Predicting Probability of Debt Recovery

Knime Spring Summit March 2018

HAYATVARLIK

An introduction to Hayat Varlık & today's presenters

Hayat Varlık one of the leading asset management companies of Turkey, was originally established in 2008 under a resolution passed by Banking Regulation and Supervision Agency.

ELİF YİĞİT AYCAN

HEAD OF DECISION SCIENCES / R&D

MESRUR BÖRÜ

SENIOR DATA SCIENTIST / R&D

The main activity of Hayat Varlık is to buy and sell receivables and other assets from banks and other financial institutions in order to turn indebtedness to the cash or to restructure and sell assets and to collect, restructure or sell to third parties the receivables and other assets by supporting and consulting services.

	Acquisition	Collection	Legal enforcement
	Portfolio acquisitions	*Collections on portfolio & cost to collect*	
Key value drivers	A cumulative total of TL 10.8 billion, 53% commercial and 47% individualized overdue receivables portfolio (as of the end of 2016) **Customer intelligence** and **data analytics** for pricing	**Customer intelligence** and **asset tracing** **Customer management** and negotiation **Data analytics** / systems	**Customer intelligence** and **asset tracing** Use of external collection partners

Innovative business model and platform with nationwide infrastructure

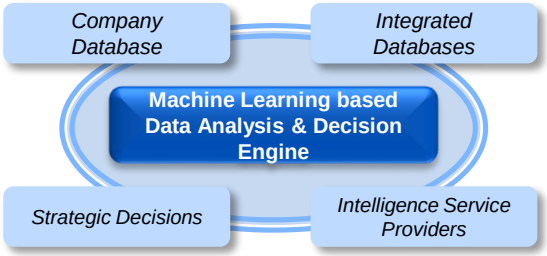
Customer-centric approach underpinned by data analysis, machine learning and nationwide infrastructure

Data and technology driven approach

- Data and analytics underpins an **automated collection platform**
- **Customer intelligence** & **in-depth tracing** capability
- **data analysis based on machine learning algorithms** and decision engine
- **Proprietary databases** supporting contact, offer and intelligence strategy

Certified R&D centre in Turkey

Tech-driven approach

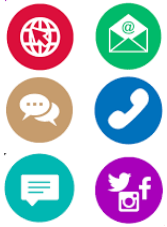


Amicable collection strategy

- Debtors treated like **customers**
- **Customised payment plans**
- Distinctive **customer scorecard system**
- **Dedicated specialist team** focused on challenging accounts

Collections mostly through settlements

Omni-channel communication



Digital and automated collection

Self-service debt repayment tool

Established and nationwide infrastructure

- **Extensive call centre network in Turkey:** in house and outsourced agents
- In-house legal team
- Only NPL AMC with **offices outside Istanbul**
- **Proximity to courts** facilitate legal follow-up

Traditional approach of debt purchasers to predict probability of debt recovery is to use customer behavior data.

Method

- Customer bahavior is encoded in the transactional data collected over years
- Traditional approach tries to uncover patterns in behavior data using predictive machine learning models

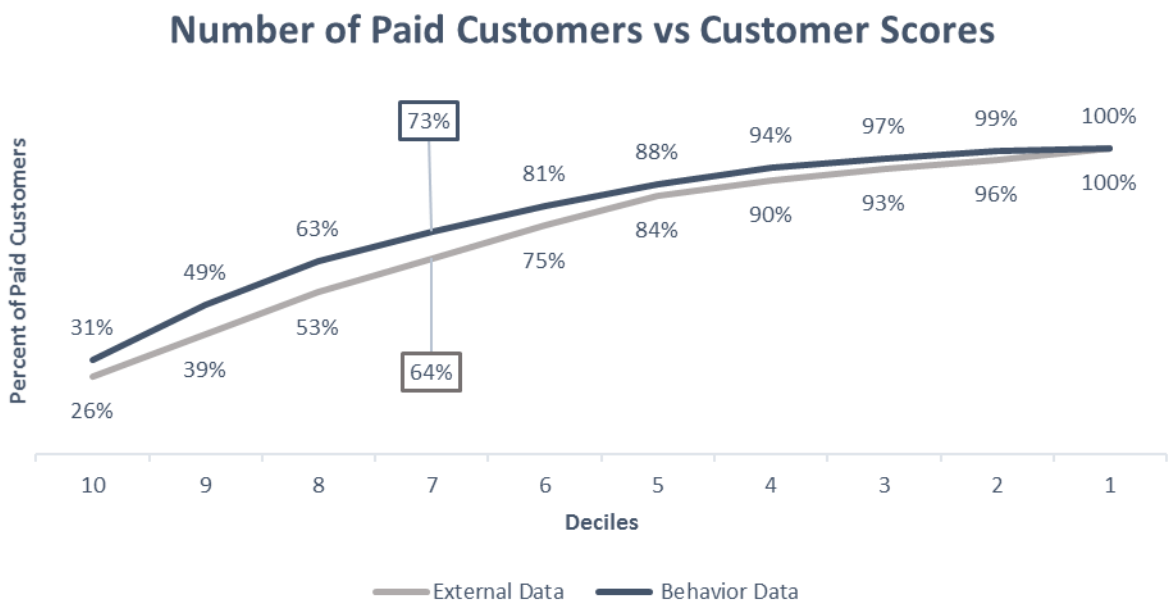
Challenges

- Behavior data is not available for new customers
- Data consists of highly correlated features
- Only small portion of customers pay their debts (Unbalanced dataset)

Solutions

- Utilize external data sources when historical data is not available
- **Use unconventional data sources (Call center recordings)**

Results obtained by traditional approaches (February 2018)



- **Predictive models that use only external data :** Customers in the top 4 deciles cover 64% of paid customers.
- **Predictive models that use historical behavior data:** Customers in the top 4 deciles cover 73% of paid customers.

To increase the performance of predictive models we used NLP methods on call center speech data.

Method

- Customer and Agent speech is converted to text using speech analytics tool
- Structured data is produced using Knime
- Predicted probability of debt recovery using last months speech data

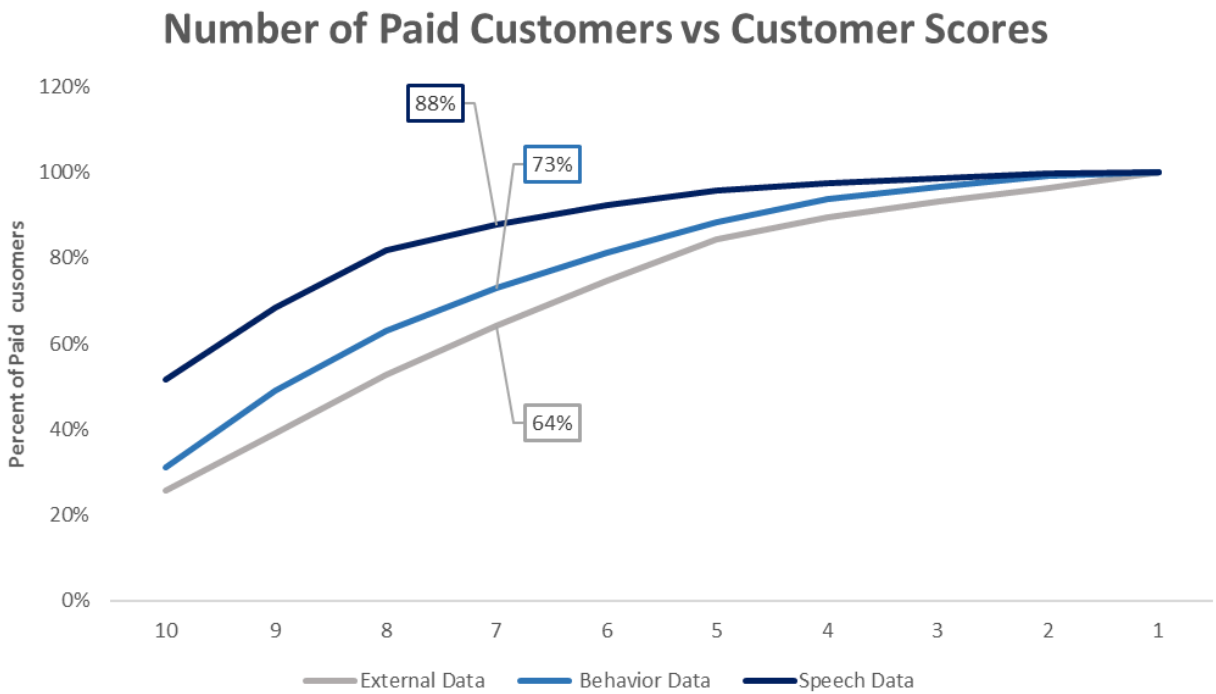
Challenges

- Lots of sparse datasets can be produced (In example: number of times a term is used, max number of times/max number of times/min number of times a word is used in speech)

Solutions

- Used stacking to combine results from models genereted using different data sets.

Results obtained by using speech data (February 2018)



- **Predictive models that use only external data :** Customers in the top 4 deciles cover 64% of paid customers.
- **Predictive models that use historical behavior data:** Customers in the top 4 deciles cover 73% of paid customers.
- **Predictive models that use speech data**: **Customers in the top 4 deciles cover 88% of paid customers.**

Preprocessing

1 *A Turkish NLP library is used for stemming and tagging words with negativity suffixes.*

Method

- Text is stemmed using a Turkish NLP Library

Challenges

- In Turkish series of suffixes added to stem may change the meaning.

Solutions

- Negativity and question suffixes are used for tagging stemmed words. (In example: pay, pay-,pay_?)

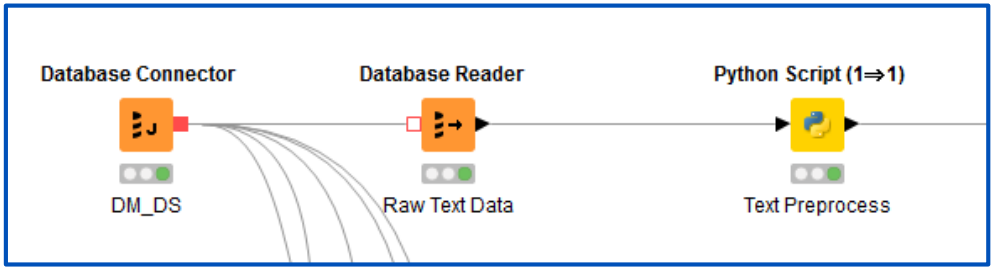
Pos tagging

- Ödemedim

Öde:	FIIL_KOK (**Root**)
me:	FIIL_OLUMSUZLUK_ME(**Negativity**)
di:	FIIL_GECMISZAMAN_DI(**Past tense**)
m:	FIIL_KISI_BEN (**First person singular**)

- I did not pay

Pay:	Root,
Not:	Negativity,
Did:	Past tense,
I:	First Person Singular




```
Suffix = "OLUMSUZ"
for i in range(len(SplitText)):
    StemList = []
    for v in range(len(SplitText[i])):
        try:
            word = SplitText[i][v].encode("utf-8")
            WordAnalysis = z.kelimeCozumle(word.decode("utf-8"))
            suffixes = str(WordAnalysis[0].ekler())
            if Suffix in suffixes:
                stem = WordAnalysis[0].kok().icerik()+"-"
            else:
                stem = WordAnalysis[0].kok().icerik()
            StemList.append(stem.encode("utf-8"))
        except: StemList.append(word)
    StemmedText.append(StemList)
```

Data preparation

2 *18 different datasets from text data and 1 dataset from customer behavior are prepared.*

Method

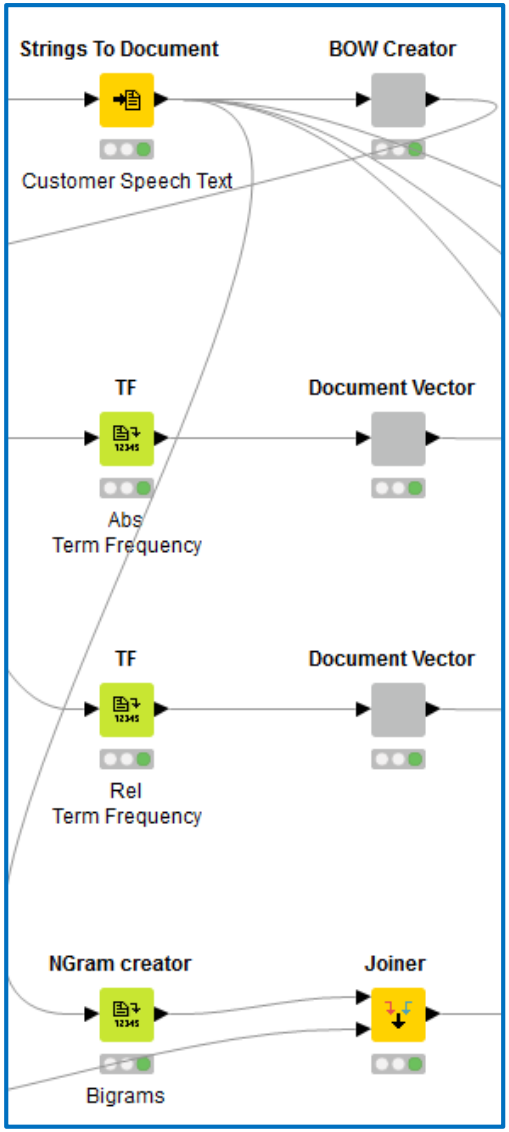
- For each customer, different datasets are prepared from customer speech text and agent speech text.
- Customer behavior data is also used in models.

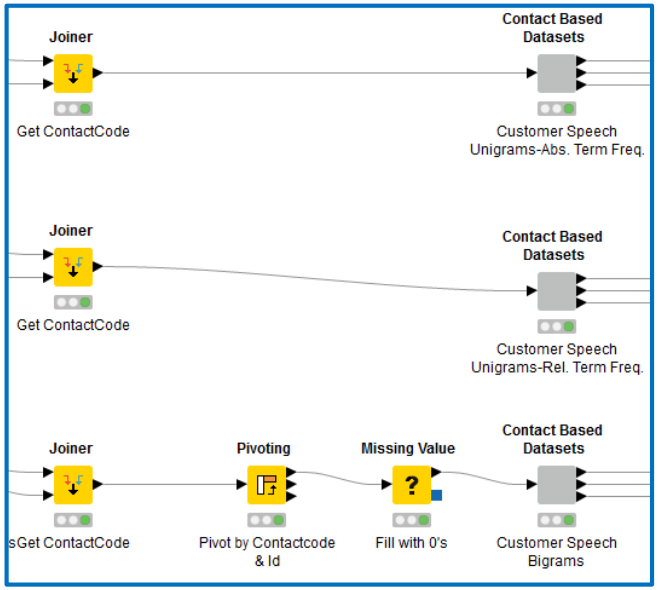
Challenges

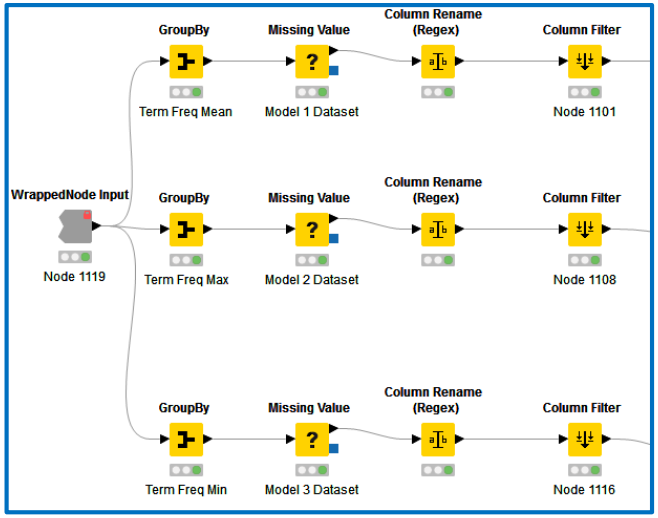
- Sparse datasets

Solutions

- Frequent unigrams/bigrams are selected.
- Remaining unigrams/bigrams are used to built 1-level tree based models. Features with low importance are droped.
- For each customer 18 different datasets are prepared.







Stacking

3 *18 different datasets are used to predict probabilty of making a payment with 2 different boosting algorithms.*

Method

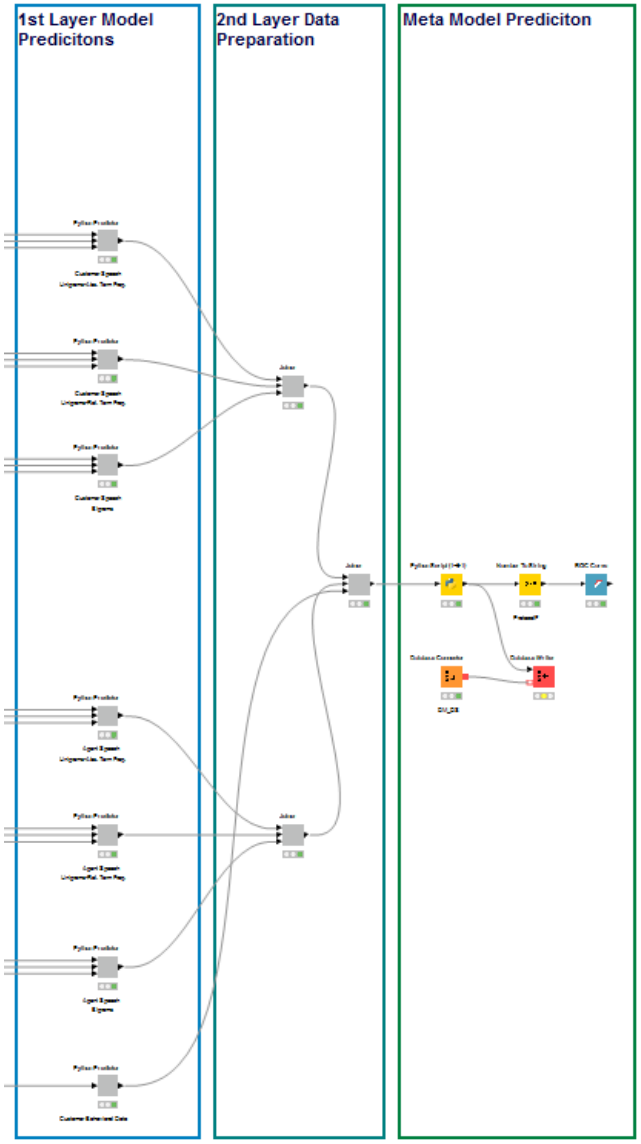
- Each dataset is used to predict probability of making a payment with 2 different boosting algorithm.

Challenges

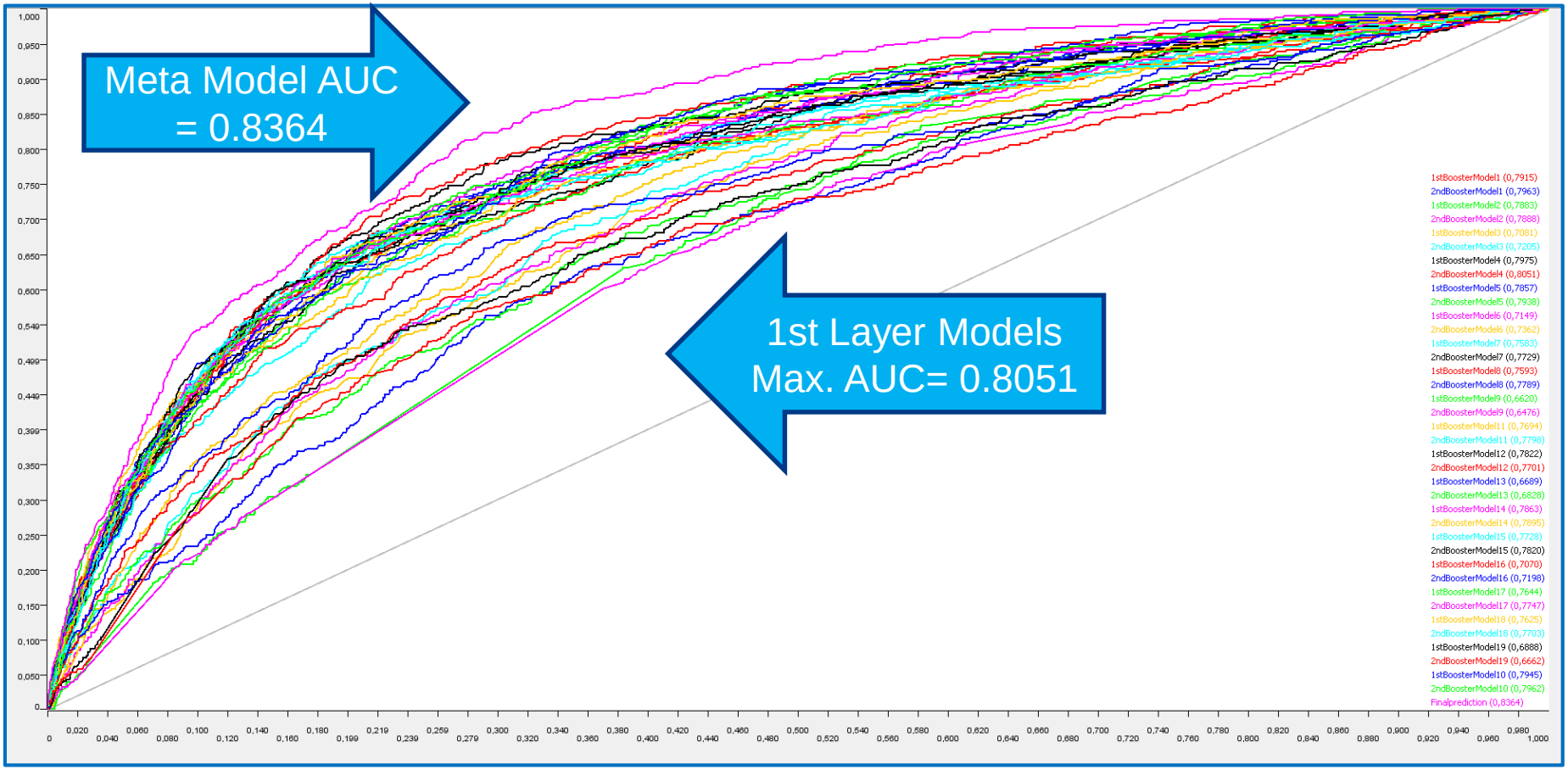
- None of the models produced significantly better results compared to customer behavior model.

Solutions

- Probabilities produced by each model is combined to generate a meta model dataset containing 38 features.
- A meta model is built to predict probability of payment



Metamodel has much better performance than 1st layer models.



Only 40% of the customers contacted in the last month are needed to generate 88% of total payment made in next 15 days.

Benefits of Project

- Only 40% of the customers contacted in the last month are needed to generate 88 % of total payment made in next 15 days.
- Collection efforts are focused on top 40% customers.

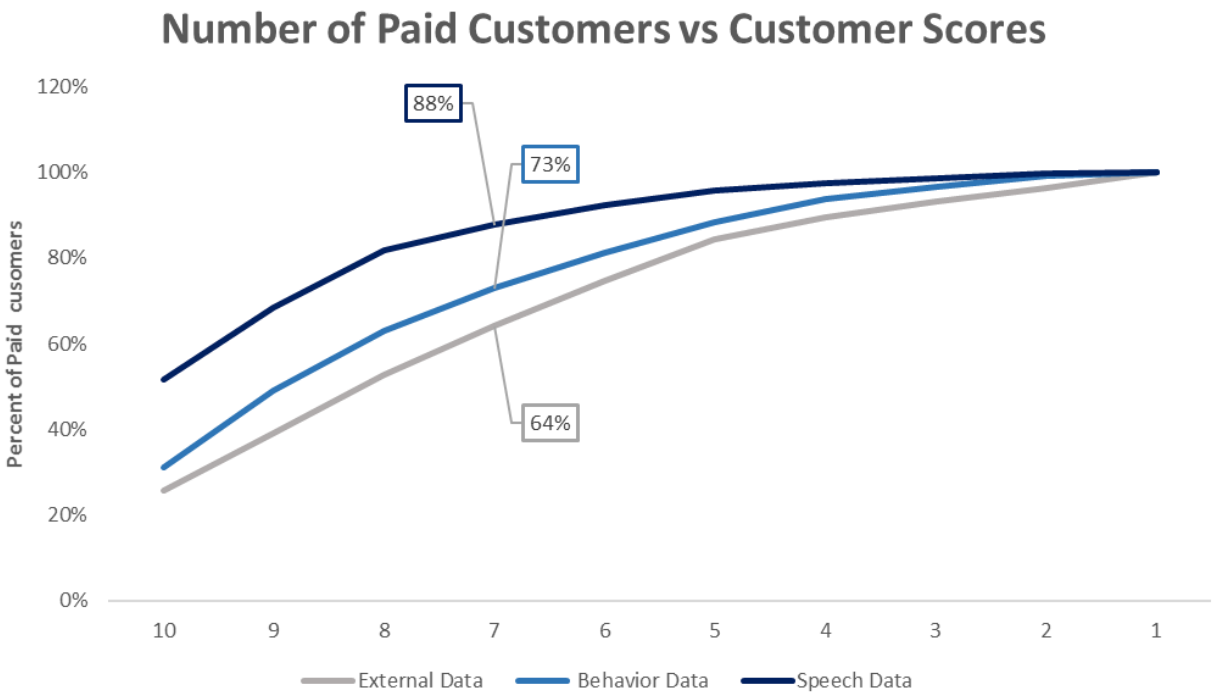
Benefits of Knime

- Provides a platform that can be easily integrated with popular programing languages.
- Can be easily scheduled and used for process automation.

Ongoing projects using Knime

- Using machine learning models to built strategies for each step of customers life cycle

Results obtained by using speech data (February 2018)



- **Predictive models that use only external data :** Customers in the top 4 deciles cover 64% of paid customers.
- **Predictive models that use customer behavior data:** Customers in the top 4 deciles cover 73% of paid customers.
- **Predictive models that use speech data**: **Customers in the top 4 deciles cover 88% of paid customers.**

Thank you…

HAYATVARLIK

Knime Spring Summit March 2018