Notes on Expanding Internal Data Mining Efforts

Tuomo Kareoja

Blackwell

September 18, 2019

Filling Missing Data with a Predictive Model

Problem: Some of the
customer data has missing
values and this could skew
further analyses if not fixed

Customer Brand Preference: Non-Missing VS Predicted

Filling Missing Data with a Predictive Model

Problem: Some of the customer data has missing values and this could skew further analyses if not fixed **Solution:** Create a model to predict the missing values from the values of the other columns

Customer Brand Preference: Non-Missing VS Predicted

Filling Missing Data with a Predictive Model

Problem: Some of the customer data has missing values and this could skew further analyses if not fixed
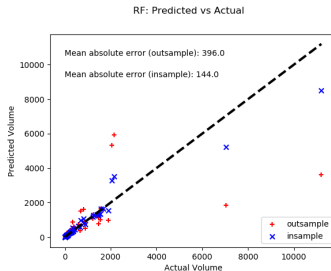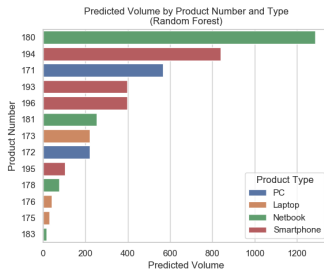**Solution:** Create a model to predict the missing values from the values of the other columns
**Knowledge Gained:** The data are missing at random so the missing values don't skew the analysis

Customer Brand Preference: Non-Missing VS Predicted
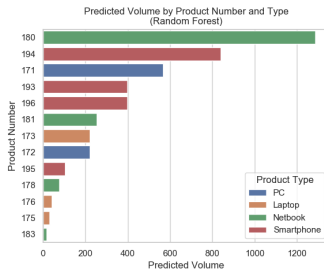
Predicting Sales of Upcoming Products

Problem: We would want to know the future sales volumes of our upcoming products

Predicted Volume by Product Number and Type (Random Forest)

RF: Predicted vs Actual

Predicting Sales of Upcoming Products

Problem: We would want to know the future sales volumes of our upcoming products

Solution: Create a model to predict the sales volume from customer reviews and product attributes

Predicted Volume by Product Number and Type (Random Forest)
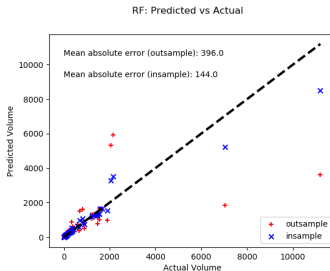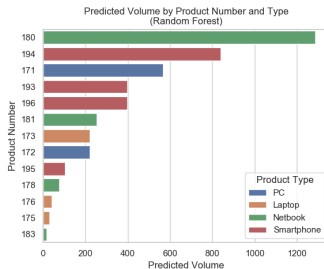

RF: Predicted vs Actual
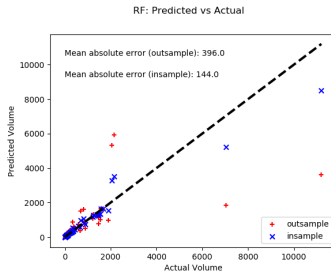
Predicting Sales of Upcoming Products

Problem: We would want to
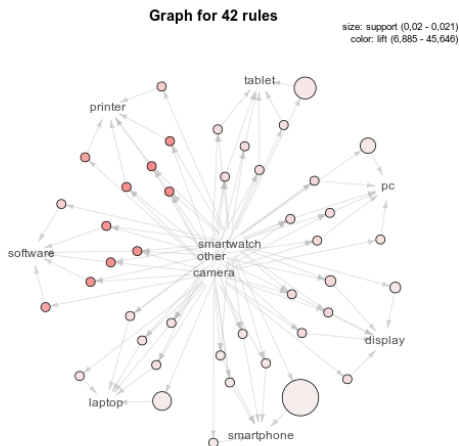know the future sales
volumes of our upcoming
products
Solution: Create a model to
predict the sales volume from
customer reviews and
product attributes
Knowledge Gained: Some
products a predicted to
perform way better than
others, but a better model
would need more data

Predicted Volume by Product Number and Type
(Random Forest)

RF: Predicted vs Actual

Assessing Electronidex Sales Profile

Problem: Is Electronidex
product line compatible to
ours and are there potential
cross-selling possibilities

Graph for 42 rules

size: support (0,02 - 0,021)
color: lift (6,885 - 45,646)

Assessing Electronidex Sales Profile

Problem: Is Electronidex product line compatible to ours and are there potential cross-selling possibilities
Solution: Find product categories that are overlapping and those that fit together and do a market basket analysis of their sales data

Graph for 42 rules

size: support (0,02 - 0,021)
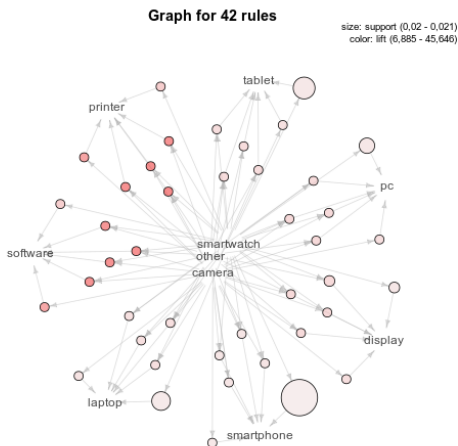color: lift (6,885 - 45,646)

Assessing Electronidex Sales Profile

Problem: Is Electronidex product line compatible to ours and are there potential cross-selling possibilities
Solution: Find product categories that are overlapping and those that fit together and do a market basket analysis of their sales data
Knowledge Gained: About 50 % of our current profits come from products that overlap with Electronidex offerings. There seems to be customer group that we do not serve at the moment and which would be interested in our products.
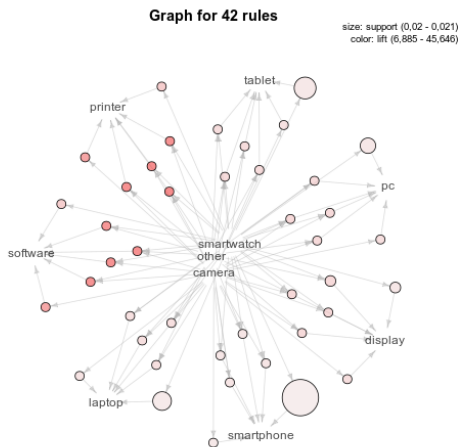
Graph for 42 rules

size: support (0,02 - 0,021)
color: lift (6,885 - 45,646)

Lessons Learned

1. **Lesson:** Data mining can be used to help further analysis of the data by fixing its problems

Lessons Learned

1. **Lesson:** Data mining can be used to help further analysis of the data by fixing its problems
   Advice: Find important datasets where we have problems with data quality and try to fix this by creating models to correct the errors

Lessons Learned

1. **Lesson:** Data mining can be used to help further analysis of the data by fixing its problems
   Advice: Find important datasets where we have problems with data quality and try to fix this by creating models to correct the errors

2. **Lesson:** We can glean predictive information from even very small datasets, but predictive models created from this kind of data can only give very rough estimates

Lessons Learned

1. **Lesson:** Data mining can be used to help further analysis of the data by fixing its problems
   Advice: Find important datasets where we have problems with data quality and try to fix this by creating models to correct the errors

2. **Lesson:** We can glean predictive information from even very small datasets, but predictive models created from this kind of data can only give very rough estimates
   Advice: More investments in gathering data of high quality from our processes and products

Lessons Learned

1. **Lesson:** Data mining can be used to help further analysis of the data by fixing its problems
   Advice: Find important datasets where we have problems with data quality and try to fix this by creating models to correct the errors

2. **Lesson:** We can glean predictive information from even very small datasets, but predictive models created from this kind of data can only give very rough estimates
   Advice: More investments in gathering data of high quality from our processes and products

Lessons Learned

1. **Lesson:** Data mining can be used to help further analysis of the data by fixing its problems
   Advice: Find important datasets where we have problems with data quality and try to fix this by creating models to correct the errors

2. **Lesson:** We can glean predictive information from even very small datasets, but predictive models created from this kind of data can only give very rough estimates
   Advice: More investments in gathering data of high quality from our processes and products

3. **Lesson:** Poorly defined questions are hard to answer especially when datasets are big

Lessons Learned

1. **Lesson:** Data mining can be used to help further analysis of the data by fixing its problems
Advice: Find important datasets where we have problems with data quality and try to fix this by creating models to correct the errors

2. **Lesson:** We can glean predictive information from even very small datasets, but predictive models created from this kind of data can only give very rough estimates
Advice: More investments in gathering data of high quality from our processes and products

3. **Lesson:** Poorly defined questions are hard to answer especially when datasets are big
Advice: Collaborate with data scientists while still at the process of formulating the business questions and before acquiring the data

Lessons Learned

1. **Lesson:** Data mining can be used to help further analysis of the data by fixing its problems
   Advice: Find important datasets where we have problems with data quality and try to fix this by creating models to correct the errors

2. **Lesson:** We can glean predictive information from even very small datasets, but predictive models created from this kind of data can only give very rough estimates
   Advice: More investments in gathering data of high quality from our processes and products

3. **Lesson:** Poorly defined questions are hard to answer especially when datasets are big
   Advice: Collaborate with data scientists while still at the process of formulating the business questions and before acquiring the data

Lessons Learned

1. **Lesson:** Data mining can be used to help further analysis of the data by fixing its problems
   Advice: Find important datasets where we have problems with data quality and try to fix this by creating models to correct the errors

2. **Lesson:** We can glean predictive information from even very small datasets, but predictive models created from this kind of data can only give very rough estimates
   Advice: More investments in gathering data of high quality from our processes and products

3. **Lesson:** Poorly defined questions are hard to answer especially when datasets are big
   Advice: Collaborate with data scientists while still at the process of formulating the business questions and before acquiring the data

4. **Lesson:** Datasets can contain information whose value is not clear from the dataset itself

Lessons Learned

1. **Lesson:** Data mining can be used to help further analysis of the data by fixing its problems
   Advice: Find important datasets where we have problems with data quality and try to fix this by creating models to correct the errors

2. **Lesson:** We can glean predictive information from even very small datasets, but predictive models created from this kind of data can only give very rough estimates
   Advice: More investments in gathering data of high quality from our processes and products

3. **Lesson:** Poorly defined questions are hard to answer especially when datasets are big
   Advice: Collaborate with data scientists while still at the process of formulating the business questions and before acquiring the data

4. **Lesson:** Datasets can contain information whose value is not clear from the dataset itself
   Advice: Pair data scientists with analysts that have relevant domain expertise