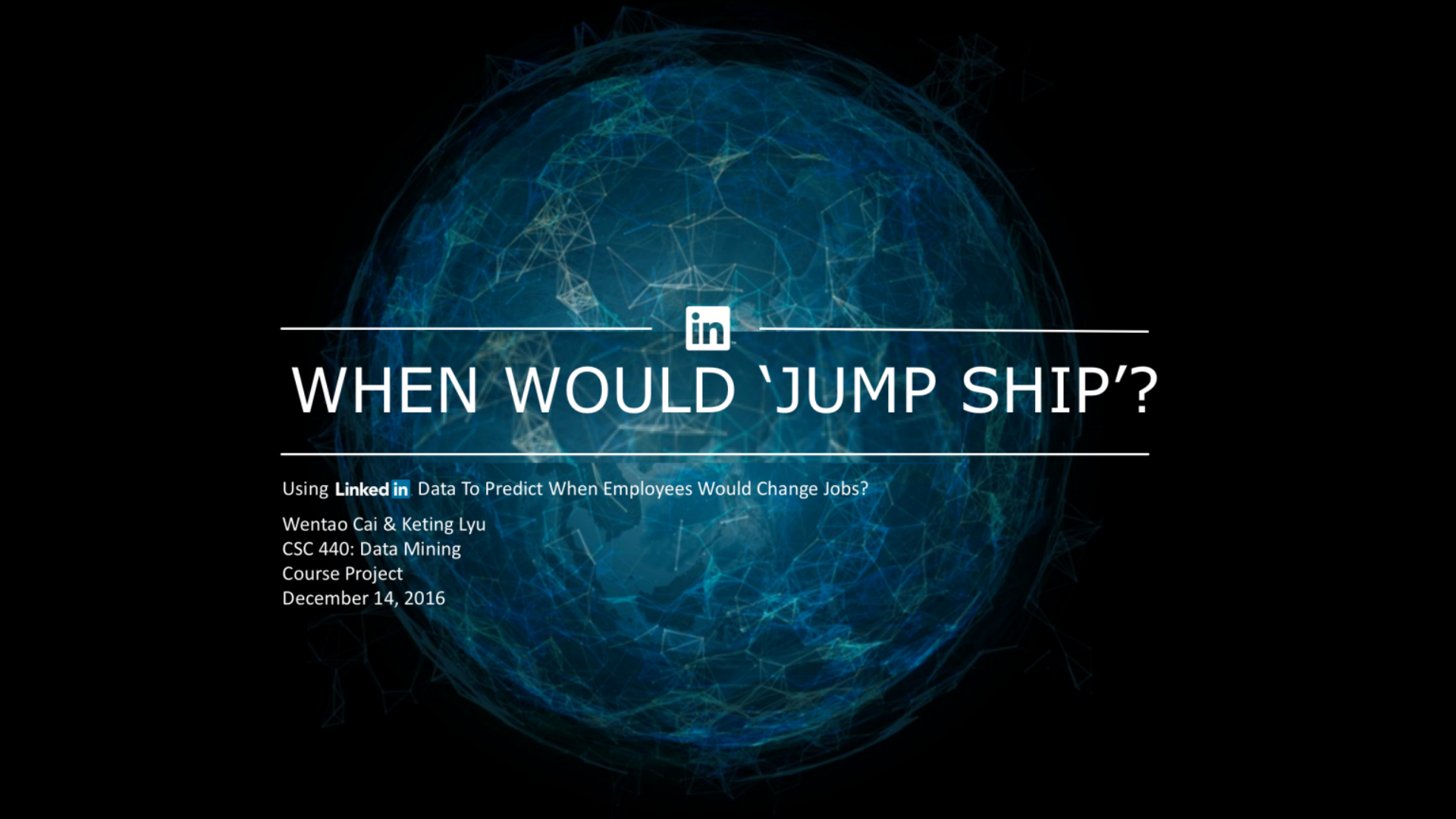


WHEN WOULD 'JUMP SHIP'?

Using LinkedIn Data To Predict When Employees Would Change Jobs?

Wentao Cai & Keting Lyu
CSC 440: Data Mining
Course Project
December 14, 2016

INDEX

Introduction

Background/Objective
/Literature Review

Dataset

Acquisition/Preprocessing
/Issue

Models

SVR/BPNN
/Random Forest

Analysis

Data Analysis/Model Analysis
/Future Work/Conclusion

INDEX

Introduction

Background/Objective
/Literature Review

Dataset

Acquisition/Preprocessing
/Issue

Models

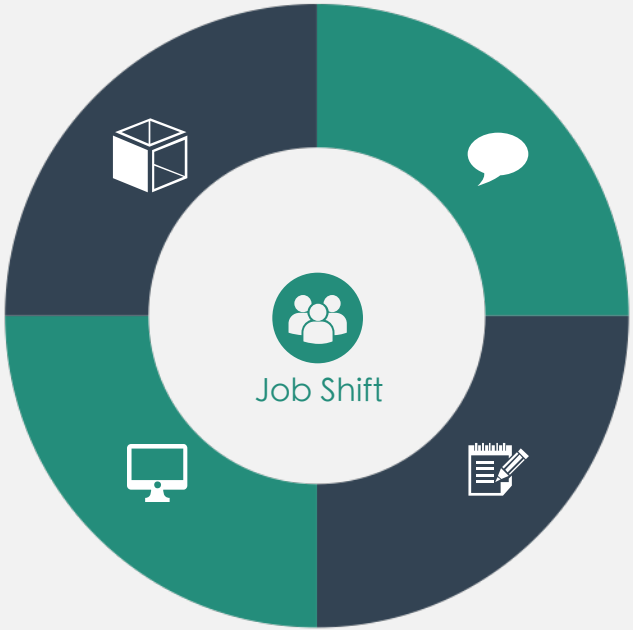
SVR/BPNN
/Random Forest

Analysis

Data Analysis/Model Analysis
/Future Work/Conclusion

INTRODUCTION
Background

Employee Turnover
Prediction on time point of job change is concerned with multiple agents in job market.



Employers
Training and adaption of employees are time and money consuming.

Employees
It’s difficult to decide when is an appropriate point to leave his/her position

Headhunters
Having the knowledge of career path of each works would release their job pressures.

INTRODUCTION

Objective

LinkedIn

An employment-oriented social networking website

Contains data of personal information, educational background and career paths

We get dataset from LinkedIn to model for the prediction when people with specific background would leave their current position.

INDEX

Introduction

Background/Objective
/Literature Review

Dataset

Acquisition/Preprocessing
/Issue

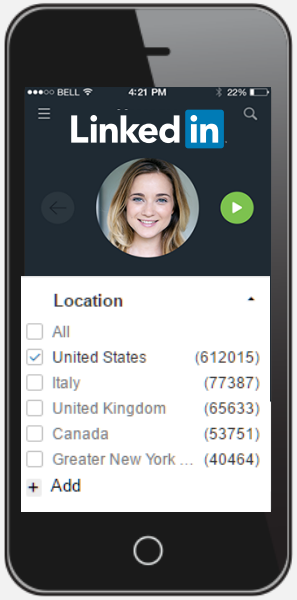
Models

SVR/BPNN
/Random Forest

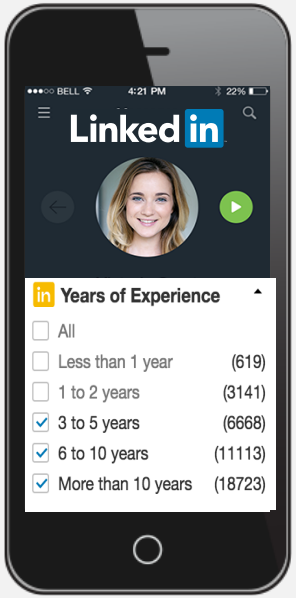
Analysis

Data Analysis/Model Analysis
/Future Work/Conclusion

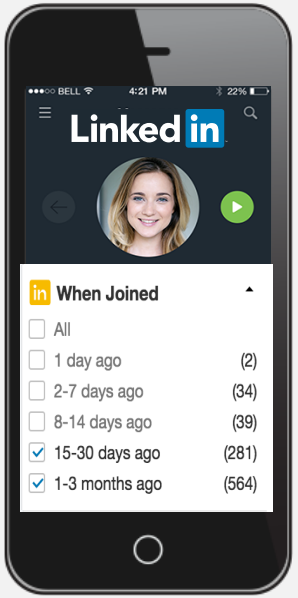
DATASET
Acquisition · Fliters



- Regional constraint



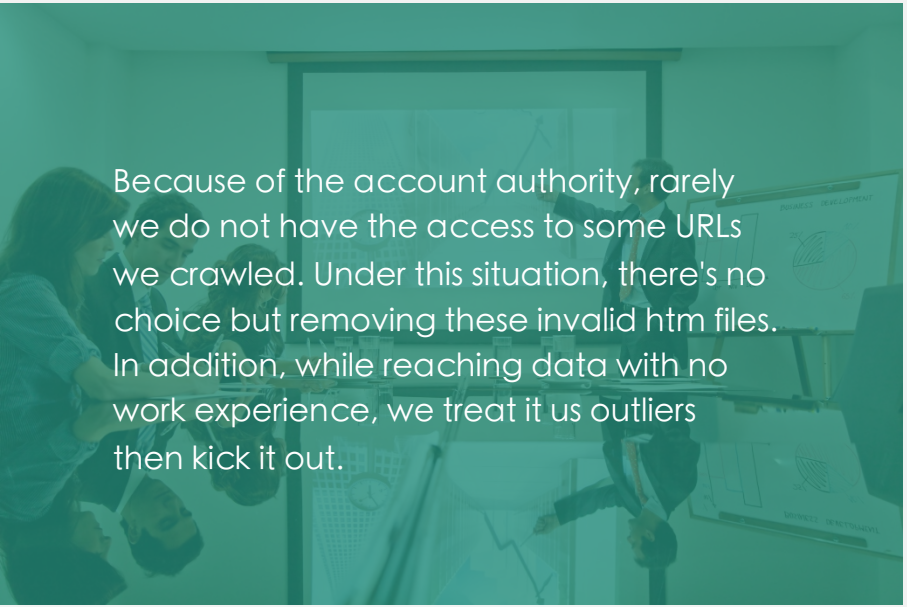
- Newbies are more likely to still stay in their first position.



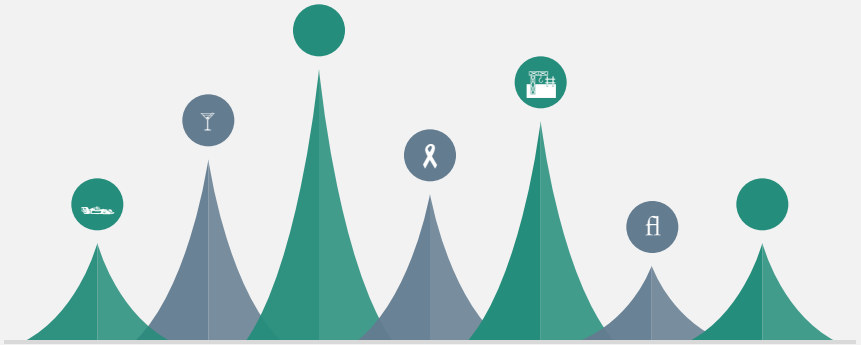
- Users tend to not fill profile completely in the first 15 days of registration

DATASET

Preprocessing · Data Cleaning

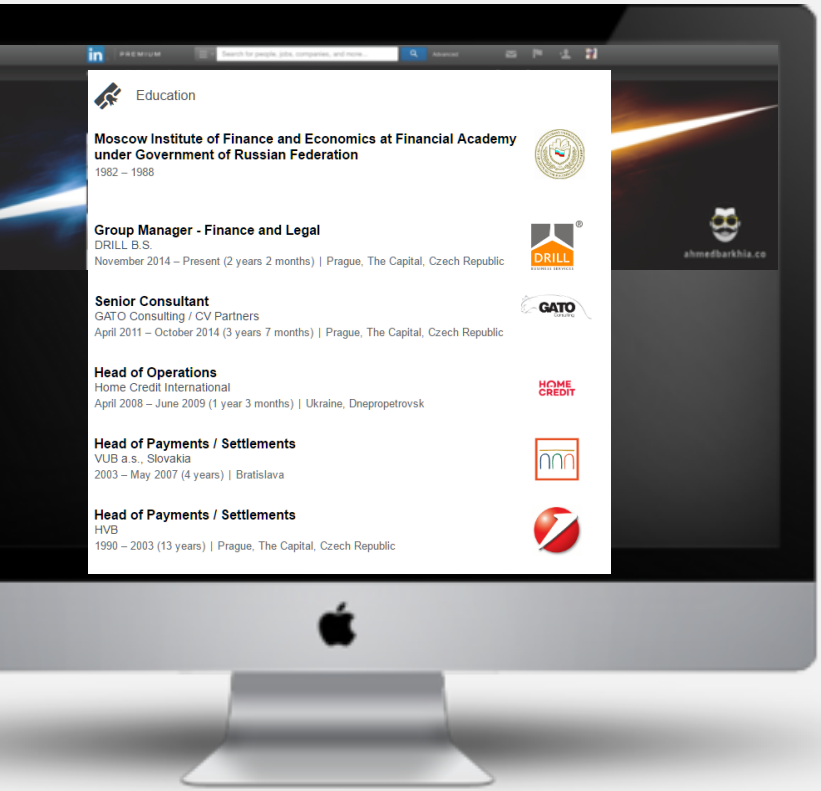


- There's no appearance of large-scale noisy data demonstrate LinkedIn users open the premium account and set their account as public tend to maintain their profile in a good status, which ensures the quality of the dataset.



DATASET

Preprocessing · Data Selection

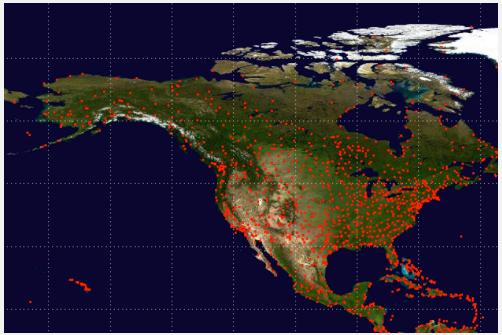


- For education, the first and the last are considered. The first education is for calculating age, namely, we abstract the current year (2016) by its start time in year. The last is to get people's highest degree and the last school they attended.
- For experience, we only take up to 6 which are most current per person. Only company name, title, start time and length staying in this position of each experience are stored.
- Other items we store are name, location, industry, and skills.

DATASET

Preprocessing · Transformation · Encode

- Encoded by the sequence number of the phrase



Location

City	Code
greater new york city area	0
san francisco bay area	1
new york new york	2
greater los angeles area	3
.....	...

Industry

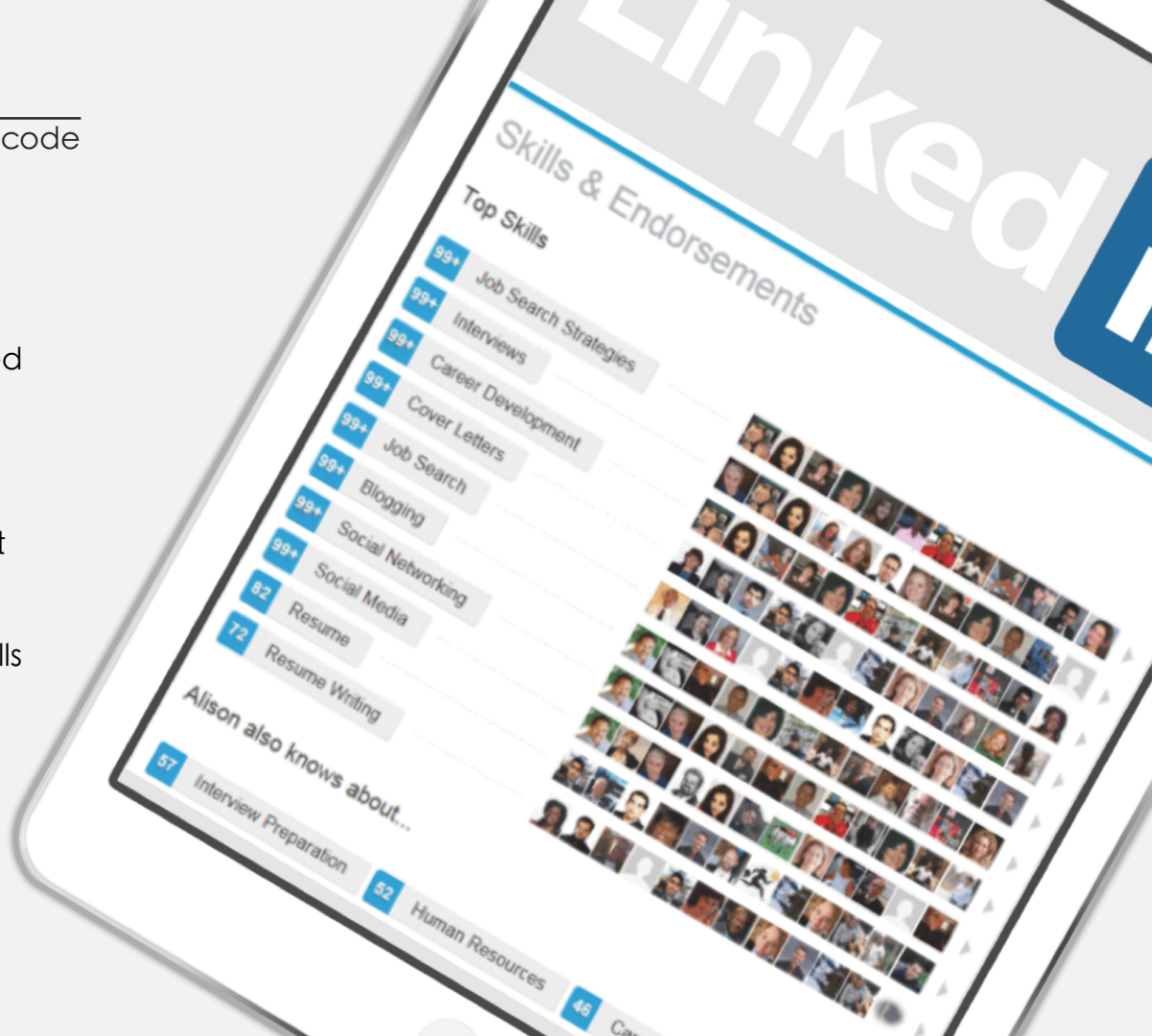
Industry	Code
marketing and advertising	0
information technology	1
financial services	2
computer software	3
.....	...

DATASET

Preprocessing · Transformation · Encode

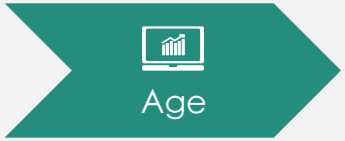
Skills

- First pick top 128 frequent skills sorted and numbered from 0 to 127. Then we classify these skills into 7 classes. Note that if a person has several of these 128 frequent skills, then we encode it into binary and convert it to integer.
- For example, say Sam only has 3 skills of management and 1 skills of internet, then his value of skill is 33, which is equal to binary 0100001.



DATASET

Preprocessing · Transformation · Encode



Difference between current date and the date he or she graduated from the first degree (in year). 0 stands for missing or questionable.

Degree	Code
bachelor degree	0
other	1
master degree	2
associate degree	3
doctor degree	4

Encoded by USNEWS rank of universities. 220 if not in the ranking.

DATASET

Preprocessing · Transformation · Encode

Degree	Code
Us army	0
Microsoft	1
LinkedIn	2
IBM	3
Bank of America	4
.....	...

Company: 1-6

C

Start Time: 1-6

O

- Difference between 1950-01-01 and begin date of the experience (in month).
- 0 stands for missing or questionable

Title	Code
Software Engineer	0
Reporter	1
President	2
Intern	3
Owner	4
.....	...

D

Title: 1-6

E

Length: 1-5

- Distribute between begin date and end date of the experience (in month).
- 0 stands for missing or questionable or unknown.

DATASET

Preprocessing · Transformation · Encode

Y: the target output, standing for the last known length of the experience (in month)

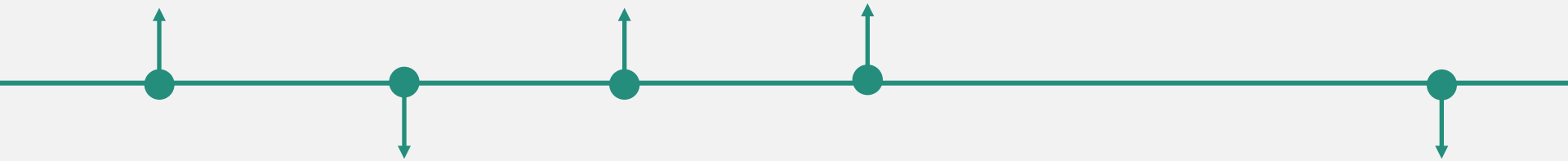
The final data attributes we obtained including Location, Industry, Skills, Age, Degree, School, Company (1-6), Title (1-6), Start Time (1-6), Length (1-5), and Y

DATASET
Issue

Regular LinkedIn account are only accessible to recruiter users

Dataset size limit

Example of Raw data					
Person A	Personal Info.	1st job	2nd job	3rd	3rd
Person B	Personal Info.	1st job	2nd job		



Many LinkedIn users choose to keep their data in privacy.

Data for training and testing					
	Input				Output
Person A	Personal Info.	1st job			Start time of 2nd job
Person A	Personal Info.	1st job	2nd job		Start time of 3rd job
Person A	Personal Info.	1st job	2nd job	3rd job	Start time of 4th job
Person B	Personal Info.	1st job			Start time of 2st job

INDEX

Introduction

Background/Objective
/Literature Review

Dataset

Acquisition/Preprocessing
/Issue

Models

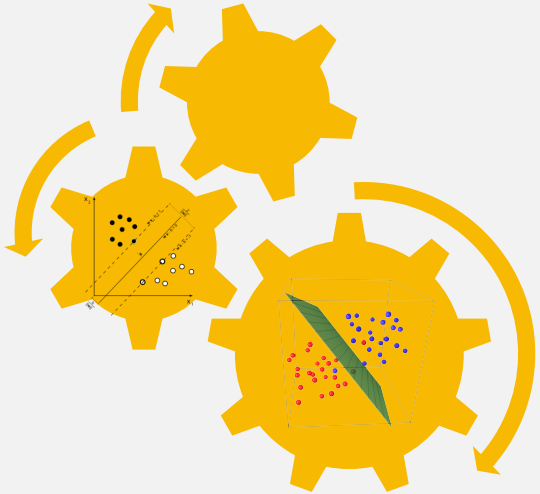
SVR/BPNN
/Random Forest

Analysis

Data Analysis/Model Analysis
/Future Work/Conclusion

MODELS

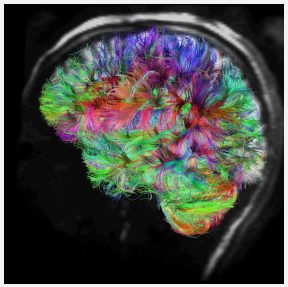
Support Vector Regression

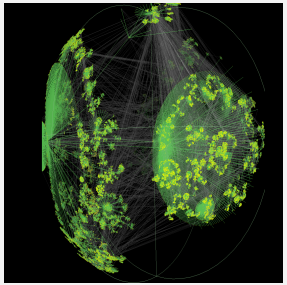


- Support Vector Regression performs regression on continuous variables rather than only the discrete, like Support Vector Machine does.
- SVM aims at finding the maximum separating hyperplane so that points of different classes have the largest margin among each other.
- To train SVR model for our data, we choose Radial Basis Function (RBF) as kernel, setting

MODELS

Back-Propagation Neural Network

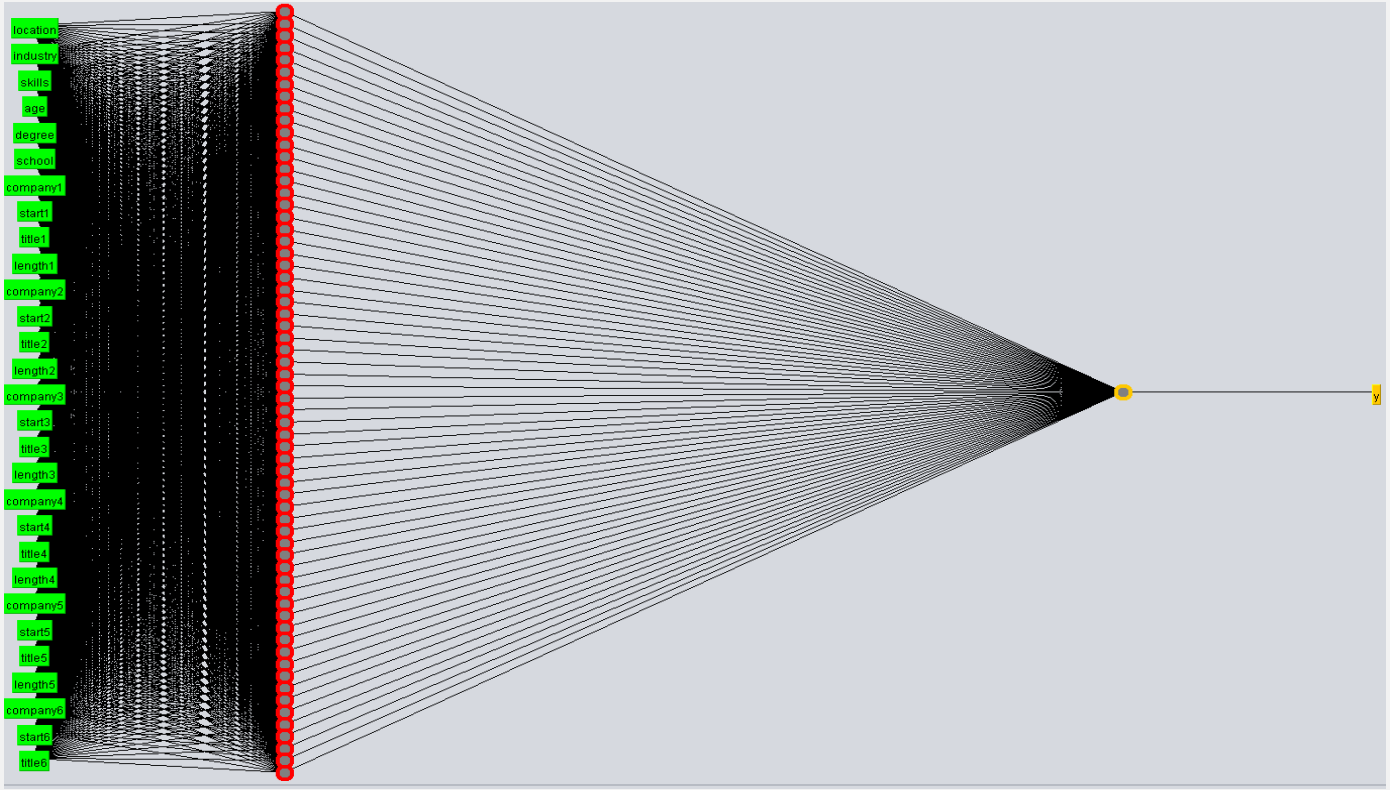




- Back-Propagation Neural Network is one kind of Artificial Neural Network trained by Back-Propagation algorithm with optimization method such as stochastic gradient descent.
- A net is constructed by several layers, each of which contains a number of units.
- In general, the net with three layers, which are input, hidden, and output layer, is used most frequently.
- In our project, we use 3-layer net, where the hidden layer has 64 units using sigmoid as kernel.

MODELS

Back-Propagation Neural Network



Visualized Neural Network model with 64 units in hidden layer

MODELS
Random Forest

Random Forest is an ensemble learning method for classification and regression.

The idea behind this method is constructing a bunch of decision trees by bagging and taking the mean of their result as the output.

INDEX

Introduction

Background/Objective
/Literature Review

Dataset

Acquisition/Preprocessing
/Issue

Models

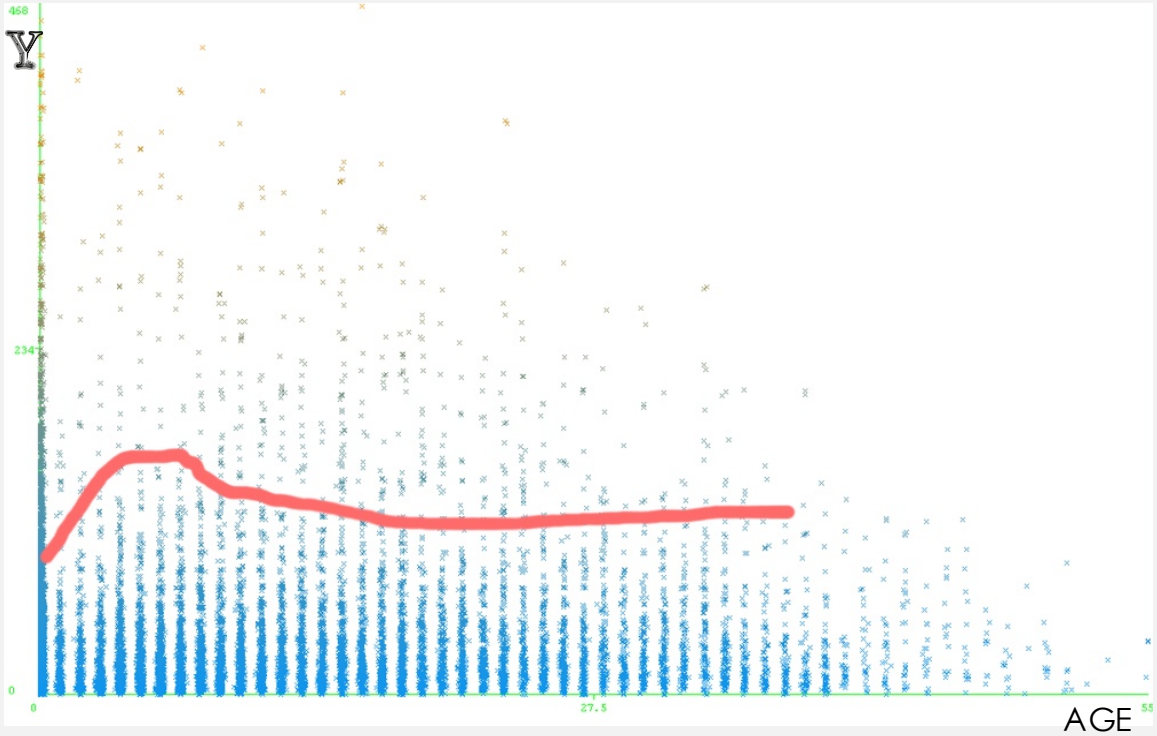
SVR/BPNN
/Random Forest

Analysis

Data Analysis/Model Analysis
/Future Work/Conclusion

Analysis

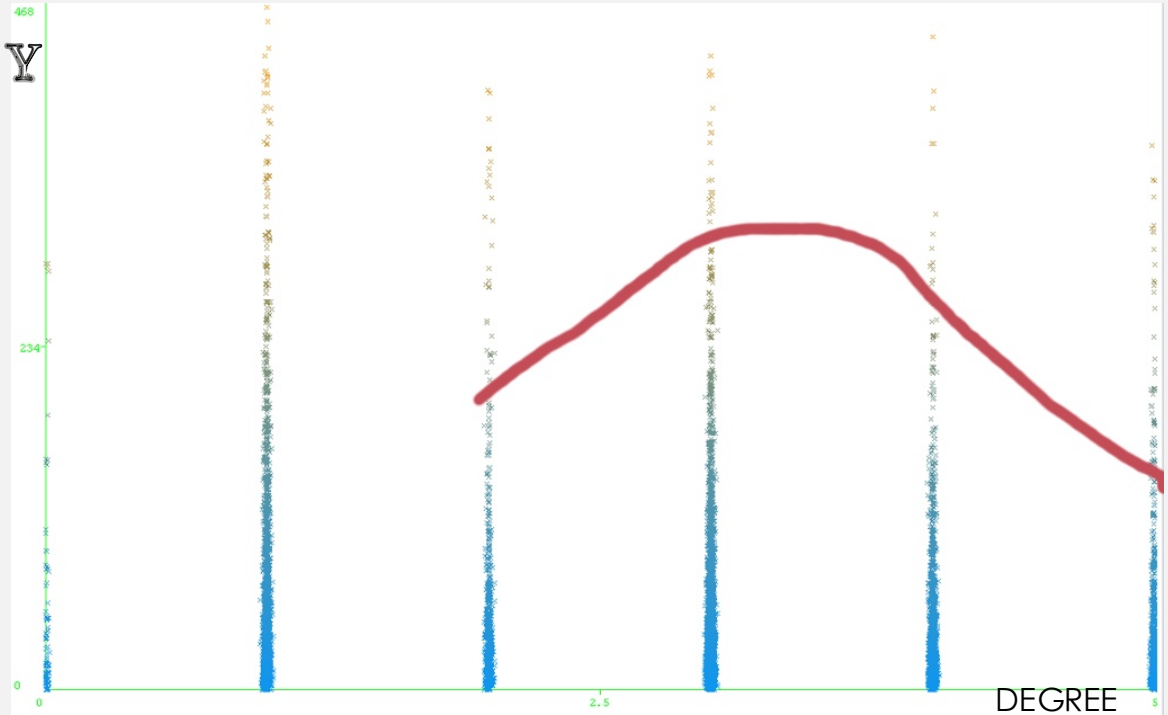
Data Analysis · Age



- New graduates are more likely to leave the current position
- While with 10 years after graduating tend to be slightly more stable in the position.
- After 10 years Y goes unrelatedly with ages.

Analysis

Data Analysis · Degree

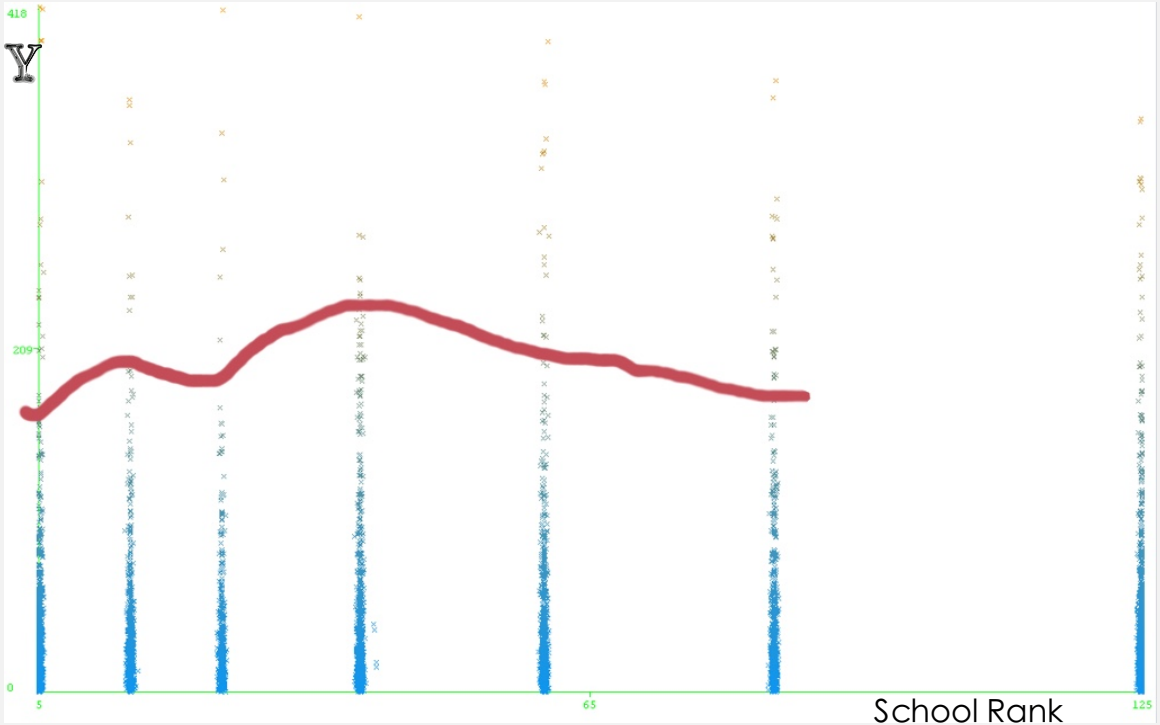


- People with Bachelor as highest are more stable and have bigger target value.
- People with Doctor Degree, on the contrary, somehow tend to stay in the position for a shorter time.

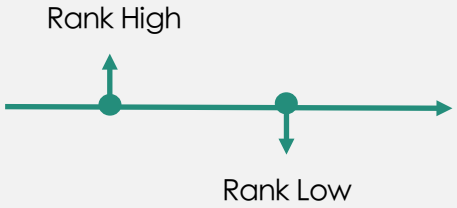
degree	code	degree	code
unknown	0	others	1
Associate	2	Bachelor	3
Master	4	Doctors	5

Analysis

Data Analysis · School

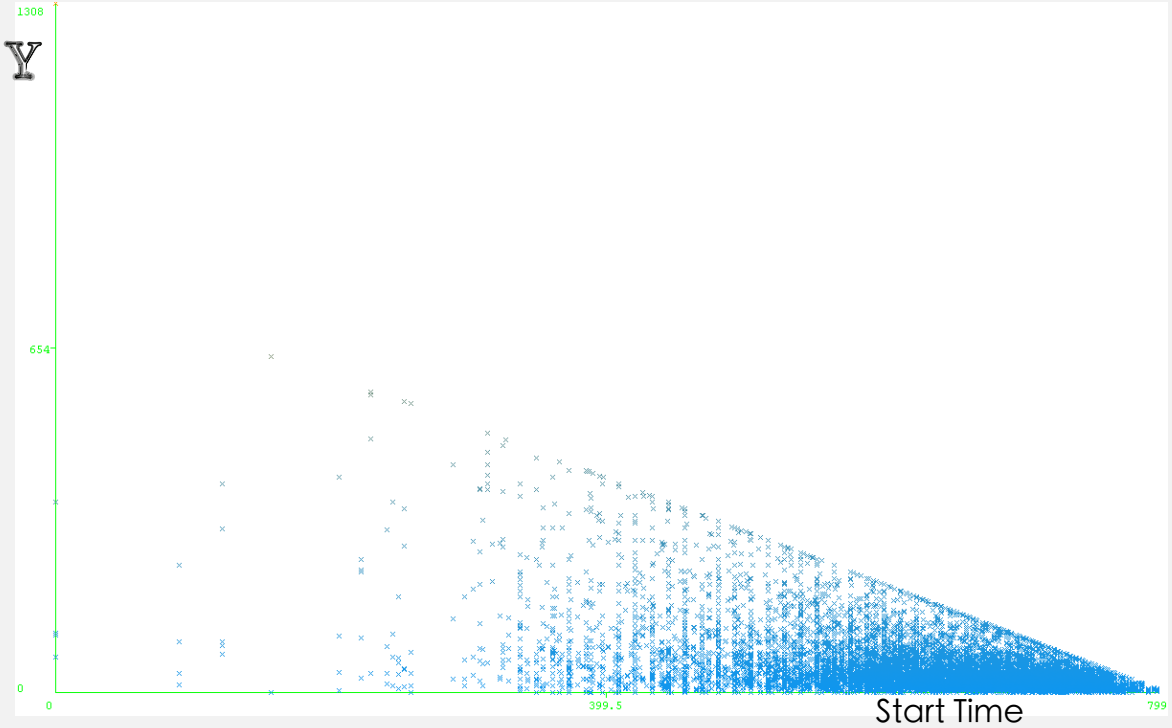


- Graduate from universities ranked from 30 to 70 tend to stay in their current job for a bit long time.



Analysis

Data Analysis · Start Time

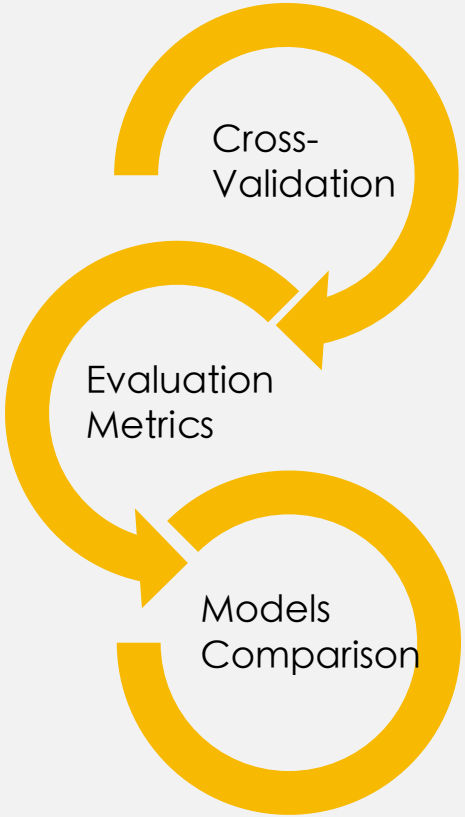


- Points upper line should be treated as outliers.
- Since they cover only tiny part of whole data, we can just throw it if such outlier occurs in some pieces of data.

Analysis
Model Analysis

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

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

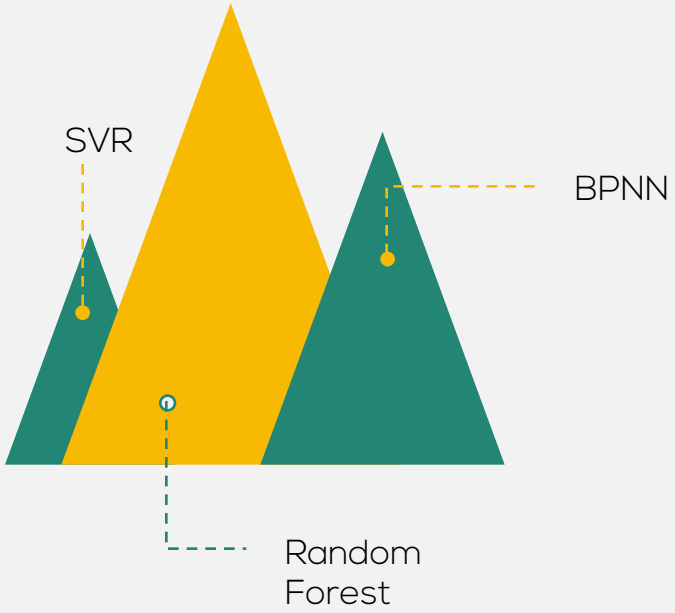


- We use 10-fold cross-validation to evaluate models

Model	Time (sec)	MAE	RAE
SVR	148.69	28.7513	88.4751%
BPNN	4905.14	19.0443	58.6024
RF	18.7	18.5468	57.0731%

Analysis

Model Analysis · Model Comparison



Random Forest

Costs significantly short time to build and gets a surprising result, where MAE is just about 18.

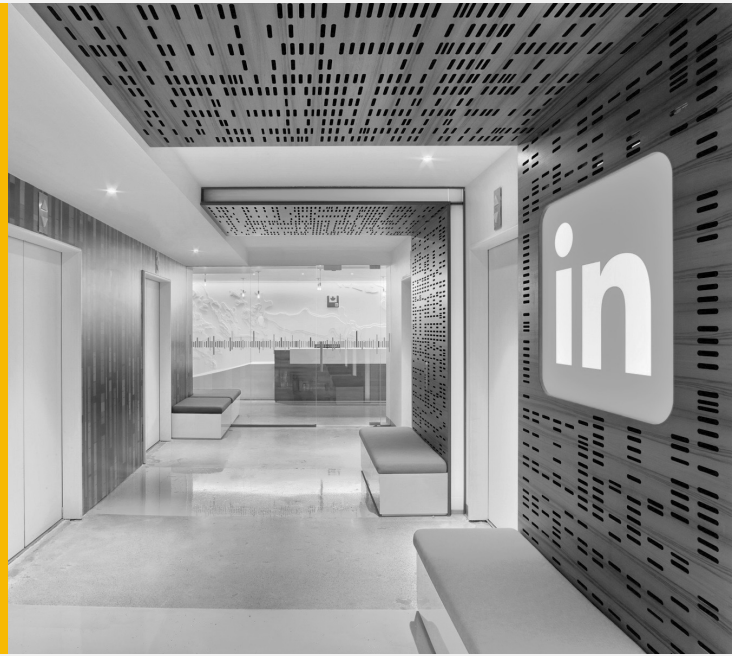
It means we can predict time a person staying in its current position with the average error of 18 months, i.e. one year and a half.

Future Work

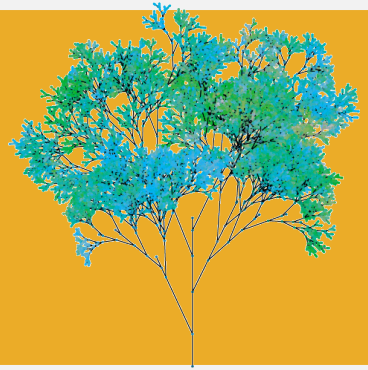
For dataset, we could extract more featured information as dimension, such like times of job shift or amount of connection.

For models, we can adjust structure and optimize parameters for neural networks to build model with better accuracy.

Seeking for additional data sources is also helpful. We can try to find any substitutes of LinkedIn, or get other method to access more in LinkedIn.



Conclusion



SVR

Unexpectedly it has bad effect and is hard to evolve anymore

BPNN

Requires awfully long time and machine resource to train and evolves slow.

Random Forest

It takes kind of random process in training, gets the best effect within an unexpected short time.

REFERENCES

1. Amir Mohammad Esmaieeli Sikaroudi, RouzbehGhousi, Ali EsmaieeliSikaroudi. *A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing)* [J]. Journal of Industrial and Systems Engineering, 2015(8.4): 106-121
2. Mai Ngoc Khuong, Bui Diem Tien. *Factors influencing employee loyalty directly and indirectly through job satisfaction - A study of banking sector in Ho Chi Minh City* [J]. International Journal of Current Research and Academic Review, 2013(14):
3. Stefan Lessmann, Stefan Voß. *A reference model for customer-centric data mining with support vector machines* [J]. Computational Intelligence and Information Management, 2009(12): 520-530
4. Wenseng Dai, Yang-Yu Chuang, Chi-Jie Lu. *A Clustering-based Sales Forecasting Scheme Using Support Vector Regression for Computer Server* [J]. Procedia Manufacturing, 2015(2): 82-86
5. N.B. Chaphalkar, K.C. Iyer, Smita K. Patil. *Prediction of outcome of construction dispute claims using multilayer perceptron neural network model* [J]. International Journal of Project Management, 2015(33.8): 1827–1835
6. Richard K. Zimmerman, G. K. Balasubramani, Mary Patricia Nowalk. Classification and Regression Tree (CART) analysis to predict influenza in primary care patients [J]. BMC Infectious Diseases, 2016(16): 503

THANK YOU!

Have a Nice Day!