Understanding Mediation effects through a questionnaire dataset:

The following analysis aims to establish the dependence of Person-Value fit in establishing how likely or probable is the person, on pursuing a bridge employment. The relevant analysis is made on a Likert scale questionnaire.

Feature Engineering:

To quantify and analyze the dependence, and intermediate mediation effects, relevant hypotheses are defined.

H1. PV fit is positively related to the intention to pursue bridge employment within the same organization

Indirect effects

H2. PJ fit mediates the positive relationship between PV fit and SOBE

H3. PO fit mediates the positive relationship between PV fit and SOBE

H4. The positive relationship between PV fit and SOBE is serially mediated by PJ fit and AOC

H5. The positive relationship between PV and SOBE is serially mediated by PO and AOC

With the questionnaire proving unstructured, specified feature engineering is done, in order to account for the hypothesized metrics, and make the dataset analysis ready.

The answers are each grouped with regards to the metrics. In Order to select the relevant answers, and the ideal combination to account for regression.

The filter is done by performing the Spearman's rank order correlation. For each grouped metric, piecewise correlation results and the overall picture, help decide if the answer could prove judgemental, without leading to unnecessary over fitting. The filter is thus made with respect to the hypothesized parameters.

The resulting variables are however still ordinal, and cannot be used for regression directly. Fortunately due to the present scaling, we combine like answers to account for the hypothesis, and assume continuity of the resultant, though a minor information loss is incurred.

The target variable is, however, not combined to ensure a better fit.

Modelling:

In accordance to the desired mediation tests from the path diagram, we define the individual regression equations (structural equations) as such.

```
model1 = 'PVFIT~POFIT

    AOC~POFIT+PVFIT

    SOBE~POFIT+PVFIT+AOC'
```

```
model2='PJFIT~PVFIT
    AOC~PJFIT+PVFIT
    SOBE~PJFIT+PVFIT+AOC'

model3='POFIT~PVFIT
    SOBE~PVFIT+POFIT'

model4='PJFIT~PVFIT
    SOBE~PVFIT+PJFIT'

model5='SOBE~PVFIT'
```

The models are subjected to linear and multiple regression testings, and the beta weights of each regression equation give us the path coefficients, in lieu with Bryman and Cramer's methods.The SME analysis is performed by coming in each model in the lavaan package in R. We obtain the weights of each path, by bootstrapping the results. This accounts for the indirect influence as well.

Model Performance:

Model	Adjusted R^2
Model 1	0.2995
Model 2	0.1386
Model 3	0.2875
Model 4	0.1386
Model 5	0.1711

The weights are added on to the path diagram, as such.

BOOTSTRAPPED WEIGHTS OF EACH CONNECTION

LHS	RHS	VALUE	STD.ERROR
PVFIT	POFIT	0.299539	0.003434
PVFIT	PJFIT	0.329482	0.010036
POFIT	AOC	0.219876	0.021833
PJFIT	AOC	0.221597	0.029303
AOC	SOBE	0.517558	0.231262
PVFIT	SOBE	0.679869	0.104761

PJFIT	SOBE	0.094462	0.075701
POFIT	SOBE	0.146694	0.065404

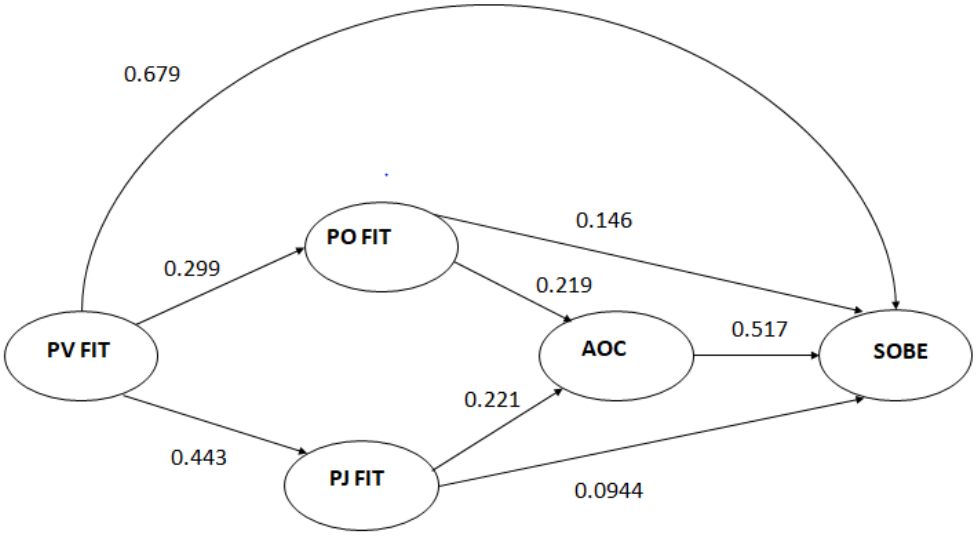


Fig:Path Diagram with the Coefficients

The Path coefficients help us to quantify the total effect of PV FIT upon SOBE (Intention to perform Bridge employment)

Direct Influence of PVFIT on SOBE: 0.679

Indirect Influence: 0.299*0.146 + 0.299*0.219*0.517 + 0.443*0.0944 + 0.443*0.21*0.517 = 0.146

Total effect = 0.679 + 0.146 = 0.825

This score justifies a positive impact of PV Fit on the intention to pursue bridge employment within the same organization. The Hypothesis H1 is thus accepted.

The mediations can be quantified, by means of the Sobel test. The Sobel test is used to determine if an attribute or a set of attributes mediate the effect of an independent variable to the dependent variable, the desired outcome. We test the mediation across each hypothesized paths under a significance level of 0.05.

The corresponding formula is given by:

Single mediating variable: **z-value = $x*y/ SQRT(y^2*SE_x^2 + x^2*SE_y^2)$**

Serial Mediation: z-value = $x*y*z / SQRT(x^2y^2*SE_z^2 + x^2z^2*SE_y^2 + y^2z^2*SE_x^2)$,

*where x,y,z are path coefficients, and SE_i denotes the corresponding standard error.

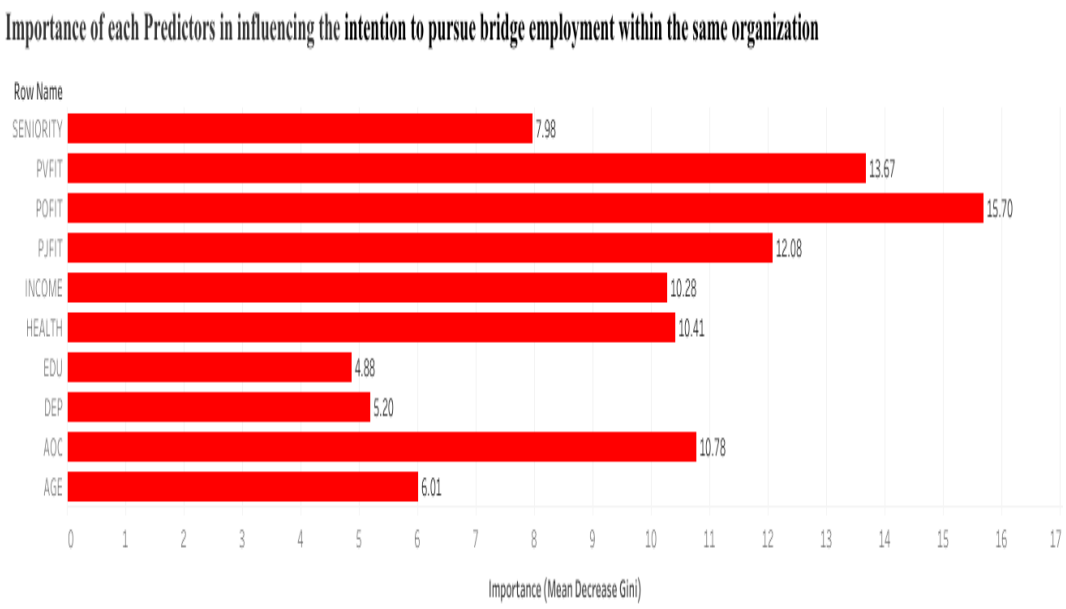
Mediation test	Z value	p value	Corresponding Hypothesis
PO FIT MEDIATES THE RELATION BEWEEN PVFIT AND SOBE	2.242137	0.024953	H3
PO FIT AND AOC SERIALLY MEDIATE THE RELATION BEWEEN PVFIT AND SOBE	2.183994	0.028963	H5
PJ FIT MEDIATES THE RELATION BEWEEN PVFIT AND SOBE	1.246934	0.212422	H2
PJ FIT AND AOC SERIALLY MEDIATE THE RELATION BEWEEN PVFIT AND SOBE	2.141401	0.032242	H4

Testing at a significance level of 0.05, we find that the mediation is significant, corresponding to hypothesis H3,H2,H4. However, Hypothesis H5 is not accepted, and the mediating effect is not significant enough.

To Assert how important each variable is with regards to the Outcome variable (SOBE) ,we accomplish a feature transformation on the combination of variables corresponding to SOBE. A sum total above 8 is taken as a threshold, beyond which the dependence is asserted as a YES. This binary categorization helps us to perform a classification test, thereby asserting the variable importance. Owing to the accomplishment of a Bootstrapping nature, we use a Random Forest Classification, so as to quantify the predictor importance, based on the Mean Decrease in Gini across each splits.

The corresponding results are visualized as such:

Fig:Variable importance



Classification Model:

The following are checked for compliance, wrt a Machine learning classifier approach. By virtue of supervised learning methods, we are able to establish a classifier algorithm that establishes predictions with a suitable accuracy.

Individual likert scale responses, are combined depending on the model representation, by reducing dimensions. A dimensionality reduction helps in combining dimensions, without incurring significant information loss. Owing to the ordinal nature of the data (5 point Likert Scale), we employ a correlation analysis based on the maximum likelihood estimate. A polychoric correlation matrix is layered on, and dimensions are further combined, by a Principal component Analysis approach. We arrive at a point with exactly 4 predictors.

With the target variable, categorized into a binary outcome, based on a threshold, we are set for a classification setup.

Two classifiers, in the form of a Support Vector Machine algorithm, and a Bernoulli outcome Gradient Boosting Machines, are used .

A 25:75 test train split is used, and a 10 fold cross validation is done, to select the best fit, and consequently quantify the model.

Cross Validation:

	Cross Validation Results		
Method	K value	Partition percentage	Error
SVM	5	80:20	1.293891
SVM	10	90:10	1.289048
GBM	5	80:20	1.160332
GBM	10	90:10	1.174139

Model Evaluation

Model	Accuracy	Precision	Recall	F1
GBM	0.8381	0.8928571	0.8156757	0.8692308
SVM	0.8116	0.8684211	0.804878	0.835443