenditures.



(a) Count of EDUCA



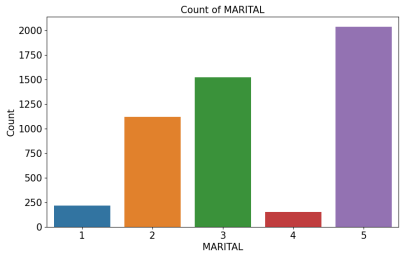
(b) Boxplot of TOTEXPPQ by EDUCA

Figure 7: EDUCA plots

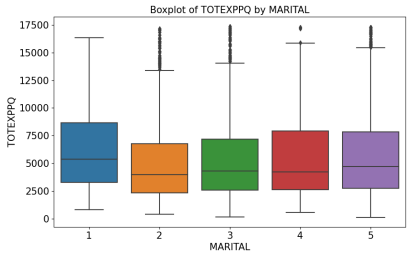
MARITAL

We get the distribution of "MARITAL" variable and the boxplot of "TOTEXPPQ" by "MARITAL". In this data, most people's marital level is in the range of 2, 3, and 5.

Group 1 has a higher median, implies that, on average, people in this group have a higher expenditure. The height of box for group 2 is narrower than boxes for other groups, showing that the expenditures in group 2 are more tightly clustered around the median. The whiskers for group 2 is shorter than those for other groups, it indicates that people in this groups have fewer extreme expenditures. Group 2 has more outliers than other groups, it suggests that group 2 may have more people with exceptionally high expenditures.



(a) Count of MARITAL



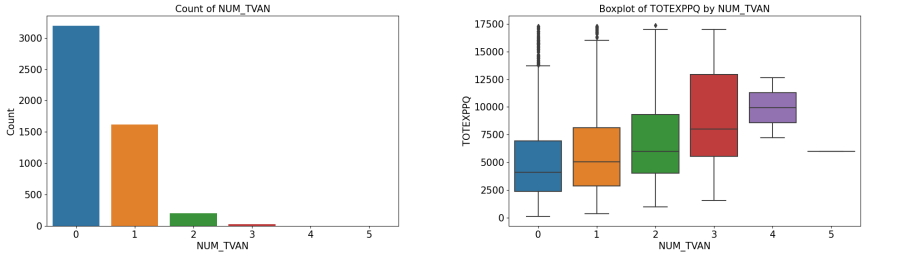
(b) Boxplot of TOTEXPPQ by MARITAL

Figure 8: MARITAL plots

NUM_TVAN

We get the distribution of "NUM TVAN"(Total number of owned trucks and vans) variable and the boxplot of "TOTEXPPQ" by "NUM TVAN". In this data, most people have 0 to 1 cars.

Group 4 has a higher median, implies that, on average, people in with 4 cars have a higher expenditure. The height of box for group 4 is narrower than boxes for other groups, showing that the expenditures in group 4 are more tightly clustered around the median. The whiskers for group 4 is shorter than those for other groups, it indicates that people in this groups have fewer extreme expenditures. Group 0 has more outliers than other groups, it suggests that group 0 may have more people with exceptionally high expenditures.



(a) Count of NUM_TVAN (b) Boxplot of TOTEXPPQ by NUM_TVAN

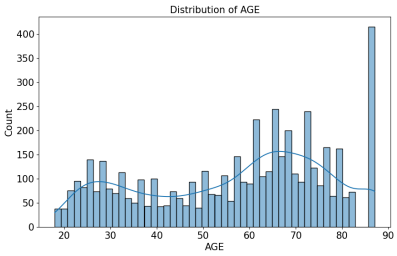
Figure 9: NUM_TVAN plots

2.2.2 Numerical Variable

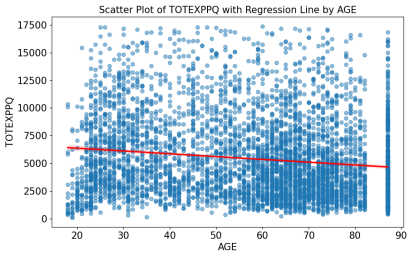
AGE

We get the distribution of "AGE" variable and the scatter Plot of "TOTEXPPQ" with Regression Line by "AGE". From the plot we find that there are more people in old age in this data, which might involves in retirement influences.

From the scatter plot we could see a negative slope trend indicating a negative correlation between "TOTEXPPQ" and "AGE". Also the points are not cluster together to the regression line, showing that this relationship is not strong enough. As there are more data in the old age, there might be outliers that have influence to this correlation.



(a) Distribution of AGE



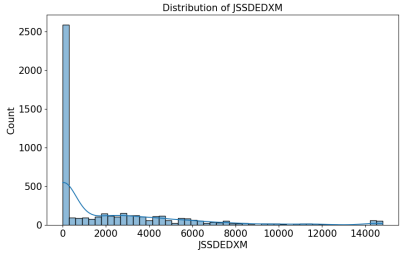
(b) Scatter Plot of TOTEXPPQ with Regression Line by AGE

Figure 10: AGE plots

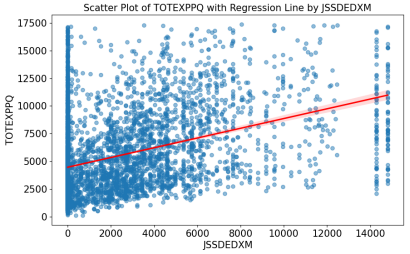
JSSDEDXM

We get the distribution of "JSSDEDXM"(Social Security payment during the past 12 months) variable and the scatter Plot of "TOTEXPPQ" with Regression Line by "JSSDEDXM". From the plot we find that older people has more retirement deposit than median people in this data.

From the scatter plot we could see a positive slope trend indicating a positive correlation between "TOTEXPPQ" and "JSSDEDXM". Also the points are cluster together to the regression line, showing that this relationship is quite strong. As there are more data in the old age, there might be outliers that have influence to this correlation.



(a) Distribution of JSSDEDXM



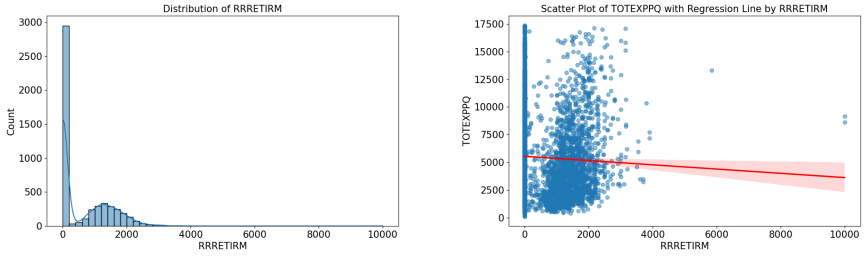
(b) Scatter Plot of TOTEXPPQ with Regression Line by JSSDEDXM

Figure 11: JSSDEDXM plots

RRRETIRM

We get the distribution of "RRRETIRM"(Amount of last Social Security or Railroad Retirement check) variable and the scatter Plot of "TOTEXPPQ" with Regression Line by "RRRETIRM". From the plot we find that most people have no social security deposit last quarter.

From the scatter plot we could see a negative slope trend indicating a negative correlation between "TOTEXPPQ" and "RRRETIRM". Also the points are cluster together to the regression line, showing that this relationship is quite strong. As there are some few data extremely away from the regression line in the plot, they might be outliers that have influence to this correlation.



(a) Distribution of RRRETIRM

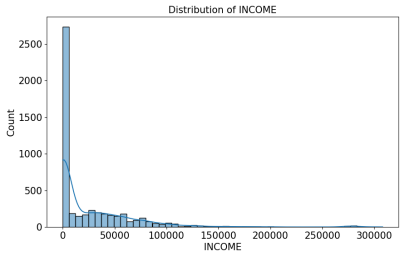
(b) Scatter Plot of TOTEXPPQ with Regression Line by RRRETIRM

Figure 12: RRRETIRM plots

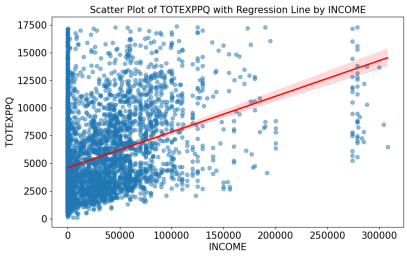
INCOME

We get the distribution of "INCOME" variable and the scatter Plot of "TOTEXPPQ" with Regression Line by "INCOME". From the plot we find that most people have no social security deposit last quarter.

From the scatter plot we could see a positive slope trend indicating a positive correlation between "TOTEXPPQ" and "INCOME". Also the points are cluster together to the regression line, showing that this relationship is quite strong. This slope is much steeper than plots from other features, indicates that "INCOME" is a big influence which make normal sense as more people you can get, more expenditure you could make.



(a) Distribution of INCOME



(b) Scatter Plot of TOTEXPPQ with Regression Line by INCOME

Figure 13: INCOME plots

2.3 Heat Map

We also create a heat map to check the correlation between each numerical variables.

Note that "SOCRRXM" and "RRRETIRM" has a high correlation of 0.93. It makes sense as "SOCRRXM" is annual social security retirement benefit while "RRETIRM" is the last social security retirement check.

Also we have "INC_HRSQ" and "INCWEEKQ" has a high correlation of 0.88. It makes sense as "INC_HRSQ" is the number of hours worked per week while "INCWEEKQ" is the number of weeks worked full time or part time.

"INC_HRSQ" and "AGE" has a high negative correlation of -0.61. It makes sense as people with old age will not be able to work more hours per week than young people.

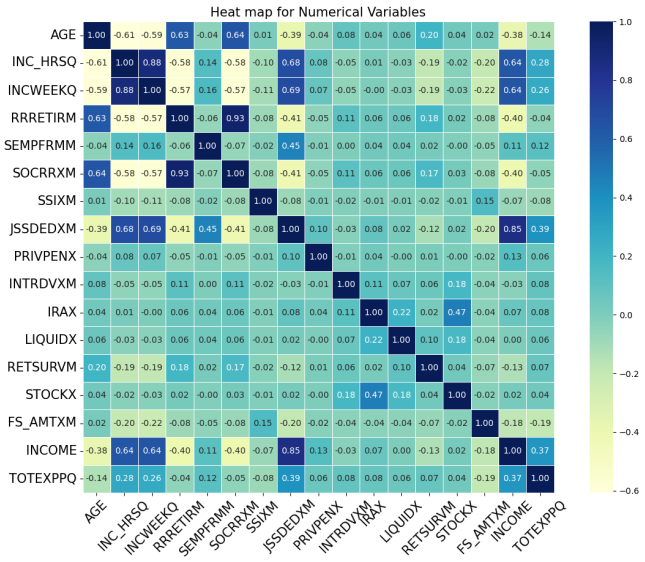


Figure 14: Heat map for Numerical Variables

3 Modeling

3.1 Models

3.1.1 Linear Regression

The simple linear regression assumes that the predictive variables and the response variable have linear relationships and different attributes have different weights (coefficients) in explaining the outcome. Given the fact that the expenditure is skewed, we would consider performing some kind of transformation on our response variable to make it much normally distributed. Thus, we use the Box-Cox transformation on the expenditure in which A Box-Cox transformation is a transformation of a non-normal dependent variable into a normal shape.

3.1.2 Ridge and Lasso regression

Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting, which may result from simple linear regression. In other words, these two models add penalty to the weights of the variables.

3.1.3 Random Forest

Random Forest is an ensemble learning method for regression that operates by constructing a multitude of decision trees at training time and using the average prediction of the individual trees as the final prediction. The model contains several important parameters such as the number of trees, max number of features, max number of levels, etc. We perform a random search to find the optimal values of hyper-parameters. Specifically, we use Scikit-Learn's RandomizedSearchCV method since it is faster than the grid search method, which allows us to try a larger range of parameter combinations. We first define a grid of hyper-parameter ranges in Table 4, and randomly sampled from the grid, performing 10-Fold CV with each combination of values. The best parameters are list in Table 5.

Table 4: Search grid for the tuning hyperparameters

Hyperparameter	Search Space
n_estimators	500, 800, 1000, 1500, 2000
max_depth	10, 15, 25, 30, 50
max_features	auto, sqrt, log2
min_sample_split	2, 10, 20
min_sample_leaf	2, 10

Table 5: Hyperparameter Tuning Result

Hyperparameter	Tuning Result
n_estimators	1000
max_depth	25
max_features	auto
min_sample_split	2
min_sample_leaf	2

3.1.4 LightGBM

LightGBM is a gradient boosting framework that employs tree-based learning algorithms. The hyper-parameters of LightGBm, such as the percent of features selected on each iteration, the max number of leaves, and the max depth in one tree are adaptively tuned with Bayesian hyper-parameter optimization and used to train the model. We constructed the objective function Bayesian Optimization for LightGBM to minimize the MSE of cross-validation. The predefined search space, which is the range of input parameters of the objective function is in Table 6. The best parameters after performing the Bayesian optimization for 200 iterations are listed in Table 7.

Table 6: Search grid for the tuning hyperparameters

Hyperparameter	Tuning Result
boosting_type	gbdt
objective	regression
feature_fraction	[0.6, 1.0]
learning_rate	[0.01, 0.2]
num_leaves	10, 20, ..., 100
max_depth	2, 4, ..., 30
reg_alpha	[0, 1.0]
reg_lambda	[0, 1.0]
subsample	[0.5, 1]

Table 7: Hyperparameter Tuning Result

Hyperparameter	Tuning Result
boosting_type	gbdt
objective	regression
feature_fraction	0.668423240033442
learning_rate	0.009912254641514136
num_leaves	100
max_depth	30
reg_alpha	0.7249290948226097
reg_lambda	0.7504450119058688
subsample	0.7912974290986281

3.2 Evaluation Metrics

For evaluating the performance of our regression models, since the outcome is continuous, we choose these four numeric metrics to calculate the error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$PE = \frac{|\hat{y}_i - y_i|}{y_i} * 100$$

And here is the summary of the result we get

Validation Error of Different Models

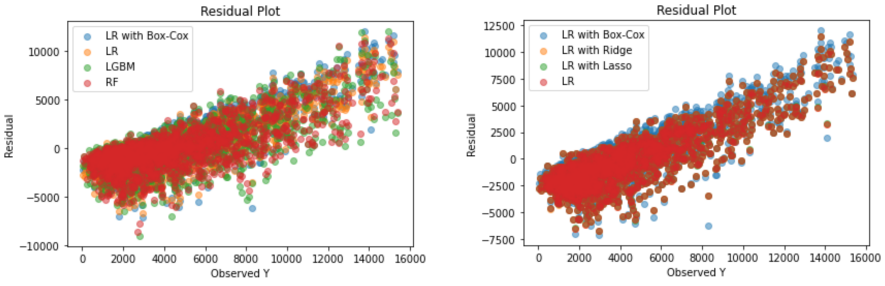
Model	RMSE	MAE	ME	PE
Standard Linear Regression	3189.5725	2500.2026	86.0477	46.2736
Ridge Regression	3187.7936	2498.5749	84.9046	46.2934
Lasso Regression	3188.3346	2498.6211	85.5951	46.2754
Linear Regression with Box-Cox	3276.5607	2454.9044	-563.5302	26.6354
Random Forest	2504.4378	1814.5928	78.3172	14.6379
LightGBM	2565.5247	1894.4659	73.3383	16.8491

Note that the Random Forest Regression model has the overall best performance with the lowest RMSE 2504.4378 and the lowest MAE 1814.5928.

3.3 Residual Plots

The residual is defined as the difference between the observed value and the predicted value. We plot the residual of each observation in the test data, and the smaller residual indicates that the predictions are closer to the ground truth and the model less prediction error.

Here We create two separate residual plots for more direct comparison. One is the comparison between linear models and tree-based models, and the other one is the comparison of different kinds of linear models.



(a) Residual plot

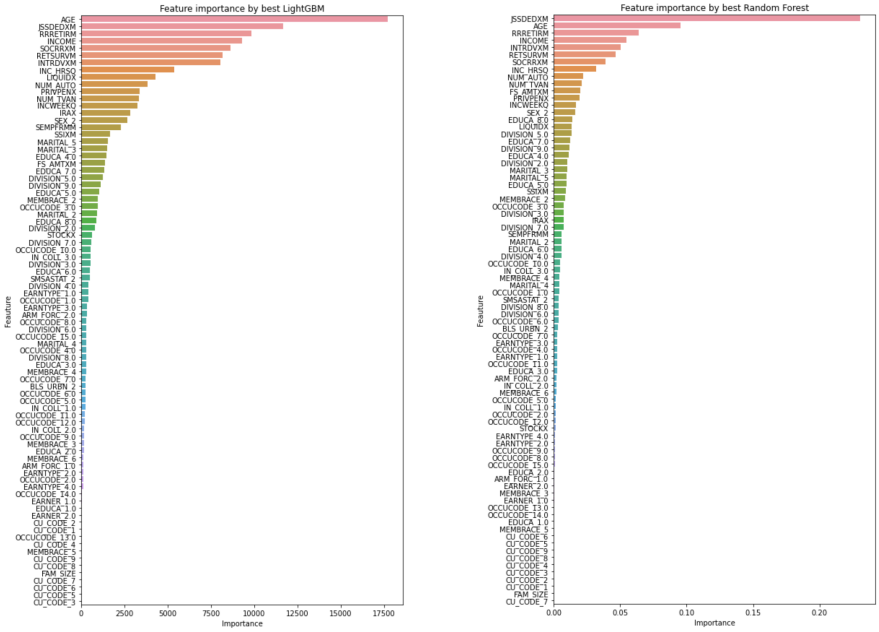
(b) Residual plot LR

Figure 15: Residual plots

Figure 15 shows that the random forest model has the best performance among all the models. Also, tree-based models are better than linear models, and they tend to have smaller residuals when it comes to the observations with relatively large Y.

3.4 Feature Importance

To get an intuition of which features are most related to the total expenditure, we plotted the feature importance of tree-based models. The feature importance plots generated from our random forest and lightgbm model.



(a) Feature Importance by LightGBM (b) Feature Importance by Random Forest

Figure 16: Feature Importance

We can see that the ranks of feature importance are basically similar between the two models. Another important takeaway is that, apart from age, the most important features are generally related to the amount of money, including the social security payment, the income from salary and self-employment, income from interest and dividends, income received from retirement, and so on.