IDEA

Rensselaer Institute for Data Exploration and Applications

Project Outmatch

Authors/Group Members:

Brian Maher
Sonny Chauhan
Ruoshi Zhao

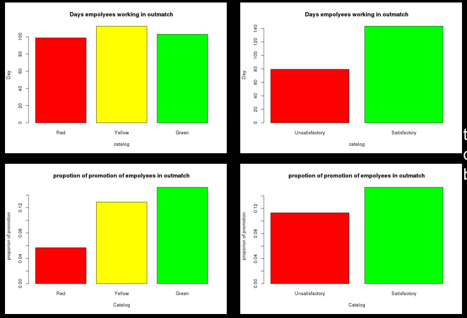
Overview & Problems Tackled: -

We are working with a company called Outmatch, who gave us a dataset of current and prior employees for a large discounter retailer. Our goal is to help them using data analytics to improve their hiring process by hiring employees who will:

- Perform well in their roles
- Have the potential for promotion to managerial roles
- Stay with the company for longer

Data Description:

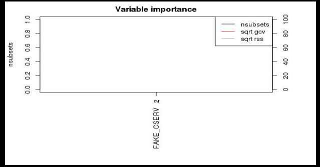
- The dataset has 130792 rows corresponding to current or prior employees in the company
- The 177 features represent
  - Results of a personality questionnaire administered during the application process
  - Results of personality questionnaire
- Added in a new feature Satisfactory/ Unsatisfactory:
  - Employee there less than 90 days Unsatisfactory iff terminated and ineligible for rehire
  - Satisfactory iff terminated and eligible for rehire OR currently employed for >= 90 days
  - Otherwise, unknown



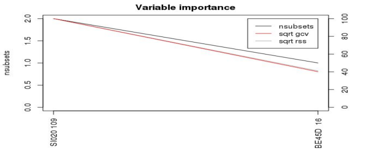
these plots show the difference between band and satisfactory

Data Analytics Methods:

- Experimental design
  - Eligible for rehire
  - Promotions
  - 90 day turnover
  - Satisfactory/Unsatisfactory
- Feature selection methods
  - Earth
  - Information value and weight of evidence
  - Correlation
  - RandomForest
- Machine learning
  - Logistic regression
  - Knn
  - Kknn
  - RandomForest
  - Decision trees



The above plot is an example of the earth model with respect to aggregates for Satisfactory/ Unsatisfactory.



This is an example of the earth model with eligible for rehire for explanatory.

Experimental Results 1:

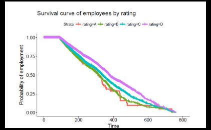
- Applied modeling above with the experiments used above
- for all of tasks models performed poorly based on accuracy
- models highly overfitting the data
- patterns in data difficult to learn

References/Citations:

- Presentation 2
- R packages used:

1. rpart 2. class 3. randomForest 4.ggplot

Rating	% with rating who are satisfactory	% with rating who are unsatisfactory	% of sample with rating
A	68.9%	31.1%	1.1%
B	64.1%	35.9%	9.9%
C	53.7%	46.3%	36.1%
D	44.4%	55.6%	52.9%

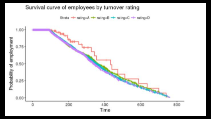


- **Left, above, is a chart of the satisfactory ratings class on a held out validation set. Right, below is survival curves for the satisfactory ratings on a held out validation set. They show that more likely to be satisfactory employees stay at the company less time.**

Experimental Results 2:

- **We extract insight from our low accuracy models with a thresholding method: we divide samples into multiple groups based on their predicted probability of belonging to the desirable class (satisfactory or >= 90 days)**
- **We use the follow thresholds to assign ratings**
  - **For probability P of being satisafactory:**
    - **A: P >= 0.7**
    - **B: 0.6 <= P < 0.7**
    - **C: 0.5 <= P < 0.6**
    - **D: P < 0.5**
  - **For probability P of staying >= 90 days:**
    - **A: P >= 0.6**
    - **B: 0.5 <= P < 0.6**
    - **C: 0.4 <= P < 0.5**
    - **D: 0.4 > P**
  - **Left, below, is a chart of the turnover ratings class on a held out validation set. Right, below is survival curves for the turnover ratings on a held out validation set**

Rating	% >= 90 days	% < 90 days	% with Rating	Average days employed
A	58.8%	41.2%	0.5%	158.0
B	52.7%	47.3%	11.7%	135.0
C	44.0%	56.0%	63.6%	113.9
D	37.7%	62.3%	24.2%	95.4



Discussion:

- Models don't perform well when evaluated by accuracy, but business insights can still be derived by looking at more extreme samples
- many features derived showed great aspects of being a good employee and we can see what features are most important

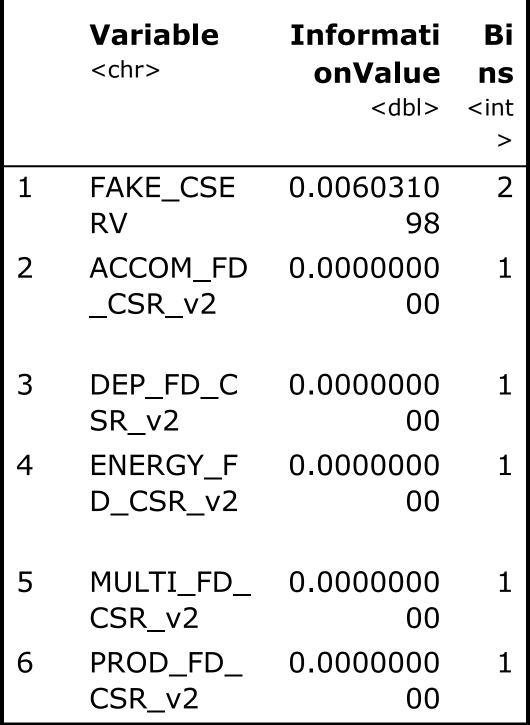
Conclusion & Recommendations:

- It is difficult to extract meaningful insights on data
- more varisation responses may be able to better model patterns
- should include people they didn't hire in dataset as well
- To encourage the company 4 or 6 options on all the questionnaire items instead of 2
- Deep Learning
- Try different defintions of Satisfactory/Unsatisfactory
- See if any of the features are highly correlated with each other besides response variables

Project Outmatch

Feature Selection Methods	Independent Variables	Feature tested	Important variables
Earth	Aggregates	Satisfactory/ Unsatisfactory	1. FAKE CSERV
Information Value and weight of evidence	Aggregates	Satisfactory/ Unsatisfactory	1.FAKE CSERV
Correlation	Aggregates	Satisfactory/ Unsatisfactory	most positively correlated(Satisfactory): 1.PROD_FD_CSR_v2 2.SELF_CONT_FD_CSR_v2 3.SOC_FD_CSR_v2
Correlation	Aggregates	Satisfactory/ Unsatisfactory	most negatively correlated (Unsatisfactory): 1.FAKE_CSERV 2.ENERGY_FD_CSR_v2 3.MULTI_FD_CSR_v2

Feature Selection Methods	Independent Variables	Feature tested	Important variables
Earth	Explanatory	Satisfactory/ Unsatisfactory	1. SI020 2. SI016 3. GG1029 4. BHL20
Information Value and weight of evidence	Explanatory	Satisfactory/ Unsatisfactory	1.SI020 2.SI018 3.SI016 4.SI059 5.SI017
Correlation	Explanatory	Satisfactory/ Unsatisfactory	most positively correlated(Satisfactory): 1. SI020 2. SI018 3. SI059 4. SI016 5.SI017 -all personality questionnaires 1-Agree, and 2-Disagree
Correlation	Explanatory	Satisfactory/ Unsatisfactory	most negatively correlated (Unsatisfactory): 1.FAKE_CSERV 2.EE2198 3.MT2102 4.EE2218 5.ENERGY_FD_CSR_v2



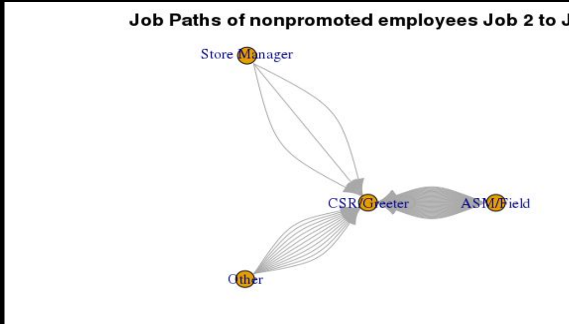
	Variable <chr>	InformationValue <dbl>	Bins <int>
1	FAKE_CSERV	0.006031098	2
2	ACCOM_FD_CSR_v2	0.000000000	1
3	DEP_FD_CSR_v2	0.000000000	1
4	ENERGY_FD_CSR_v2	0.000000000	1
5	MULTI_FD_CSR_v2	0.000000000	1
6	PROD_FD_CSR_v2	0.000000000	1

This is an example of a table for information value and weight of evidence.

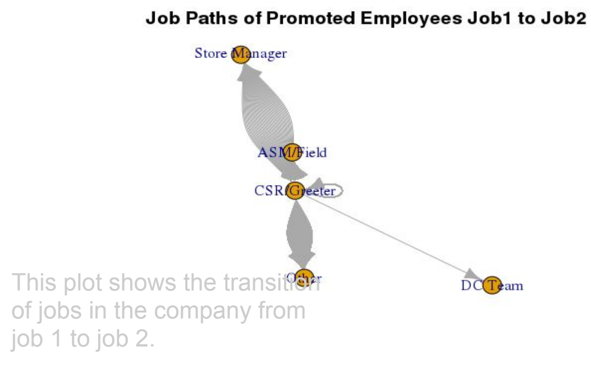
Project Outmatch

Feature Selection Methods	Independent Variables	Feature tested	Important variables
Earth	Aggregates	Promotions	1.MULTI_FD_CSR_v2 2.FAKE_CSERV
Information Value and weight of evidence	Aggregates	Promotions	1.MULTI_FD_CSR_v2
Correlation	Aggregates	Promotions	most positively correlated(people likely to get Promoted): 1.MULTI_FD_CSR_V2 2.P_FD_CSR_v2 3.SOC_FD_CSR_v2
Correlation	Aggregates	Promotions	most negatively correlated (not Promotable): 1.FAKE_CSERV 2.SELF_CONT_FD_CSR_v2

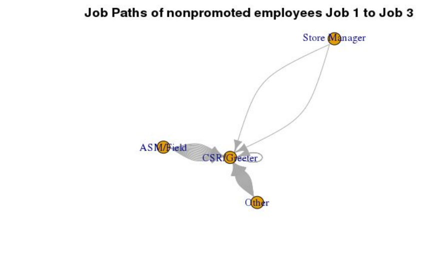
Feature Selection Methods	Independent Variables	Feature tested	Important variables
Earth	Explanatory	Promotions	1. PP2215 2. SI052 3.FAKE012 4.MT2108
Information Value and weight of evidence	Explanatory	Promotions	1.PP2215 2.MT2108 3.LC27D 4.SI052 5.BE45D
Correlation	Explanatory	Promotions	most positively correlated(Promotable): 1.PP2215 2.MT2108 3.LC27D 4.BE45D 5.MULTI_FD_CSR_v2
Correlation	Explanatory	Promotions	most negatively correlated (not Promotable): 1.SI052 2.SI051 3.SI525 4.BHL20 5.FAKE_CSERV



This plot shows the transition in jobs from job 2 to job 3 of people who are not employed in the company.



This plot shows the transition of jobs in the company from job 1 to job 2.



This is an example of an arcplot showing which transitions are not promotions in the company.

Feature Selection Methods	Independent Variables	Feature tested	Important variables
Earth	Aggregates	ELIGIBLE FOR REHRE	1FAKE_CSERV
Correlation	Aggregates	ELIGIBLE FOR REHIRE	Some of the most positively correlated (people likely to be ELIGIBLE FOR REHIRE): 1PROD_FD_CSR_v2 2.SOC_FD_CSR_v2 3.SELF_CONT_FD_CSR_v2
Correlation	Aggregates	ELIGIBLE FOR REHIRE	Some of the most negatively correlated (mot eligible for rehire)): 1.FAKE_CSERV 2.ENERGY_FD_CSR_v2 3.MULTI_FD_CSR_v2

Feature Selection Methods	Independent Variables	Feature tested	Important variables
Earth	Explanatory	ELIGIBLE FOR REHIRE	1. SI020 2. BE45D
Correlation	Explanatory	ELIGIBLE FOR REHIRE	Some of the most positively correlated(employees eligible for rehire): 1.SI020 2.SI018 3.SI059 4. SI060 5. SI017
Correlation	Explanatory	ELIGIBLEFORREHIRE	Some of the most negatively correlated (not eligible for rehire): 1.FAKE_CSERV 2.TEN05 3.MT2102 4.MULTI_FD_CSR_v2 5.ENERGY_FD_CSR_v2

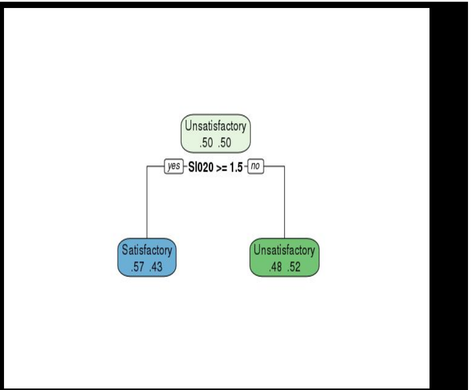
Feature Selection Methods	Independent Variables	Feature tested	Important variables
Earth	Aggregates	ninety day turnover	1.P_FD_CSR_v2
Information Value and weight of evidence	Aggregates	ninety day turnover	1.P_FD_CSR_v2
Correlation	Aggregates	ninety day turnover	most positively correlated(people likely to stay longer than 90 days): all aggregates have negative correlation with staying longer than 90 days.
Correlation	Aggregates	ninety day turnover	most negatively correlated (ninety day turnover): 1.P_FD_CSR_v2 2.MULTI_FD_CSR_v2 3.ENERGY_FD_CSR_v2

Feature Selection Methods	Independent Variables	Feature tested	Important variables
Earth	Explanatory	ninety day turnover	1.SI016 2.MT2102
Information Value and weight of evidence	Explanatory	ninety day turnover	1.SI016 2.SI020 3.MT2102 4.SI060 5.SI525
Correlation	Explanatory	ninety day turnover	Some of the most positively correlated(ninety day turnover): 1. SI016 2.SI020 3.SI525 4.EE1190 5. SI060
Correlation	Explanatory	ninety day turnover	Some of the most negatively correlated (not staying 90 days): 1.MT2102 2.P_FD_CSR_v2 3.MULTI_FD_CSR_v2 4.GG2049 5. ENERFY_FD_CSR_v2

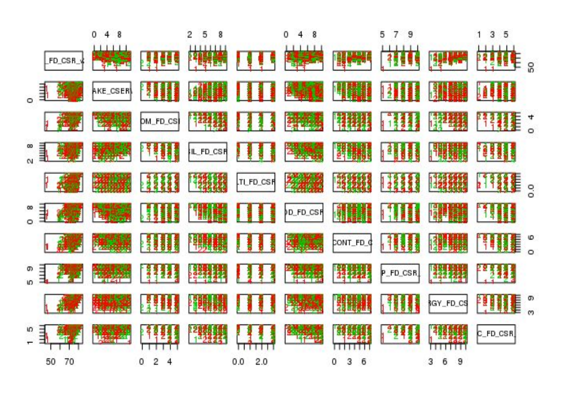
IDEA Rensselaer Institute for Data Exploration and Applications

Project Outmatch

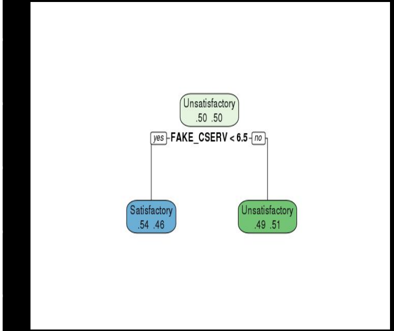
Model Type	Independent Variables	Sensitivity /Recall train/test (true positive rate)	Acceptance Rate (pred P/total)	Overall Accuracy train/test
Assume "Pass" is satisfactory	PassFail	99.50%	99.54%	48.42%
Baseline model.	NA	48.44% / 48.44%	100% / 100%	48.44% / 48.44%
Logistic regression (90 day)	Aggregates	31.95% 31.89%	29.28% 29.29%	53.23% 53.16%
Logistic regression (90 day)	All explanatory	44.29% 44.18%	37.93% 38.01%	56.54% 56.35%
Logistic regression (124 day)	Aggregates	31.95% 31.89%	29.28% 29.29%	53.23% 53.16%
Logistic regression (124 day)	All explanatory	44.29% 44.18%	37.93% 38.01%	56.54% 56.35%
Random Forest	Aggregates	11.25% 11.58%	51.55% 51.31%	48.63% 48.32%
Random Forest	All explanatory	11.68% 12.00%	51.19% 51.19%	50.04% 50.07%
Decision tree	Aggregates	50.02%, 49.79%	52.2%, 52.02%	52.13%, 51.95%
Decision tree	All Explanatory	60.67%, 58.81%	53.98%, 52.88%	54.00 %, 52.93%
knn	Aggregates	74.86%, 51.63%	75.02% , 50.51%	74.98%, 50.54%
knn	Explanatory	75.13%, 50.87%	50.76% , 50.11%	74.63 % ,50.76%
kknn (Weighted knn)	Aggregates	70.51% , 50.88%	49.98% , 49.98%	70.18%, 50.89%
kknn (Weighted knn)	Explanatory	99.18%, 51.88%	50.2%, 50.11%	99.25%, 51.81%



This is an example of the decision tree output in R with Satisfactory/Unsatisfactory with respect to explanatory variables.If an employee has SI020 greater than or equal to 1.5 than there is a 57% they are Satisfactory and if SI020 is less than 1.5 then there is a 48% they are Unsatisfactory

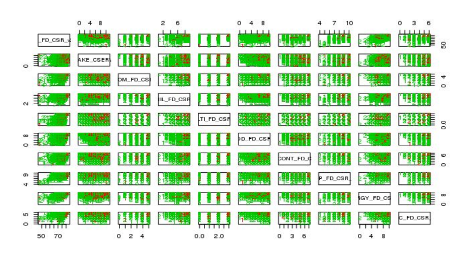


This is a kknn plot similar to the one on bottom but this is with respect to training.



This is an output from the decision tree with respect to the training set with aggregates, any employee who is less than 6.5 has a 54% chance of being Satisfactory.

Go Back



This is an example of the pairs plot output from the kknn algorithm. The different panes represent the data in different dimensions. The green points in the panes represent data points that are labeled correctly and red points are points labeled incorrectly.

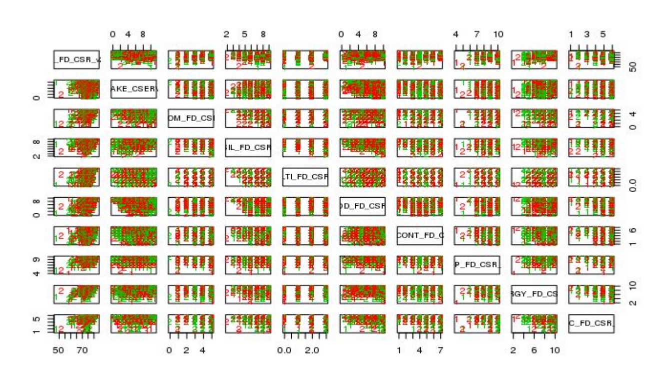
IDEA Rensselaer Institute for Data Exploration and Applications

Project Outmatch

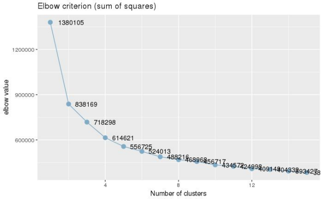
Go Back

90 day turnover prediction models

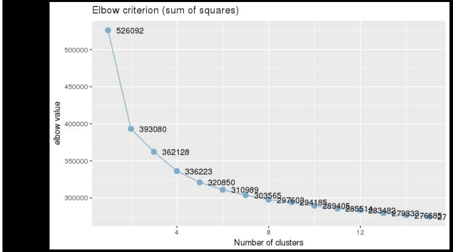
Model Type	Independent Variables	train/test (true positive rate)	Acceptance Rate (pred P/total)	Overall Accuracy train/test
logistic regression	Aggregates	0.74% 0,70%	0.64% 0,74%	56.43% 56.29%
Logistic regression	All explanatory	13.86% 14.79%	11.30% 12.19%	57,21% 57.13%
Random Forest	Aggregates	40.77% 16.84%	19.30% 15.99%	72.66% 55.12%
Random Forest	All explanatory	95.66% 16.40%	41.71% 14.40%	98.08% 56.31%
Decision trees	Aggregates	100%, 100%	100%, 100%	61.39%, 61.39%
Decision trees	All explanatory	100%, 100%	100%, 100%	61.39%, 61.39%
knn	Aggregates	72.83%, 63.41%	62.92%, 62.,99%	65.1%, 53.47%
knn	All explanatory	82.31%, 65%	64.23%, 64.94%	75.29%, 53.72%
kknn (Weighted knn)	Aggregates	74.49%, 61.33%	61.39%, 61.39%	71.96%, 54.26%
kknn (Weighted knn)	All explanatory	99.52%, 68.92%	61.48%, 68.68%	99.11%, 54.85%



The plot above is a pairs plot with respect to the aggregates. This plot shows the classification of kknn, the green parts of the plot are the parts that are labeled correctly and the red parts are the parts labeled incorrectly. The different panes show the data in different dimensions. This pairs plot is with respect to the testing set.

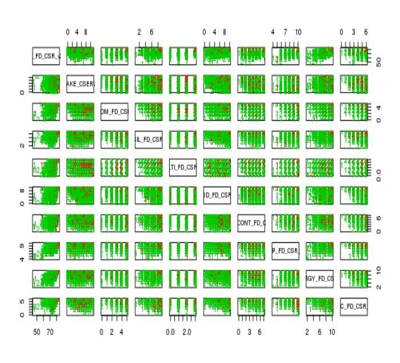


This plot is the plot of the elbow plot for knn that is used to find the optimal value of k. The y axis is sum of squared errors, x is number of clusters



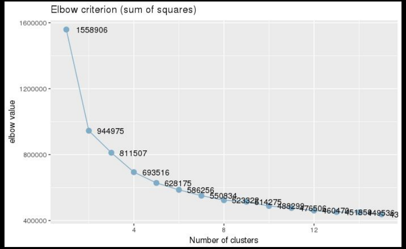
This plot is a plot for knn used to find the optimal value of k.

The plot to the right is the pairs plot with respect to the aggregates, as well.
This time it is with respect to the training dataset.

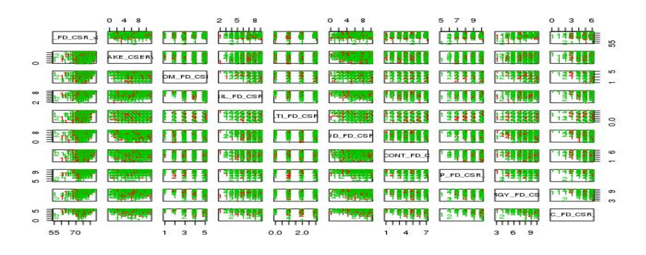


Promotion to ASM/Field models

Model Type	Independent Variables	train/test (true positive rate)	Acceptance Rate (pred P/total)	Overall Accuracy train/test
logistic regression	Aggregates	19.60% 18.75%	14.89% 14.98%	76.00% 75.66%
Logistic regression	All explanatory	22.80% 23.22%	15.00% 15.00%	76.86% 76.98%
Decision trees	Aggregates	100%, 100%	100%, 100%	85.03%, 85.03%
Decision trees	Explanatory	100%, 100%	100%, 100%	85.03%, 85.03%
knn	Aggregates	90.05%, 86.94%	86.74%, 86.97%	81.37% 75.94%
knn	All explanatory	92.7%, 87.2%	86.36%, 87.27%	86.14%, 76.7%
kknn	Aggregates	98.13%, 95.89%	95.45%, 95.76%	86.4%, 82.27%
kknn	All explanatory	99.98%, 91.86%	85.51%, 91.87%	99.35%, 80.11%
Random Forest	Aggregates	100% 100%	100% 100%	86.74% 84.95%
Random Forest	All explanatory	100% 100%	100% 100%	97.72% 85.03%

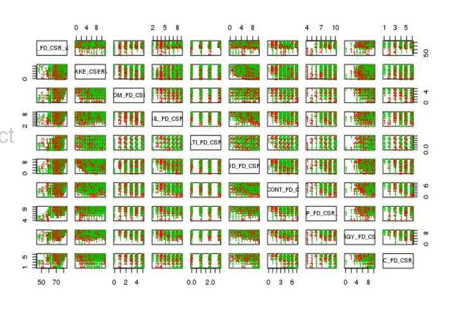


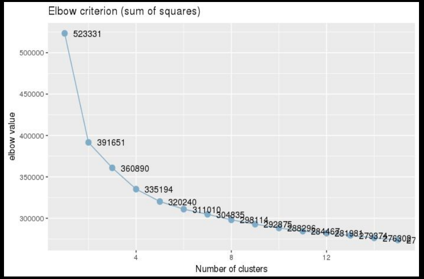
This is an elbow plot used for knn in order to find the optimal value of k with respect to the aggregates.



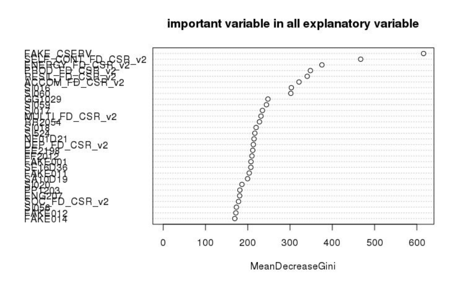
This pairs plot shows the classification of kknn. This pairs plot is of kknn with respect to training set of aggregates. The different panes show the data in different dimensions. The green parts of the plot are the data points that are labeled correctly and red parts are data points that are labeled incorrectly

This plot is a similar pairs plot to the plot above the only difference is that this pairs plot is with respect to the testing set.





This is an elbow plot for knn that is used to find optimal value of k with respect to explanatory variables.



This plot show the importance of each variable base on Random Forest