Capstone Project II
Predicting individual annual cost of medical insurance by ML method

By
Lok Hang Ronald, Wong

Background Information

- Medical cost is one of the largest burdens of Americans
- Health care spending per person surpassed $10,000 in 2016
- March steadily higher to $14,944 in 2023.

Statistic source from Centers for Medicare and Medicaid Services (CMS)

Aims and Objectives

- Establish a ML model for predicting medical insurance cost
- Understand essential and non-essential factors for insurance cost
- Aid start-up companies to set up better marketing strategy
- Aid customers to have better planning on their medical insurance

Who might Care?

- Start-up Companies
  - Provide services such as consulting
  - Finding potential customers
- General Public
  - Evaluation of their current planes
  - Better planning and strategy

Photo source: http://canonprintermx410.blogspot.com

Dataset

- Data source from Kaggle - a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners

- Contains information of 1337 customers with the following features (1337 rows, 7 columns)

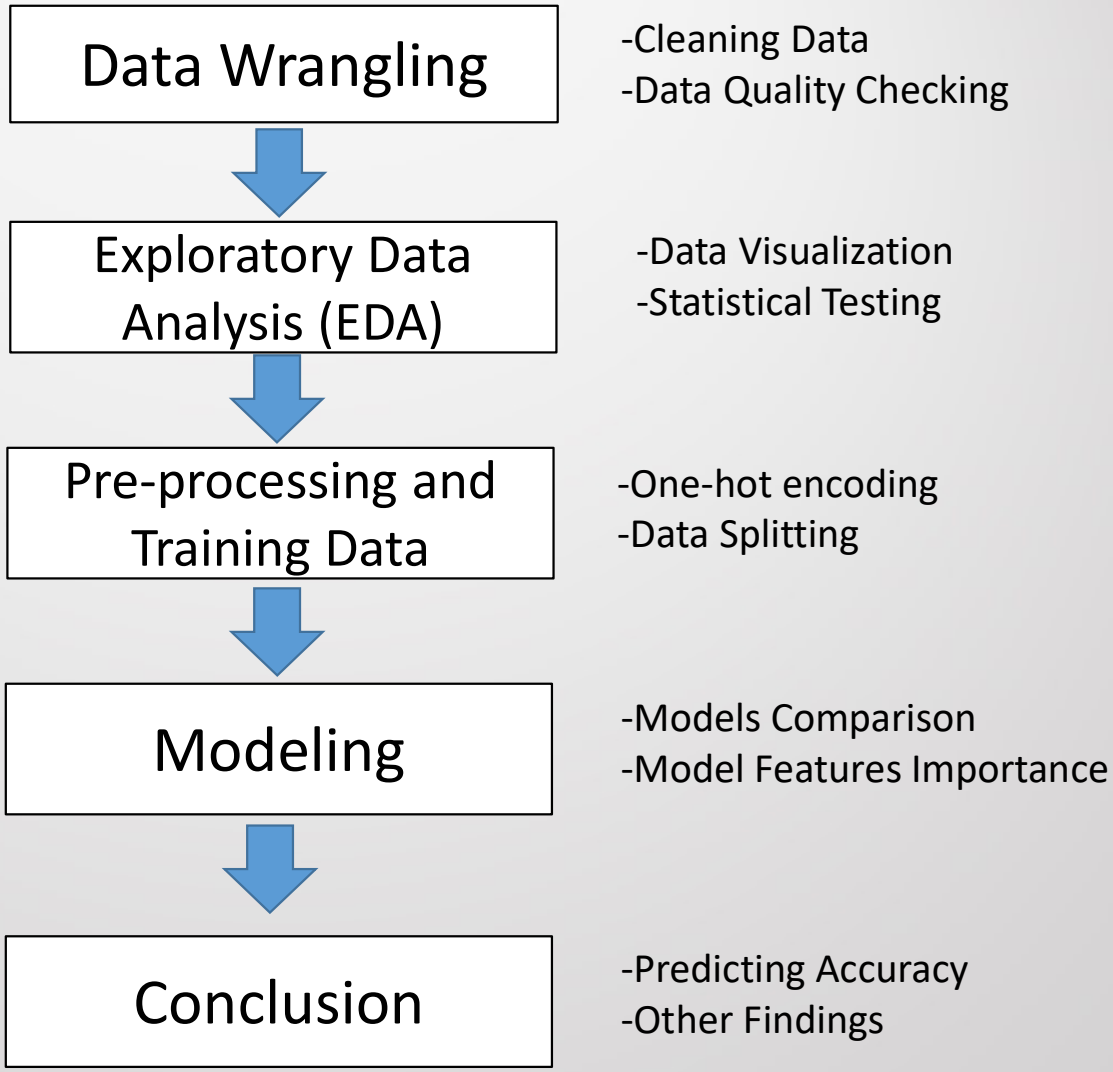
- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, (kg / m^2) ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, NE, SE, SW, NW
- charges: Individual medical costs billed by health insurance

Dataset

Data Preview:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Project Outline



Exploratory Data Analysis (EDA) – t-test

- For comparing categorical variable with numerical variable
- For 2 unique values in categorical variable
- Null Hypothesis: Mean of two groups are the same
- Adopted for gender and smoking behavior

Exploratory Data Analysis (EDA) – One-way ANOVA

- For comparing categorical variable with numerical variable
- For 3 or more unique values in categorical variable
- Null Hypothesis: Mean of all groups are the same
- Adopted for region and family size (0 to 5 children)

Exploratory Data Analysis (EDA) – chi-squared test

- For checking correlation between categorical variables
- Similar with heatmap for numerical variables
- Null Hypothesis: The categorical variables are not correlated
- Findings: Smoker and Gender, bmi and Gender are correlated

Modelling

- 2 ML approaches for predicting continuous variable

- Ordinary least squares (OLS)

  - A type of linear least squares method for estimating the unknown continuous variable in a linear regression model

- Random Forest Regressor (RFR)

  - An ensemble learning method for regression operate by constructing a multitude of decision trees

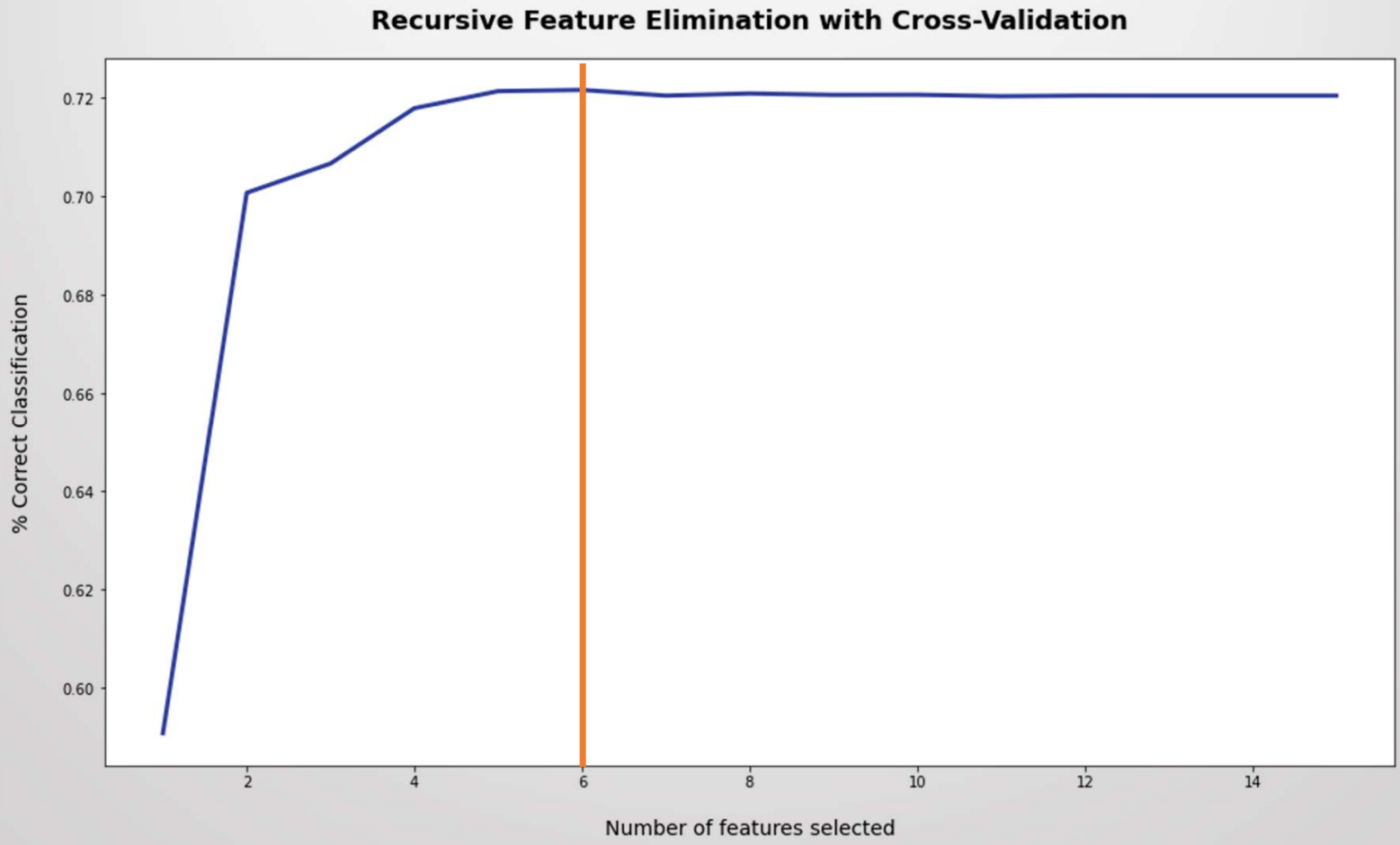
Modelling – Feature Selection

OLS Summary Report

	coef	std err	t	P>|t|	[0.025	0.975]
const	6462.7392	305.143	21.179	0.000	5863.884	7061.594
age	3686.3005	202.075	18.242	0.000	3289.720	4082.881
bmi	1772.7240	212.844	8.329	0.000	1355.008	2190.440
sex_1	70.4981	402.658	0.175	0.861	-719.735	860.731
smoker_1	2.268e+04	508.044	44.648	0.000	2.17e+04	2.37e+04
children_0_1	42.0528	416.949	0.101	0.920	-776.227	860.333
children_1_1	263.7529	474.098	0.556	0.578	-666.684	1194.190
children_2_1	1563.5188	523.927	2.984	0.003	535.290	2591.747
children_3_1	873.4268	600.356	1.455	0.146	-304.796	2051.650
children_4_1	2677.4248	1172.073	2.284	0.023	377.184	4977.666
children_5_1	1042.5631	1334.109	0.781	0.435	-1575.679	3660.805
SW_1	1426.1022	358.018	3.983	0.000	723.478	2128.726
SE_1	1242.4228	364.633	3.407	0.001	526.816	1958.030
NW_1	2113.1498	364.828	5.792	0.000	1397.160	2829.139
NE_1	1681.0644	364.648	4.610	0.000	965.427	2396.702

Modelling – Feature Selection

Recursive Feature Elimination and Cross-Validation Selection (RFECV)



Result – Metrics

- R2 Score

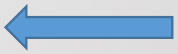
R2 is a statistic that will give some information about the goodness of fit of a model. (Best model will have R2 Score:1)

- Root Mean Square Error(RMSE)

A frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.

Result – OLS Model

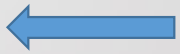
Model	R2 Score	RMSE
All Features	0.786	6033
Best 6 Features	0.643	7793
Drop 5 Features	**0.785**	**6042**



Best OLS Model

Result – RFR Model

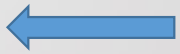
Model	R2 Score	RMSE
All Features	0.862	4840
Best 6 Features	**0.862**	**4844**



Best RFR Model

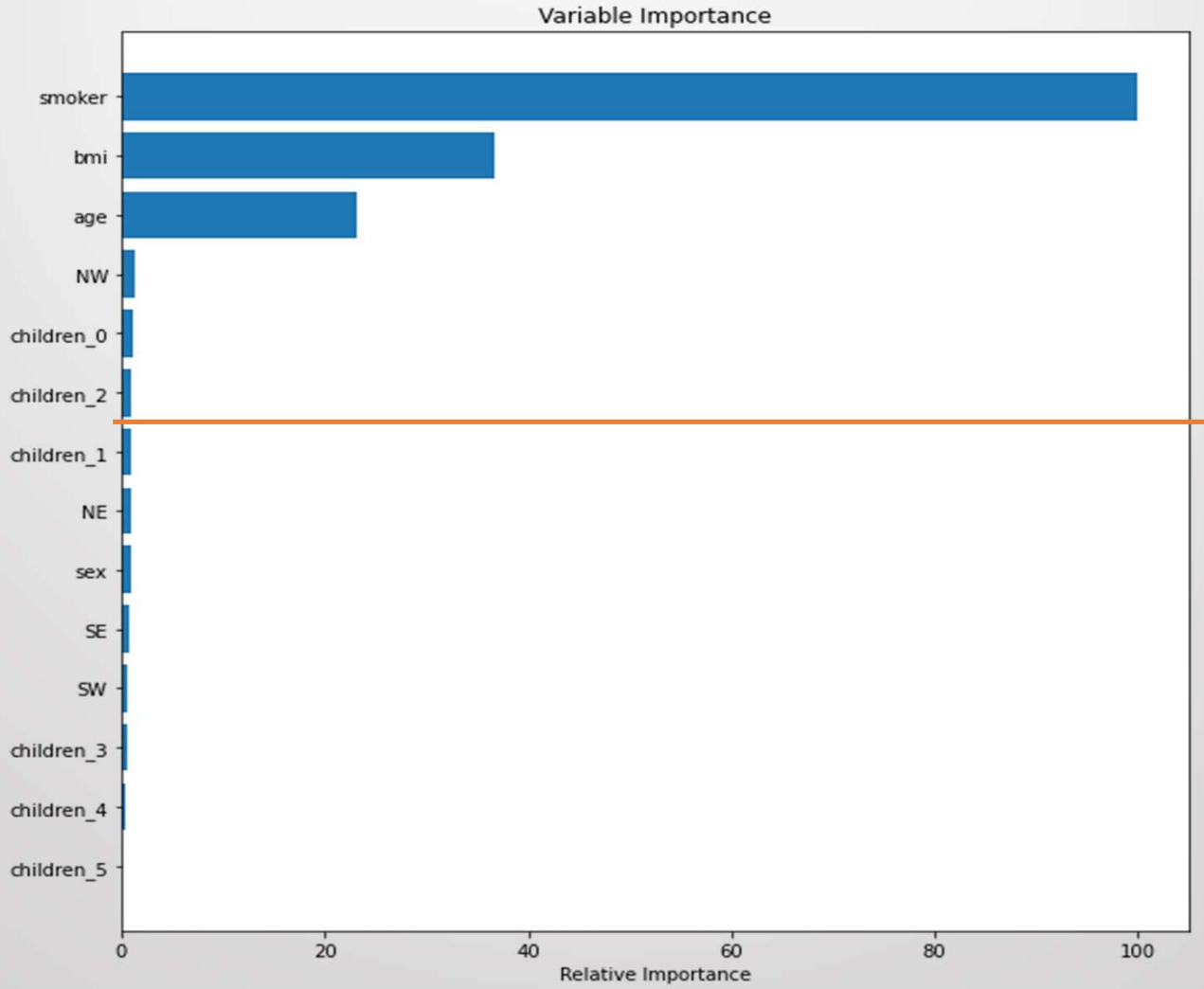
Result –Model Selection

Model	R2 Score	RMSE
Best OLS	0.785	6042
Best RFR	**0.862**	**4844**



Selected Model

Result –Feature Importance



Adopted in selected model

Not adopted in selected model

Data constraints and limitations

- Data size is small (1337 customers)
- Only 6 features have been incorporated for prediction
- Only 2 regression models have been adopted and compared

Future Studies

- Increasing data size to prevent over-fitting
- Comprise more customer information (More features can be incorporated such as medical history)
- Comparing more ML methods such as K-NN, Neural Network and non-linear model

Conclusion

- Established a model for predicting medical insurance cost
- Explored important and non-essential factors
- Importance to start-up companies and general public

Recommendation

- For Start-up companies:
  - Can provide consultation and prediction service to customers
  - Can set-up their long-term marketing strategy
  - Can explore potential customers

- For general public:
  - Can compare current medical insurance with predicted results
  - Can forecast future insurance cost and set-up better planning (such as changing smoking behavior and target BMI)
  - Can understand better the features affecting their insurance cost and the features can be ignored

Q & A