11th Industrial Problem Solving Workshop

SOCIÉTÉ GÉNÉRALE PROBLEM

Topic

Categorical variables selection in risk modeling

Team members: Yichao Chen; William Désilets; Faith Lee; Bruno Monsia; Ahmed SidAli

Supervisors: Adrian Gonzalez-Sanchez; Helena Liu; Alejandro Murua; Jiaxin Yang

Introduction

Risk analysis is very important for financial institutions to identify different types of dangers and to make decisions:

- be in a position to take a decision as to whether to enter into a relationship or maintain an exisiting relationship with a client;
- evaluate the legitimacy of trasactions instructed by a customer regard to the information financial institutions have of such client.

Tools: Risk modeling to calculate the risk rating of clients based on their risk profiles.

Introduction:

Each risk profile (response variable) is assessed on the basis of 4 risk levels: *Low* , *Med-Low*, *Med-High* and *High*

There are continuous and categorical predictors on dataset.
All categorical variables are one-hot encoded before entering the model.

Questions :

1. How should the continuous variables be discretized and encoded, if at all?
2. How can-we select predictors for risk modeling so that monotonic relationship among the dummy coefficients be preserved ?
3. Which categorical predictor should be included in the model?
4. Which levels within one categorical predictor should be distinguished?

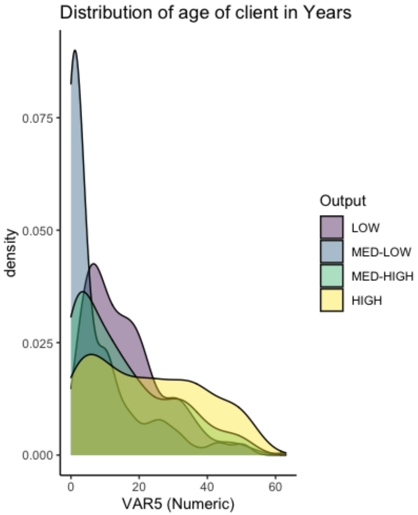
Descriptive statistics

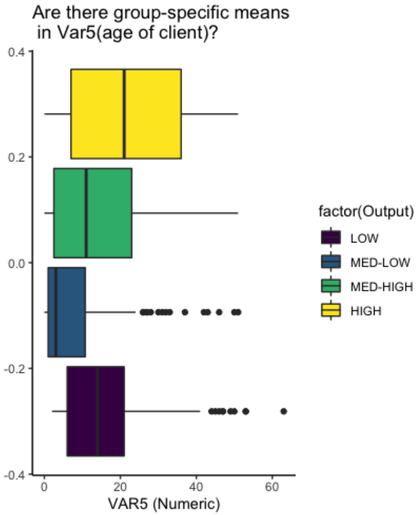
- Dataset contains 740 subjects, 12 predictors and output variable
- 119 cases of High risk; 296 of Low risk, 179 cases of Med-High and 146 cases of Med-Low (Does not seem strong case of Imbalanced classes)
- Variables 5,6,10,12,13 are numeric. The rest of the variables are ordered categorical variables

Is it relevant to discretize certain continuous variables ?

Cont. variables	Output	mean	std	min	25%	50%	75%	max
VAR10	HIGH	0.72	1.85	0	0	0	0	7
	LOW	0.45	0.63	0	0	0	1	3
	MED-HIGH	1.00	1.83	0	0	0	1	7
	MED-LOW	1.35	1.80	0	0	1	1	7
VAR12	HIGH	0.06	0.10	0	0	0.03	0.06	0.58
	LOW	0.01	0.07	0	0	0	0	0.58
	MED-HIGH	0.06	0.13	0	0	0	0.05	0.55
	MED-LOW	0.11	0.18	0	0	0	0.155	0.58
VAR13	HIGH	577.95	3960.60	0	0	0	2	41831
	LOW	1152.73	11244.04	0	0	0	27.25	175878
	MED-HIGH	117.88	581.94	0	0	0	8.5	6022
	MED-LOW	67.40	344.12	0	0	0	0	3482

Challenges during exploratory analysis (monotonicity)





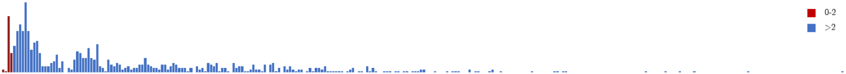
Challenges: (non-monotonicity) Higher number of years does not translate to lower risk.

In the original dataset, 4 variables[1] are encoded as one-hot categorical encoding

Each variable is split into only two classes, whose ranges appear to be mostly arbitrary. As a result, **the classes are strongly imbalanced**.

Distribution and discretization of **variable 5** – Age of the client (years), bucket width = 1 year

Distribution and discretization of **variable 6** – Length of business relationship (months), bucket width = 2 months



Distribution and discretization of **variable 10** – Number of risky product, bucket width = 1 product

Distribution and discretization of **variable 12** – Proportion of risky crossborder payments (%), bucket width = 10%

[1] Variables 5, 6, 10 and 12. Variable 13 is also continuous, but it was not discretized.

For each of the variables, 8 new discretization schemes were tested

3 parameters were considered to create new schemes:

1. Number of categories
   - 2, 3, 4 and 5 classes were considered
2. Type of categorical encoding[2]
   - Mean encoding - the mean of each class is used, and the variable is considered continuous in the modeling
   - One-hot "dummy" encoding - using K-1 binary variables
3. Categorical distribution[3],[4]
   - Equal distribution between classes - every class has the same number of observations[5]

---

[2] Harmonic mean encoding was considered, but could not be implemented within the time allowed

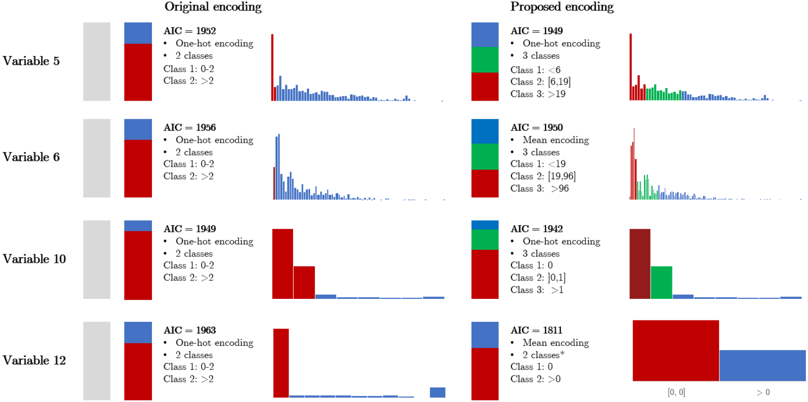
[3] The class distribution proposed by Pasta (2009) was implemented in the code but not tested

[4] Other possible distributions not explored within this workshop include: inferring class ranges from visual observations, clustering analysis, treating 0's as special values

[5] If the distribution is highly asymmetric and a value accounts for more than $\frac{100}{Nb\ categories}$ % of the values, classes will be imbalanced.

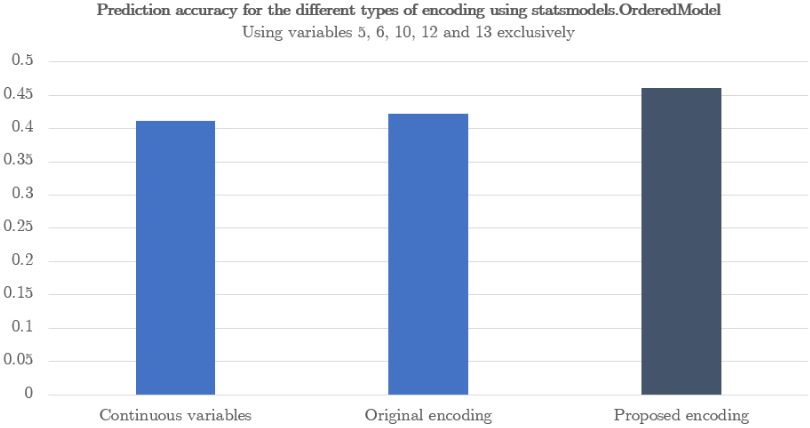
New discretization schemes were selected based on the AIC value of single-variable cumulative logit models

A cumulative logit model was adjusted for each of the discretization schemes. **The variable discretization scheme that yielded the model with the smallest AIC value was selected.**



The new discretization proposed increased prediction accuracy by 3.9%

However, the proposed discretization scheme proposed does not take into account possible interactions between variables and as such, should be extensively tested before being used.



Ordinal Logistic Regression

- We first tackle the problem by trying various classical algorithms
- The **Ordinal Logistic Regression** is a classical method used when the response variable is Ordinal
- We use the Python library **OrderedModel** from **scipy.stats** following these steps:
  - Data transformation ⇒ Creating dummy variables to make use of the nominal independent variables
  - Fitting the model with different sets of independent variables
  - Variable selection by comparing the performances of the different models using the **Akaike Information Criterion (AIC)** and the **Bayesian Information Criterion (BIC)**

Ordinal Logistic Regression-Complete model

OrderedModel Results

Dep. Variable:	Output	Log-Likelihood:	-332.19
Model:	OrderedModel	AIC:	708.4
Method:	Maximum Likelihood	BIC:	809.7
Date:	Thu, 26 Aug 2021		
Time:	18:25:58		
No. Observations:	740		
Df Residuals:	718		
Df Model:	22		

	coef	std err	z	P>|z|	[0.025	0.975]
VAR1_LOW	-9.5903	1.060	-9.043	0.000	-11.669	-7.512
VAR1_MED-LOW	-11.5317	1.149	-10.037	0.000	-13.784	-9.280
VAR1_MED-HIGH	-9.3654	1.062	-8.820	0.000	-11.446	-7.284
VAR2_NO	0.8400	0.428	1.962	0.050	0.001	1.679
VAR2_YES EQUIVALENT	-3.9074	0.507	-7.700	0.000	-4.902	-2.913
VAR3_LOW	-0.2891	0.213	-1.355	0.176	-0.707	0.129
VAR4_NO	2.2185	1.171	1.895	0.058	-0.076	4.513
VAR5_Discrete_1	1.2280	0.329	3.738	0.000	0.584	1.872
VAR5_Discrete_2	0.5070	0.333	1.524	0.128	-0.145	1.159
VAR6_Discrete_MEAN	-0.0062	0.002	-3.305	0.001	-0.010	-0.003
VAR7_YES NOT MATERIAL	-9.6501	1.047	-9.221	0.000	-11.701	-7.599
VAR7_NO	-10.3330	1.026	-10.071	0.000	-12.344	-8.322
VAR8_YES POTENTIAL	8.5025	1.347	6.313	0.000	5.863	11.142
VAR8_NO	-1.1847	0.379	-3.130	0.002	-1.927	-0.443
VAR9_NO	-9.9385	1.011	-9.833	0.000	-11.920	-7.957
VAR9_PRESENCE	-8.2455	0.991	-8.320	0.000	-10.188	-6.303
VAR10_Discrete_1	-1.9991	0.326	-6.124	0.000	-2.639	-1.359
VAR10_Discrete_2	-2.1351	0.378	-5.652	0.000	-2.875	-1.395
VAR12 (Numeric)	1.9468	0.815	2.390	0.017	0.350	3.544
LOW/MED-LOW	-32.1294	2.906	-11.055	0.000	-37.826	-26.433
MED-LOW/MED-HIGH	1.0820	0.077	13.968	0.000	0.930	1.234
MED-HIGH/HIGH	2.1665	0.110	19.675	0.000	1.951	2.382

Ordinal Logistic Regression-Reduced model 1: remove VAR3, VAR4, VAR5

OrderedModel Results

Dep. Variable:	Output	Log-Likelihood:	-342.62
Model:	OrderedModel	AIC:	721.2
Method:	Maximum Likelihood	BIC:	804.2
Date:	Fri, 27 Aug 2021		
Time:	07:15:37		
No. Observations:	740		
Df Residuals:	722		
Df Model:	18		

	coef	std err	z	P>|z|	[0.025	0.975]
VAR1_LOW	-8.9415	1.008	-8.866	0.000	-10.918	-6.965
VAR1_MED-LOW	-10.8260	1.090	-9.935	0.000	-12.962	-8.690
VAR1_MED-HIGH	-8.7208	1.009	-8.644	0.000	-10.698	-6.743
VAR2_NO	1.0088	0.419	2.409	0.016	0.188	1.830
VAR2_YES EQUIVALENT	-3.5876	0.487	-7.361	0.000	-4.543	-2.632
VAR6_Discrete_MEAN	-0.0097	0.002	-5.782	0.000	-0.013	-0.006
VAR7_YES NOT MATERIAL	-9.1676	1.045	-8.776	0.000	-11.215	-7.120
VAR7_NO	-9.8402	1.019	-9.661	0.000	-11.836	-7.844
VAR8_YES POTENTIAL	8.2074	1.354	6.063	0.000	5.554	10.861
VAR8_NO	-1.0519	0.369	-2.852	0.004	-1.775	-0.329
VAR9_NO	-9.3276	0.941	-9.907	0.000	-11.173	-7.482
VAR9_PRESENCE	-7.6601	0.927	-8.268	0.000	-9.476	-5.844
VAR10_Discrete_1	-1.9452	0.317	-6.127	0.000	-2.567	-1.323
VAR10_Discrete_2	-1.9707	0.367	-5.377	0.000	-2.689	-1.252
VAR12 (Numeric)	1.9475	0.803	2.425	0.015	0.373	3.522
LOW/MED-LOW	-32.9839	2.802	-11.772	0.000	-38.476	-27.492
MED-LOW/MED-HIGH	1.0482	0.077	13.560	0.000	0.897	1.200
MED-HIGH/HIGH	2.1268	0.111	19.223	0.000	1.910	2.344

Ordinal Logistic Regression-Reduced model 2: remove VAR3, VAR4, VAR12

OrderedModel Results

Dep. Variable:	Output	Log-Likelihood:	-337.55
Model:	OrderedModel	AIC:	713.1
Method:	Maximum Likelihood	BIC:	800.6
Date:	Fri, 27 Aug 2021		
Time:	07:57:11		
No. Observations:	740		
Df Residuals:	721		
Df Model:	19		

	coef	std err	z	P>|z|	[0.025	0.975]
VAR1_LOW	-9.4570	1.031	-9.176	0.000	-11.477	-7.437
VAR1_MED-LOW	-11.3452	1.115	-10.178	0.000	-13.530	-9.160
VAR1_MED-HIGH	-9.1998	1.033	-8.907	0.000	-11.224	-7.175
VAR2_NO	0.8942	0.425	2.106	0.035	0.062	1.726
VAR2_YES EQUIVALENT	-3.6829	0.494	-7.460	0.000	-4.651	-2.715
VAR5_Discrete_1	1.1801	0.325	3.629	0.000	0.543	1.817
VAR5_Discrete_2	0.4588	0.327	1.402	0.161	-0.183	1.100
VAR6_Discrete_MEAN	-0.0064	0.002	-3.458	0.001	-0.010	-0.003
VAR7_YES NOT MATERIAL	-9.1556	1.027	-8.916	0.000	-11.168	-7.143
VAR7_NO	-9.9559	1.011	-9.844	0.000	-11.938	-7.974
VAR8_YES POTENTIAL	8.0509	1.294	6.222	0.000	5.515	10.587
VAR8_NO	-1.3819	0.368	-3.750	0.000	-2.104	-0.660
VAR9_NO	-9.7369	0.979	-9.945	0.000	-11.656	-7.818
VAR9_PRESENCE	-7.9329	0.955	-8.309	0.000	-9.804	-6.062
VAR10_Discrete_1	-1.9938	0.323	-6.174	0.000	-2.627	-1.361
VAR10_Discrete_2	-2.1912	0.374	-5.863	0.000	-2.924	-1.459
LOW/MED-LOW	-33.6562	2.835	-11.870	0.000	-39.214	-28.099
MED-LOW/MED-HIGH	1.0644	0.077	13.765	0.000	0.913	1.216
MED-HIGH/HIGH	2.1461	0.109	19.702	0.000	1.933	2.360

Ordinal Logistic Regression-Summary

Models	Log-Likelihood	AIC	BIC
Complete Model	-332.19	708.4	809.7
Reduced Model 1	-342.62	721.2	804.2
Reduced Model 2	-337.55	713.1	800.6

Note: The same strategy can be used to test the other models!

Modeling analysis

Because the response variable is ordered, we proposed to use fused Lasso regularization (Tibshirani et al., 2005) with has the property of ordering the predictors and the metrical responses.

- proportional odds cumulative logit model
- Cumulative ordinal logistic regression model with fused Lasso regularization

The Proportional Odds Model

The proportional odds cumulative logit model is one of the commonly used methods for fitting ordinal response data. For an outcome with j=4 levels in increasing order and an n x p covariate matrix $\mathbf{X}$,

$$\text{logit}P(Y \leq j|\mathbf{X}) = \alpha_j + \beta^T\mathbf{X}, \text{j} = \text{1, ..., J-1} \tag{1}$$

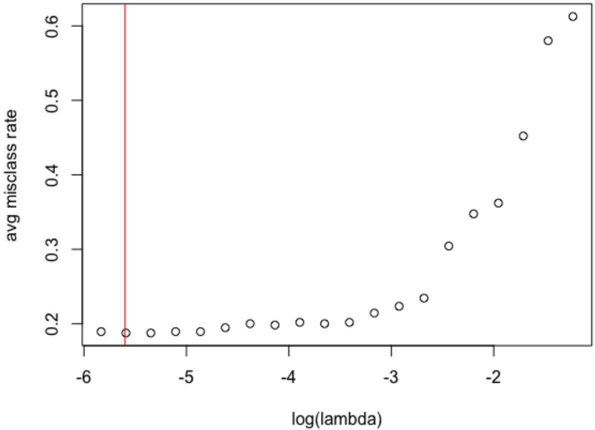
.

This means that the **cumulative probability** for a certain level of response **j**

$$P(Y \leq j|\mathbf{X}) = \frac{\exp(\alpha_j + \beta^T\mathbf{X})}{1 + \exp(\alpha_j + \beta^T\mathbf{X})} \tag{2}$$

The Proportional Odds Model: Variable Selection

We use `ordinalNet` for variable selection. At the best λ ($\lambda = 0.0037$) returned by the ordinalNet, it shows that there is some penalization



but not much.

Proportional Odds - Original

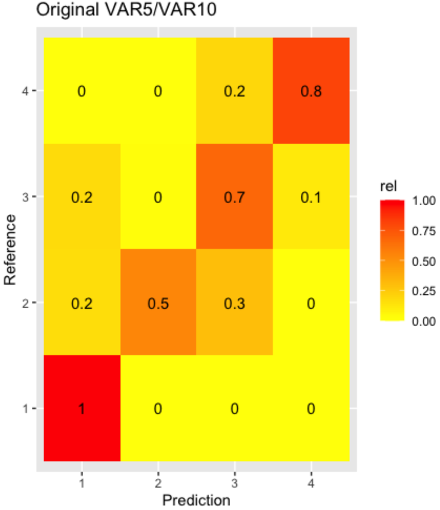
```
vglm(formula = as.numeric(Output) ~ VAR1 + VAR3 + VAR4 + `VAR5 (Discrete)` +
    VAR7 + VAR8 + VAR9 + `VAR10 (Discrete)`, family = cumulative(parallel = TRUE),
    data = IPSW_dictionary_vF)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept):1           0.9062     1.4558   0.622 0.533626
(Intercept):2           2.9374     1.4626   2.008 0.044611 *
(Intercept):3           9.6919     1.7022   5.694 1.24e-08 ***
VAR1MED-LOW             0.6446     0.2636   2.445 0.014466 *
VAR1MED-HIGH           -2.5501     0.2386 -10.689  < 2e-16 ***
VAR1HIGH               -9.4036     0.9139 -10.290  < 2e-16 ***
VAR3MED-LOW            -0.1508     0.1856  -0.812 0.416607
VAR4NO                 -0.6775     1.4093  -0.481 0.630702
`VAR5 (Discrete)`0-2   -1.4931     0.2219  -6.729 1.71e-11 ***
VAR7YES NOT MATERIAL    0.1473     0.2523   0.584 0.559292
VAR7YES MATERIAL       -7.5952     0.9141  -8.309  < 2e-16 ***
VAR8YES LIMITED        -1.0665     0.3336  -3.197 0.001391 **
VAR8YES POTENTIAL      -6.9746     1.1626  -5.999 1.99e-09 ***
VAR9PRESENCE           -0.6770     0.2262  -2.993 0.002764 **
VAR9CLIENT             -7.0487     0.8650  -8.148 3.69e-16 ***
`VAR10 (Discrete)`0-2   1.1623     0.3306   3.516 0.000438 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]), logitlink(P[Y<=

Residual deviance: 919.632 on 2204 degrees of freedom

Log-likelihood: -459.816 on 2204 degrees of freedom
```

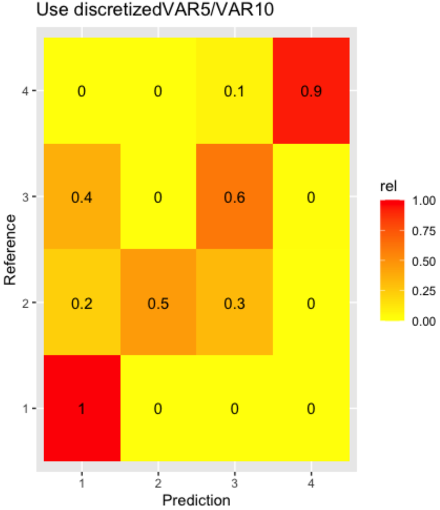
Proportional Odds Model- Discrete Variables

Ordinal Regression: Retain the order of outcome risk variables.
We use K -fold cross validation on the following variables: VAR1, VAR3, VAR4, VAR5(DISCRETE), VAR7, VAR8, VAR9, VAR10(DISCRETE) on a 80/20 split on the dataset

K	Train	Test	Train (Discretized)	Test (Discretized)
5	80,0	78,5	77.5	76,8
10	79,6	78,1	77,2	76,1
20	79,9	78,6	77,3	76,2

Confusion Matrix





Non Proportional Odds Model-Discrete Original Variables

To the same subset of variables selected from the ordinalNet previously, we fit an non-proportional odds model and found that it fares poorly than before.

Number of folds, K	Train	Test
5	76,8%	75,9%
10	73,8	71,5
20	73,6	71,5

Cumulative multinomial logistic regression penalized with Fused Lasso (Tibshirani et al., 2005)

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ l\left(\boldsymbol{\beta}\right) - FL_\lambda(\beta) \right\} \text{ with} \tag{3}$$

$$FL_\lambda(\beta) = \lambda \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \tag{4}$$

- select predictors with more influence on the response variable
- order predictors and the metrical responses.
- After the reparametrization of model ($\delta_j = \beta_j - \beta_{j-1}$) as suggested by Gertheiss & Tutz we used used R package *OrnalNet* with ordinary Lasso penalization.

Results of Fused Lasso regularization

Var.	intercept (MED-LOW)	intercept(MED-HIGH)	intercept(HIGH)	VAR1-HIGH	
Coef.	-4.53	-2.01	4.70	-5.33	
Var.	VAR1-MED_LOW	VAR1-MED_HIGH	VAR2-NO	VAR2-YES_EQUIVALENT	VAR3-MED_LOW
Coef.	1.68	2.74	-0.93	1.58	-0.23
Var.	VAR4-YES_EQUIVALENT	VAR5-Discrete_1	VAR5-Discrete_2	VAR6-Discrete_MEAN	VAR7-NO
Coef.	1.52	0.26	0.39	0.00	1.14
Var.	VAR7-YES_MATERIAL	VAR8-NO	VAR8-YES_POTENTIAL	VAR9-NO	VAR9-CLIENT
Coef.	-5.20	1.42	-3.12	1.80	-1.67
Var.	VAR10-Discrete_1	VAR10-Discrete_2	VAR12-Discrete_MEAN	VAR-13	
Coef.	0.00	-0.12	-7.69	0.00	

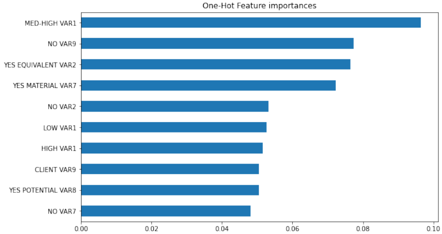
- Variables that have zero coefficients mean that they are irrelevant to the model
- dummy variables have zero coefficients mean the associated dummy variables modalities should have the same label as the reference modality.

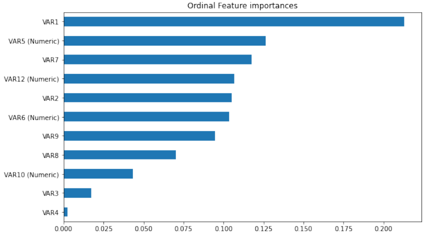
Random Forest Model

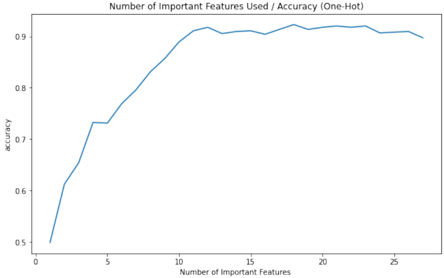
We tested some classic machine learning model like random forest model with two schemes of encoding: one-hot and ordinal. One-hot encoding outperforms.

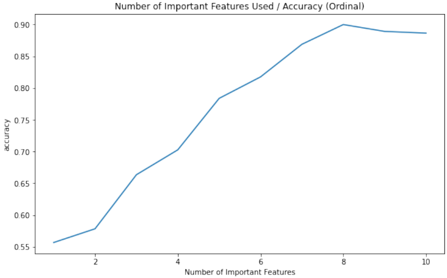
# of folds, K	Test Accuracy (One-Hot)	Test Accuracy (Ordinal)
3	88.6%	87.7%
5	90.1%	89.5%
10	91.6%	90.0%

Feature Importance

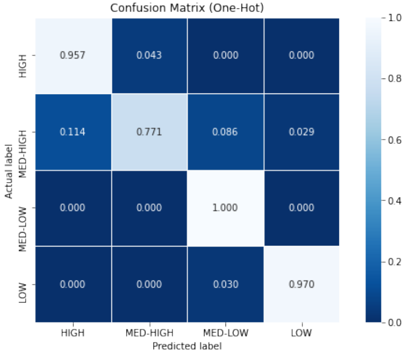


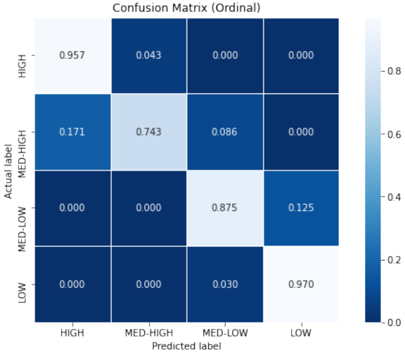






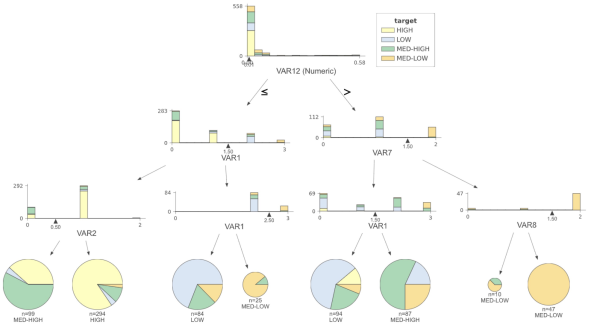
Confusion Matrix





Single Decision Tree

We then trained a single decision tree model. The trained decision tree model has maximum depth of 12 and test accuracy of 87.3% with 5-fold CV. For the purpose of illustration, we trained another decision tree of depth 3 which gave test accuracy 71.9% with 5-fold CV.



Conclusion

- In this project we would like to perform efficient feature selection that respects the intrinsic monotonicity within each categorical variable while giving reliable prediction accuracy for the ordinal response.
- We performed exploratory analysis on the dataset and tested various discretization schemes for numeric variables and selected a best scheme that boosts the performance of a benchmark model.
- We then implemented various models for comparison including Ordinal Logistic Regression, Proportional Odds Cumulative Logit Model, Cumulative Multinomial Logistic Regression Penalized with Fused Lasso and Random Forest Model
- Balance between the interpretability and prediction accuracy need to be found for the candidate models.