

The background is a dark blue grid with various data elements. On the left, there are two overlapping sine waves, one solid and one dashed, with callout letters 'a', 'b', and 'c' pointing to specific points. The main title 'RANK MINER' is centered, with 'PREDICTIVE ANALYTICS' below it. The letters 'I' and 'M' in the title are replaced by vertical bars of different heights and colors (blue, grey, blue). Faint, semi-transparent text in the background includes agent IDs (e.g., 111563213, 4551455, 155123456, 4628626, 41263623, 19653253, 56123165, 4926365), dates (AUG 14, NOV 16, FEB 18, AUG 21, SEP 19, MAY 22, DEC 15), durations (3:02, 4:42, 1:43, 5:04, 3:32), and agent IDs (1763, 2563, 4252, 1462, 1482, 1652, 1562).

RANK MINER

PREDICTIVE ANALYTICS

Leveraging voice-based emotions and behaviors
to predict agent and customer success.

Outline

- Introduction to RankMiner
- Project overview
- Data source descriptions
- Limitations of data
- Data science “in the real world”
- Glossary of terms

Introduction to RankMiner

- Predictive analytics company focused on voice-based insights
 - Serving call centers in collections, sales, customer service, and educational spaces
 - Mainly telephony-based audio
- Technologies leveraged
 - Signal manipulation (speaker diarization, speech recognition)
 - DSP feature extraction (voice) & emotional identification
 - Proprietary machine learning & pattern classification algorithms
 - (upcoming) Dynamic feature extraction & selection for real-time predicting & reporting
 - (upcoming) NLP-based feature extraction & speech analytics
 - (upcoming) Speaker identification

Introduction to RankMiner

- Applications
 - Identifying customers most likely to spend \$\$\$
 - Debt collections, up-selling & cross-selling, etc.
 - Evaluating phone agent “soft skill” performance
 - Grading behavioral quality for QA & training purposes
 - Reducing phone agent attrition
 - Predicting agents that are at risk for resigning without reason / notice
 - Pre-empting declining agent performance
 - Gauging language & reading comprehension
 - Analyzing prosodic elements to grade speaker’s acuity in vocabulary, grammar, and flow of punctuation

Project Overview

- Reducing phone agent attrition
 - For a debt collections company
 - Feature space: audio signal-derived & limited auxiliary quantities
 - Prediction targets: phone agents
 - Prediction space: spectrum of likelihood
(e.g. 0 := least likely to terminate → 1 := most likely to terminate)
- Objectives
 - Fathom data, link sources
 - Define data point formation
 - Determine feature set
 - Verify / propose new targeting scheme
 - Determine data point set
 - Train & cross-validate predictive model
 - Report performance, interpret results, identify strengths/weaknesses, suggest further research, etc.

Data source descriptions

- Agent data
 - Work record & budget items
 - Hire & termination dates, termination reason, assigned group, work shift, hours worked
 - Monthly base pay rate, revenue generated, earned commission
 - Collected over 6 mos. (Jul – Dec 2015)
 - Aggregated from 8 different sources
 - Restricted to English-speaking collectors
- Call data
 - Call date/time, duration, debtor account ID, collection group, call outcome, agent taking the call
 - Single data source collected 18 Mar – 31 Dec 2015
 - Restricted to outbound “Right Party Contacts” over 40 seconds for the agents given in agent data
- Feature data
 - 176 Audio recording-based features (RankMiner proprietary calculations)
 - Based on agent-only audio, restricted to audio files given in call data
 - Extracted from calls ranging 26 May – 31 Dec 2015
 - Assigned target values for selected feature vectors

Limitations of data

- Agent data source disconnects
 - Imperfect merging on hire dates & payroll IDs
 - Missing work schedule & budget data for some months for some agents
 - Likely: early removal from HR records erased terminated agent's last month
 - HR data sporadically missing
- Incomplete call data set
 - Doesn't contain record of all calls held for a given agent
 - Inbound calls not included
 - Other outbound calls recorded by a different data source (not included)
 - Earlier calls missing collection group, other minor fields
- No verbatim link between call & feature data
 - Audio file names altered during feature extraction (can still merge on substring)

Data science “in the real world”

- Fiddling with data is like 50% of the job
 - Most of the data collectors don't give a #&%! about the data
 - “When up to your neck in alligators, it's easy to forget that the goal is to drain the swamp”
 - Applies to ETL AND analysis!
- Combinatorial by nature
 - Tons of options at each step will require ingenuity, experience, and elan to make directed choices as what to test
- Not for the narrow-minded
 - Truly an interdisciplinary field
 - Data handling; programming skills; statistical modeling; AI theory; experimental design; domain knowledge ⇒ IT/IS; software engineering; statistics/business; applied math, computer science, EE; natural sciences; various
 - Deconstruct “established methods” – embrace Exploration vs. Exploitation

Data science “in the real world”

- ... And explaining to the laypeople is the other 50%
 - People like their “established methods”; you’ll have to spend a lot of time justifying moving out of the comfort zone
- You’re at 100%. What about the actual analysis?
 - Yeah, about that... expect to put in overtime
 - Absolutely everything is results driven.
Translation: nobody cares how long it takes
 - Resist the urge to cut corners
 - Ensure “cycle time” is accounted for
 - DO NOT MAKE ASSUMPTIONS!

Glossary of terms

- Lots of codified values in the data
 - Their meanings can help provide context & craft hypotheses
 - Some data can be linked across sources but will require context matching
 - I ran out of time; there'll be an appendix in the project plan provided to you