MODULE 3 PROJECT

STATISTICAL STUDY ON NORTHWIND DB

EXECUTIVE SUMMARY

September 2019, Nuran Kairakbayev

Hi! Let me give you a brief presentation on the main results of the statistical study of NorthWind database

Research content. Statistical Hypotheses

A. Quantity - Discount

Does DISCOUNT amount have a significant effect on the QUANTITY of a product in an order?

B. Sales - Discount

Does DISCOUNT amount have a significant effect on the CASH VALUE (sales) of a product in an order?

C. Shipper – Delivery

Does SHIPPER have a significant effect on the DELAY of product delivery in an order?

D. Discount predictors

What FACTORS have a significant effect on DISCOUNT levels?

E. DISCOUNT level classifier model

Is it possible to create a CLASSIFIER model on DISCOUNT level with robust level of prediction accuracy?

So, briefly about the hypotheses considered in the present paper. I have studied 4 main hypotheses, namely, the impact of the discount rate on the volume and value of ordered goods the impact of shipper on delay and the impact of various factors on the size of the discount. I also tried to build a predictive classification model, which should predict the size of the discount. So, let's look on those slide-by-slide

A. QUANTITY – DISCOUNT (1/5)

Key Question

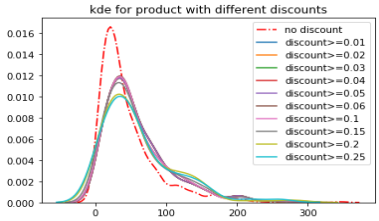
Does DISCOUNT amount have a significant effect on the QUANTITY of a product in an order?

The first hypothesis is the effect of the discount on volume of ordered products.

A. QUANTITY – DISCOUNT (2/5)

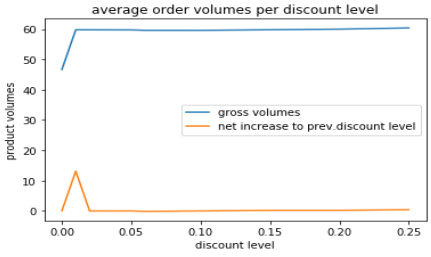
Does DISCOUNT amount have a significant effect on the QUANTITY of a product in an order?

Visual analysis



Product Quantities per Order less for Products without discounts

Marginal increase of Product Quantities:
0% - 2% discount rates



Total Effect and Marginal Effect will be tested in statistical analysis

Here is a preliminary visual analysis. On the upper chart you can see the product volume densities in distributed by groups with different discount rates. The no-discount group is clearly different from all the others. However, other groups are quite similar between each other. If you pay attention to the lower chart, you can already see the marginal effect, i.e. how much the average volume of the ordered goods differs from the same one at the previous discount level. As you can see, there is a noticeable difference only in small values.
Thus, further statistical testing will take these issues into account.

A. QUANTITY – DISCOUNT (3/5)

Does DISCOUNT amount have a significant effect on the QUANTITY of a product in an order?

	Test Version 1: testing gross difference	Test Version 2: testing marginal difference
Methodology	• Ho: E{Quantity | no discount} = E{Quantity | discount} • Testing method: Welch-test (2-tail) • Loop: varying discount groups by discount level	• Ho: E{Quantity | prev.discount} = E{Quantity | next.discount} • Testing method: Welch-test (2-tail) • Loop: varying intra-discount groups by discount level

A few words about how the studies were conducted. Here I used single AB-testing using Welch statistics with 5% p-value threshold. The null hypothesis is state that average quantity of product is the same for subgroups with different discount levels. For Version 2 I compared quantities for following subgroups: group A (discount less than X), group B (discount equal or more than X). This type of division provides to test marginal effect indicated in previous slide.

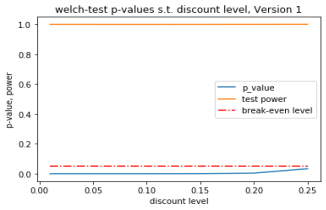
A. QUANTITY – DISCOUNT (4/5)

Does DISCOUNT amount have a significant effect on the QUANTITY of a product in an order?

Results

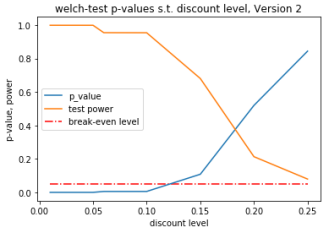
Any discount rate – affect on Product Volumes

- P-values: <5% for all discount levels
- Test power: good for discount levels



Positive marginal effect on Volumes – for 1%-10% discount rates

- P-values: <5% for discount levels in range 1%-10%
- Test power: good for discount level in range 1%-10%



First version, whether the presence of a discount affects the volume. The answer is yes, it has a positive effect, and any level of the discount rate (starting from one percent) gives positive difference with no-discount volumes.
However, if we study the marginal effect (Version 2), the influence of the discount size has a statistically significant effect only in the range from 0% to 10%. This means that if the rate rises by more than 10%, it will not have an effect different from the 10% effect of setting the rate. This is graphically shown in the graph below.

A. QUANTITY – DISCOUNT (5/5)

Does DISCOUNT amount have a significant effect on the QUANTITY of a product in an order?

Issues and Recommendations

ISSUES:

- **Discounts on prices increases product quantities in an order, but only for 1%-10% range of discount levels.**

RECOMMENDATIONS:

- **Use minimum discount rates (1%-10%) for product volumes improvement**

My recommendation is that in order to increase the volume of ordered goods it is enough to stick to the range of 1%-10%.

B. SALES – DISCOUNT (1/4)

Key Question

Does DISCOUNT amount have a significant effect on the CASH VALUE (sales) of a product in an order?

As a rule, the volume of revenue and profit is more important for the business than the physical volume of goods sold, so as an additional hypothesis I considered whether the discount rate affects the volume of revenue from goods in the basket.

B. SALES – DISCOUNT (2/4)

Does DISCOUNT amount have a significant effect on the CASH VALUE (sales) of a product in an order?

	Test Version 1: testing gross difference	Test Version 2: testing marginal difference
Methodology	• Ho: **E{Quantity | no discount} = E{Quantity | discount}** • Testing method: **Welch-test (2-tail)** • Loop: varying **discount groups** by discount level • Sales – **undiscounted original price**	• Ho: **E{Quantity | prev.discount} = E{Quantity | next.discount}** • Testing method: **Welch-test (2-tail)** • Loop: varying **intra-discount groups** by discount level

The general testing algorithm is the same as for the previous hypothesis, so I will not dwell on the methodology. I will only note that I took the initial prices of goods for the calculation of revenue, before the application of the discount.

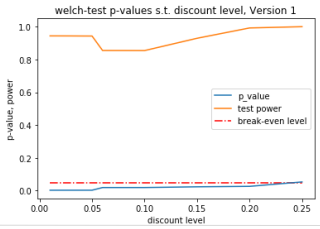
B. SALES – DISCOUNT (3/4)

Does DISCOUNT amount have a significant effect on the CASH VALUE (sales) of a product in an order?

Results

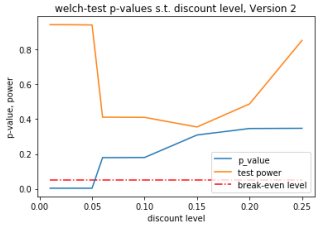
Any discount rate – affect on Product Sales

- P-values: <5% for all discount levels
- Test power: good for discount levels



Positive marginal effect on Sales – for 1%-5% discount rates

- P-values: <5% for discount levels in range 1%-5%
- Test power: good for discount level in range 1%-5%



The testing result is generally similar to the previous hypothesis, however, as it turned out, the effective and, accordingly, recommended range of discount rates for revenue is 1%-5%.

B. SALES – DISCOUNT (4/4)

Does DISCOUNT amount have a significant effect on the CASH VALUE (sales) of a product in an order?

Issues and Recommendations

ISSUES:

- **Discounts on prices increases product Sales in an order, but only for 1%-5% range of discount levels.**

RECOMMENDATIONS:

- **Use minimum discount rates (1%-5%) for sales volumes improvement**

My recommendation is that in order to increase the Sales of ordered goods it is enough to stick to the range of 1%-5%.

C. DELAY – SHIPPERS (1/4)

Key Question

Does SHIPPER have a significant effect on the DELAY of product delivery in an order?

The following hypothesis asks whether the choice of shipper affects the delivery delay.

C. DELAY – SHIPPERS (2/4)

Does SHIPPER have a significant effect on the DELAY of product delivery in an order?

	Test Version 1: Multiple AB-testing	Test Version 2: ANOVA model
Methodology	• **Delay** = ShippedDate – RequiredDate • Ho: **E(Delay | Shipper 1) = E(Delay | Shipper 2)** • Testing method: **Welch-test (2-tail)** • Loop: varying in **all combinations** of groups by shipper pairs	• Formula: **Delay ~ C(Shipper)**

Delay here means the difference between shipped and required dates. Since there are only 3 shippers in the database: Federal Shipping, United Packages and Speedy Express, I used multi-group AB testing and ANOVA model.

C. DELAY – SHIPPERS (3/4)

Does SHIPPER have a significant effect on the DELAY of product delivery in an order?

Results

Multiple AB-testing:

- **No significant difference between shippers for delay**
- P-values: <5% for 3 AB-tests (0,10 – 0,55)
- Test power: weak for all 3 AB-tests (0,09 – 0,34)

ANOVA-model

- **No significant difference between shippers for delay**
- P-value (F-stat) = 0,27

The results of both tests (multiple AB and ANOVA) showed that in general, the delay values are similar.

C. DELAY – SHIPPERS (4/4)

Does SHIPPER have a significant effect on the DELAY of product delivery in an order?

Issues and Recommendations

ISSUES:

- **There is no significant effect of shipper on level of shipping delay**

RECOMMENDATIONS:

- **Analyze and compare shipping tariffs for all shippers with same service quality**

Therefore, as a recommendation, taking to account identical service measure in terms of delivery dates, I propose to conduct a price (tariff) analysis of shippers.

D. DISCOUNT PREDICTORS (1/4)

Key Question

Are there any significant PREDICTORS for DISCOUNT rate?

In this section I tried to find factors that could affect the discount rate itself. So, it is opposite problem compare to first hypothesis

D. DISCOUNT PREDICTORS (2/4)

Are there any significant Predictors for Discount rate?

	Candidates for testing	Testing parameters

? Methodology	• **Not-ordered product in Stock**
	• **Product Category (only ANOVA)**
	• **Supplier (only ANOVA)**
		• Methodology:
		  • **Single AB-test –** for 2-class predictor
		  • **ANOVA –** for multi-class predictor
		• **Ho: E{Discount | Factor = A} = E{Discount | Factor = B}**
		• Statistics for AB-test: **Welch-test (2-tail)**

Based on some logical assumptions, I have identified the 3 most suitable candidates for this: Goods in stock but not ordered, product category, Supplier. Depending on selected variable data structure I've applied different methods (single AB-test and ANOVA).

D. DISCOUNT PREDICTORS (3/4)

Are there any significant Predictors for Discount rate?

Results

No significant predictors for Discount level:

- **Not-ordered product in Stock**:
  - p-value (AB) = 0,14
  - p-value (F-stat) = 0,75
- **Product Category:** p-value (F-stat) = 0,15
- **Product Supplier:** p-value (F-stat) = 0,06 (weak significance)

As you can see, the only factor close to the 5% significance boundary is the Supplier with 6% p-value. All other candidates has very high p-values.

D. DISCOUNT PREDICTORS (4/4)

Are there any significant Predictors for Discount rate?

Issues and Recommendations

ISSUES:

- There are no Predictors with significant relationship with Discount rate

RECOMMENDATIONS:

- Continue to analyze possible Discount rate factors

The main result of this section is that within selected candidates we didn't find significant predictor. But understanding nature and factor of discount rate is important for business planning, so, as for recommendation I propose to continue to analyze Discount rate factors.

E. DISCOUNT CLASSIFIER (1/4)

Key Question

Is it possible to create robust CLASSIFIER MODEL for DISCOUNT rate?

From the previous slides, we saw some statistical relationships between discount rate, price and sales volume, and supplier. This led me to think about trying to create a classification model that predicts the size (class) of the discount rate by price, volume and supplier factors.

E. DISCOUNT CLASSIFIER (2/4)

Is it possible to create robust classifier model for discount rate?

Methodology

Model Parameters

- Model type**: Naïve Bayes Classifier**
- Target: **Discount class**
- Model options:
  - **2-class** (discounted vs no-discounted)
  - **6-class** (no-discount, 1-5, 5-10, 10-15, 15-20, 20+)
- Model predictors:
  - **Product Price (before discount)**
  - **Quantity**
  - **Supplier**

Methodologically, I built the Naive Bayes Classifier model. Calculations were also made for two options: one guessing only the presence of the discount as such (binomial), and the other guessing the discount range (from 0 to 25% in 5% step).

E. DISCOUNT CLASSIFIER (3/4)

Is it possible to create robust classifier model for discount rate?

Results

Classifier model accuracy: moderate level

- 2-class model: accuracy score ~ 61%
- 6-class model: accuracy score ~ 60%

The modeling results are as follows: Both options showed approximately the same accuracy scores, around 60% for test subsets.

E. DISCOUNT CLASSIFIER (4/4)

Is it possible to create robust classifier model for discount rate?

Issues and Recommendations

ISSUES:

- **Accuracy level of classification model is low on given factors (~60%)**

RECOMMENDATIONS:

- **Continue to analyze possible predictors for discount rate**

60% of the forecast accuracy is quite moderate, so as a recommendation, I can say that building a predictive model may require a deeper analysis of the data, and maybe using alternative types of models.

Business Recommendations

1. Use discounts on product in small range (1%-10%) for increasing volumes of product sales

2. Use discounts on product in smaller range (1%-5%) for increasing gross revenues of product sales

3. Shipping services for all shippers are similar, so check and compare shippers' tariffs on inequality

4. Discount rate for a product could be predicted, but with low level of accuracy

Future Work

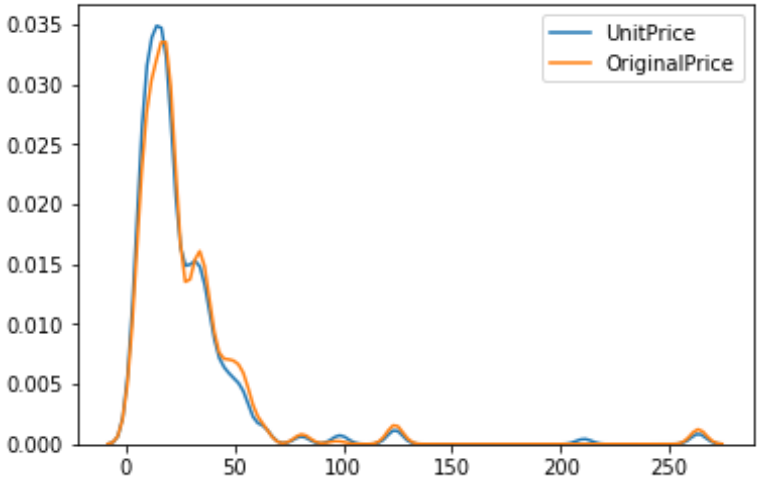
1. Analyze and test other predictors for Sales and Volumes

2. Find and understand significant factors on Discount volumes

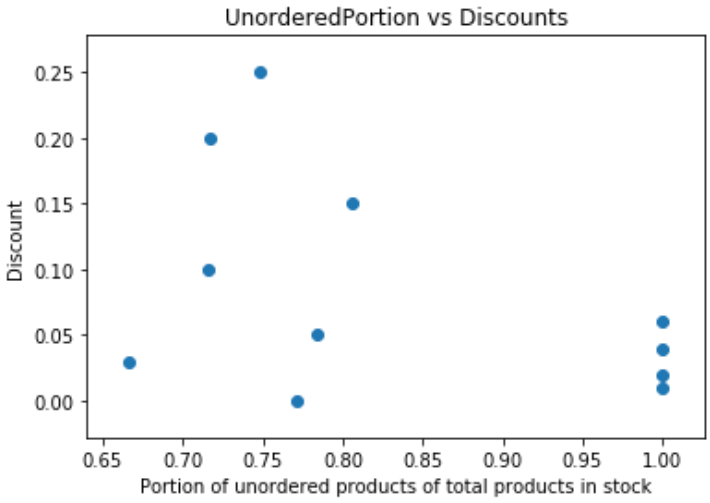
3. Based on above, create classifier model on Discount rates

APPENDIX

A1. Original price VS Discounted price



A2. Avg.Discount per Unordered product portion



A3. Avg.Discount per product category