Automating Business Modeling with the AutoModelR package

Derek McCrae Norton

InterContinental Hotels Group

Outline

Outline

Why AutoModelR?

... or should I keep my eyes open for the remainder of this talk?

- The modeling process can be lengthy.
- A standardized process can be defined.
- Many steps can be simple in implementation, if not in scope.
- A standardized report can aid in understanding.

Why AutoModelR?

... or should I keep my eyes open for the remainder of this talk?

- The modeling process can be lengthy.
- A standardized process can be defined.
- Many steps can be simple in implementation, if not in scope.
- A standardized report can aid in understanding.

Why AutoModelR?

... or should I keep my eyes open for the remainder of this talk?

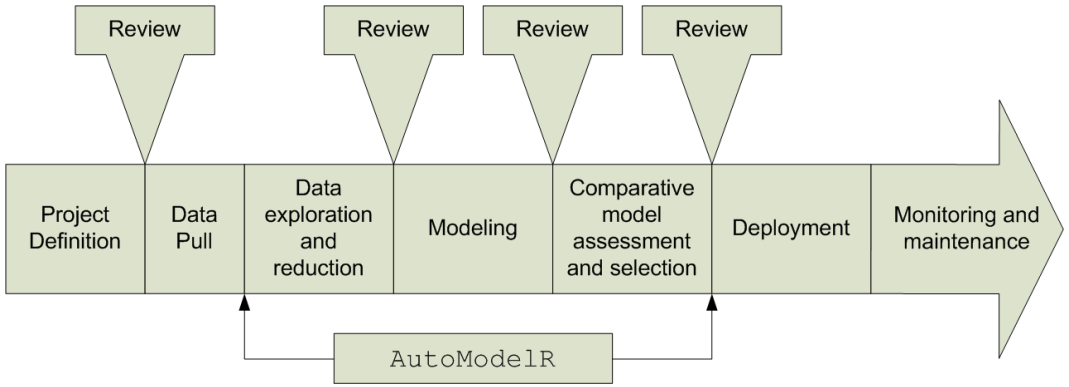
- The modeling process can be lengthy.
- A standardized process can be defined.
- Many steps can be simple in implementation, if not in scope.
- A standardized report can aid in understanding.

Why AutoModelR?

... or should I keep my eyes open for the remainder of this talk?

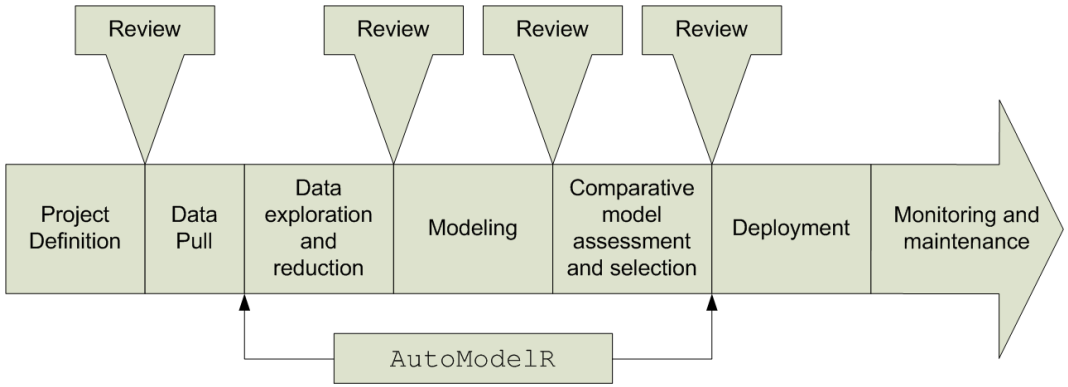
- The modeling process can be lengthy.
- A standardized process can be defined.
- Many steps can be simple in implementation, if not in scope.
- A standardized report can aid in understanding.

Modeling Process



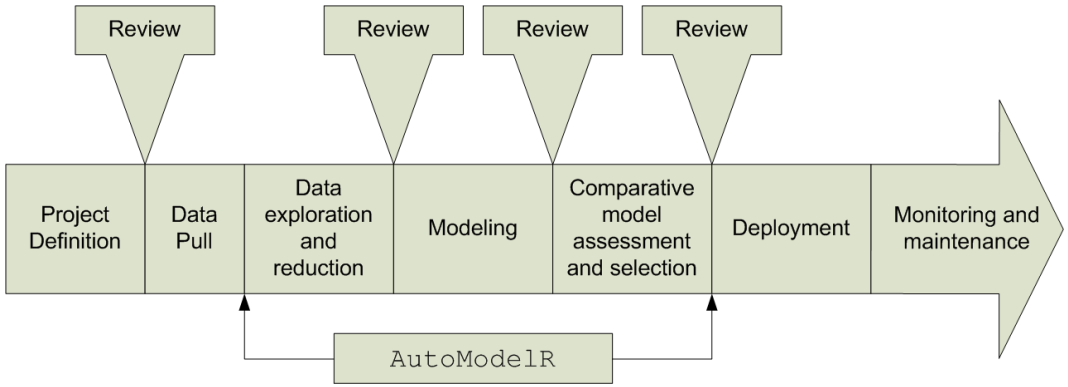
- Process is standardized in a broad sense.
- Process is followed with or without automation.
- When 40+ requests are sitting in queue, any savings of time is *greatly* appreciated.

Modeling Process



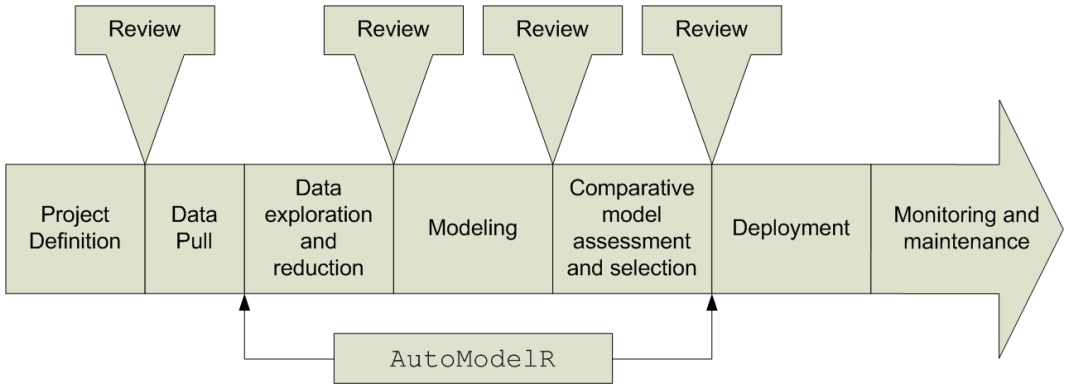
- Process is standardized in a broad sense.
- Process is followed with or without automation.
- When 40+ requests are sitting in queue, any savings of time is *greatly* appreciated.

Modeling Process



- Process is standardized in a broad sense.
- Process is followed with or without automation.
- When 40+ requests are sitting in queue, any savings of time is *greatly* appreciated.

Modeling Process



- Process is standardized in a broad sense.
- Process is followed with or without automation.
- When 40+ requests are sitting in queue, any savings of time is *greatly* appreciated.

Outline

What is AutoModelR?

- Automation of three standardized tasks.
  1. Data Exploration and Reduction
  2. Modeling
  3. Model Assesment and Selection
- Most of the focus is on task 1.
- Tasks 2 and 3 are generally first steps to more involved work.

What is AutoModelR?

- Automation of three standardized tasks.
  1. Data Exploration and Reduction
  2. Modeling
  3. Model Assesment and Selection
- Most of the focus is on task 1.
- Tasks 2 and 3 are generally first steps to more involved work.

What is AutoModelR?

- Automation of three standardized tasks.
  1. Data Exploration and Reduction
  2. Modeling
  3. Model Assesment and Selection
- Most of the focus is on task 1.
- Tasks 2 and 3 are generally first steps to more involved work.

What is AutoModelR?

- Automation of three standardized tasks.
  1. Data Exploration and Reduction
  2. Modeling
  3. Model Assesment and Selection
- Most of the focus is on task 1.
- Tasks 2 and 3 are generally first steps to more involved work.

What is AutoModelR?

- Automation of three standardized tasks.
  1. Data Exploration and Reduction
  2. Modeling
  3. Model Assesment and Selection
- Most of the focus is on task 1.
- Tasks 2 and 3 are generally first steps to more involved work.

What is AutoModelR?

- Automation of three standardized tasks.
  1. Data Exploration and Reduction
  2. Modeling
  3. Model Assesment and Selection
- Most of the focus is on task 1.
- Tasks 2 and 3 are generally first steps to more involved work.

Outline

General Steps

- Drop variables with a high percentage of missing values.
- Create indicator variables for presence in variables with a medium percentage of missing values.
- Impute variables with a low percentage of missing values.
- Drop variables with zero variation.
- Add log transformed variables.
- Conduct basic filter variable selection based on relevance and redundancy.

General Steps

- Drop variables with a high percentage of missing values.
- Create indicator variables for presence in variables with a medium percentage of missing values.
- Impute variables with a low percentage of missing values.
- Drop variables with zero variation.
- Add log transformed variables.
- Conduct basic filter variable selection based on relevance and redundancy.

General Steps

- Drop variables with a high percentage of missing values.
- Create indicator variables for presence in variables with a medium percentage of missing values.
- Impute variables with a low percentage of missing values.
- Drop variables with zero variation.
- Add log transformed variables.
- Conduct basic filter variable selection based on relevance and redundancy.

General Steps

- Drop variables with a high percentage of missing values.
- Create indicator variables for presence in variables with a medium percentage of missing values.
- Impute variables with a low percentage of missing values.
- Drop variables with zero variation.
- Add log transformed variables.
- Conduct basic filter variable selection based on relevance and redundancy.

General Steps

- Drop variables with a high percentage of missing values.
- Create indicator variables for presence in variables with a medium percentage of missing values.
- Impute variables with a low percentage of missing values.
- Drop variables with zero variation.
- Add log transformed variables.
- Conduct basic filter variable selection based on relevance and redundancy.

General Steps

- Drop variables with a high percentage of missing values.
- Create indicator variables for presence in variables with a medium percentage of missing values.
- Impute variables with a low percentage of missing values.
- Drop variables with zero variation.
- Add log transformed variables.
- Conduct basic filter variable selection based on relevance and redundancy.

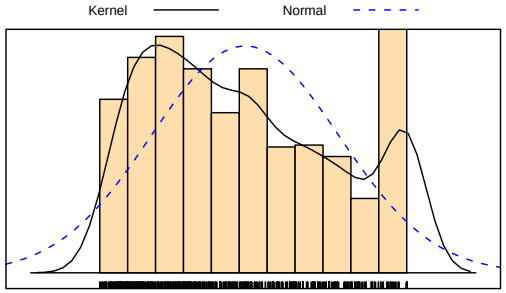
General Steps II

- Create summary tables.
- Create plots. (Thanks to Jim Porzak)

IHG_TOTAL_SHARE

MIN	MEDIAN	MAX	MEAN	STD DEV
0.0125	0.4286	1	0.4798	0.3008

N	UNIQUE	MISSING	SKEWNESS	KURTOSIS
1000	326	0	0.3837	−1.066



Histogram, Kenrnel Density, and Normal Curve

MI_EMAIL_PREF_VAL_CD

	Label	Frequency	Barchart
1	25	31.6	
2	9	18.6	
3	8	15.7	
4	0	15.2	
5	17	4.7	
6	1	4.2	
7	24	4	
8	57	1.6	
9	13	1	
10	** Others (23 Levels) **	3.4	

Outline

Modeling Goals

- As yet incomplete.
- Create a portfolio of models for particular types of dependent variables
  - Continuous
  - Binary Categorical
  - Multicategory Categorical

Modeling Goals

- As yet incomplete.
- Create a portfolio of models for particular types of dependent variables
  - Continuous
  - Binary Categorical
  - Multicategory Categorical

Modeling Goals

- As yet incomplete.
- Create a portfolio of models for particular types of dependent variables
  - Continuous
  - Binary Categorical
  - Multicategory Categorical

Summary

- The initial goal of `AutoModelR` was to automate as much as possible of the Exploratory Data Analysis in the modeling process.
  - This is achieved with a `Sweave` document that is intelligent enough to make choices similar to mine.
- The goal is now to expand this to include automated initial modeling and model comparison.

Outlook

- Continue to add initial models to be fit.
- Consider a rewrite specifically for time series data.

Summary

- The initial goal of `AutoModelR` was to automate as much as possible of the Exploratory Data Analysis in the modeling process.
  - This is achieved with a `Sweave` document that is intelligent enough to make choices similar to mine.
- The goal is now to expand this to include automated initial modeling and model comparison.

Outlook

- Continue to add initial models to be fit.
- Consider a rewrite specifically for time series data.

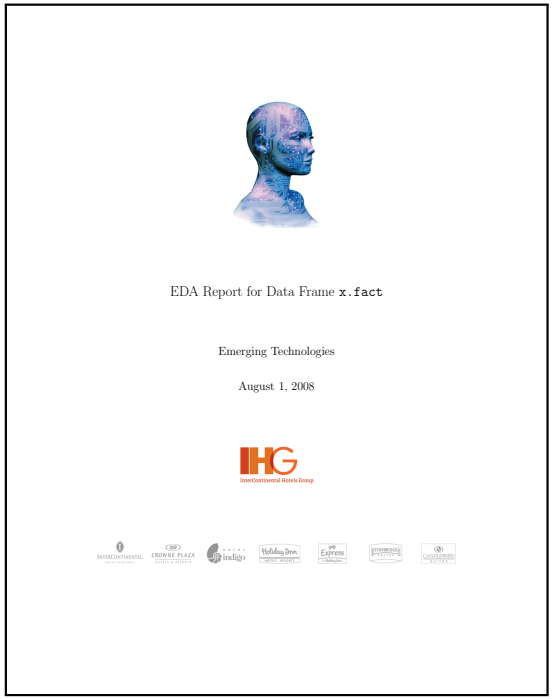
Summary

- The initial goal of `AutoModelR` was to automate as much as possible of the Exploratory Data Analysis in the modeling process.
  - This is achieved with a `Sweave` document that is intelligent enough to make choices similar to mine.
- The goal is now to expand this to include automated initial modeling and model comparison.

Outlook

- Continue to add initial models to be fit.
- Consider a rewrite specifically for time series data.

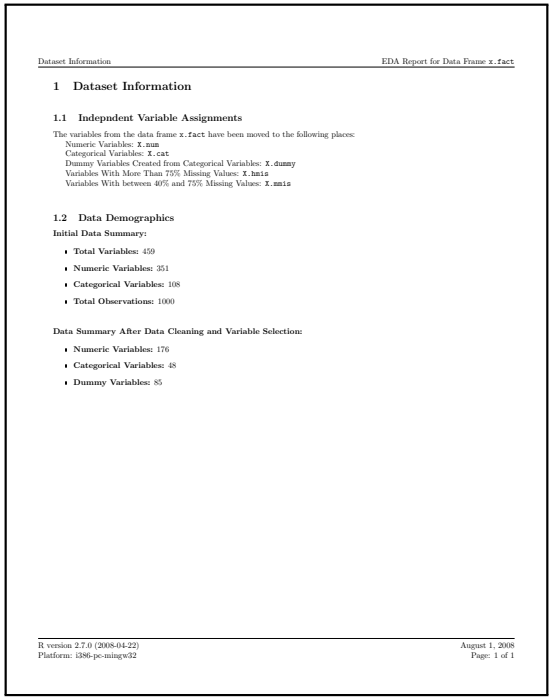
Screenshots



EDA Report for Data Frame `x.fact`

Emerging Technologies

August 1, 2008



Dataset Information EDA Report for Data Frame `x.fact`

1 Dataset Information

1.1 Indepndent Variable Assignments

The variables from the data frame `x.fact` have been moved to the following places:
Numeric Variables: `X.num`
Categorical Variables: `X.cat`
Dummy Variables Created from Categorical Variables: `X.dummy`
Variables With More Than 75% Missing Values: `X.hmis`
Variables With between 40% and 75% Missing Values: `X.mmis`

1.2 Data Demographics

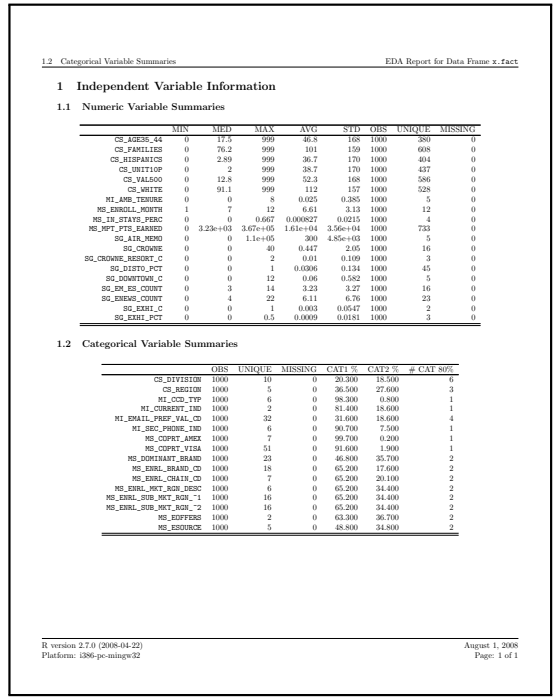
Initial Data Summary:

- **Total Variables:** 459
- **Numeric Variables:** 351
- **Categorical Variables:** 108
- **Total Observations:** 1000

Data Summary After Data Cleaning and Variable Selection:

- **Numeric Variables:** 176
- **Categorical Variables:** 48
- **Dummy Variables:** 85

R version 2.7.0 (2008-04-22) August 1, 2008
Platform: i386-pc-mingw32 Page: 1 of 1



1.2 Categorical Variable Summaries EDA Report for Data Frame `x.fact`

1 Independent Variable Information

1.1 Numeric Variable Summaries

	MIN	MED	MAX	AVG	STD	OBS	UNIQUE	MISSING
CS_AGE35_44	0	17.5	999	46.8	168	1000	380	0
CS_FAMILIES	0	76.2	999	101	159	1000	608	0
CS_HISPANICS	0	2.89	999	36.7	170	1000	404	0
CS_UNIT10P	0	2	999	38.7	170	1000	437	0
CS_VALS00	0	12.8	999	52.3	168	1000	586	0
CS_WHITE	0	91.1	999	112	157	1000	528	0
MI_AMB_TENURE	0	0	8	0.025	0.385	1000	5	0
MS_ENROLL_MONTH	1	7	12	6.61	3.13	1000	12	0
MS_IN_STAY6_PERC	0	0	0.667	0.000827	0.0215	1000	4	0
MS_MPT_PTS_EARNED	0	3.23e+03	3.67e+05	1.61e+04	3.56e+04	1000	733	0
SG_AIR_MEMO	0	0	1.1e+05	300	4.85e+03	1000	5	0
SG_CROWNE	0	0	40	0.447	2.05	1000	16	0
SG_CROWNE_RESORT_C	0	0	2	0.01	0.109	1000	3	0
SG_DISTO_PCT	0	0	1	0.0306	0.134	1000	45	0
SG_DOWNTOWN_C	0	0	12	0.06	0.582	1000	5	0
SG_EM_ES_COUNT	0	3	14	3.23	3.27	1000	16	0
SG_ENEWS_COUNT	0	4	22	6.11	6.76	1000	23	0
SG_EXHI_C	0	0	1	0.003	0.0547	1000	2	0
SG_EXHI_PCT	0	0	0.5	0.0009	0.0181	1000	3	0

1.2 Categorical Variable Summaries

	OBS	UNIQUE	MISSING	CAT1 %	CAT2 %	# CAT 80%
CS_DIVISION	1000	10	0	20.300	18.500	6
CS_REGION	1000	5	0	36.500	27.600	3
MI_CCD_TYP	1000	6	0	98.300	0.800	1
MI_CURRENT_IND	1000	2	0	81.400	18.600	1
MI_EMAIL_PREF_VAL_CD	1000	32	0	31.600	18.600	4
MI_SEC_PHONE_IND	1000	6	0	90.700	7.500	1
MS_COPRT_AMEX	1000	7	0	99.700	0.200	1
MS_COPRT_VISA	1000	51	0	91.600	1.900	1
MS_DOMINANT_BRAND	1000	23	0	46.800	35.700	2
MS_ENRL_BRAND_CD	1000	18	0	65.200	17.600	2
MS_ENRL_CHAIN_CD	1000	7	0	65.200	20.100	2
MS_ENRL_MCT_SUB_DESC	1000	6	0	65.200	34.400	2
MS_ENRL_SUB_MKT_RGN_~1	1000	16	0	65.200	34.400	2
MS_ENRL_SUB_MKT_RGN_~2	1000	16	0	65.200	34.400	2
MS_EOFFERS	1000	2	0	63.300	36.700	2
MS_ESOURCE	1000	5	0	48.800	34.800	2

R version 2.7.0 (2008-04-22) August 1, 2008
Platform: i386-pc-mingw32 Page: 1 of 1