



PREDICTING FAILURE MODE FOR FIELD CLAIMS

STEVEN GLEMBOCKI



BACKGROUND

- Quality issues with cars are logged at service centers
 - Customer complaint
 - Cause
 - Narrative
 - Description
 - Correction
 - Narrative
 - Description
 - Parts used
 - Rollup symptom
- Data is aggregated and given to Engineering to solve issues in the field

PROBLEM STATEMENT AND HYPOTHESIS

- Challenges aggregating and troubleshooting issues
 - Limited amount of information
 - Mostly free text fields
 - Can be very generic and inaccurate
 - Not standardized
- Hypothesis is that we can predict these more granular tags using the data we do have
 - Supervised machine learning using claims that have been assigned failure modes
 - Predict or limit options for assignment

DATA SET

- Obtained from service system using SQL to fetch all relevant information
- Fields include:
 - VIN
 - Model ID
 - Complaint Narrative
 - Cause Narrative
 - Cause Description
 - Correction Narrative(s)
 - Fault ID (what we are predicting)

DATA SET (CONT.)

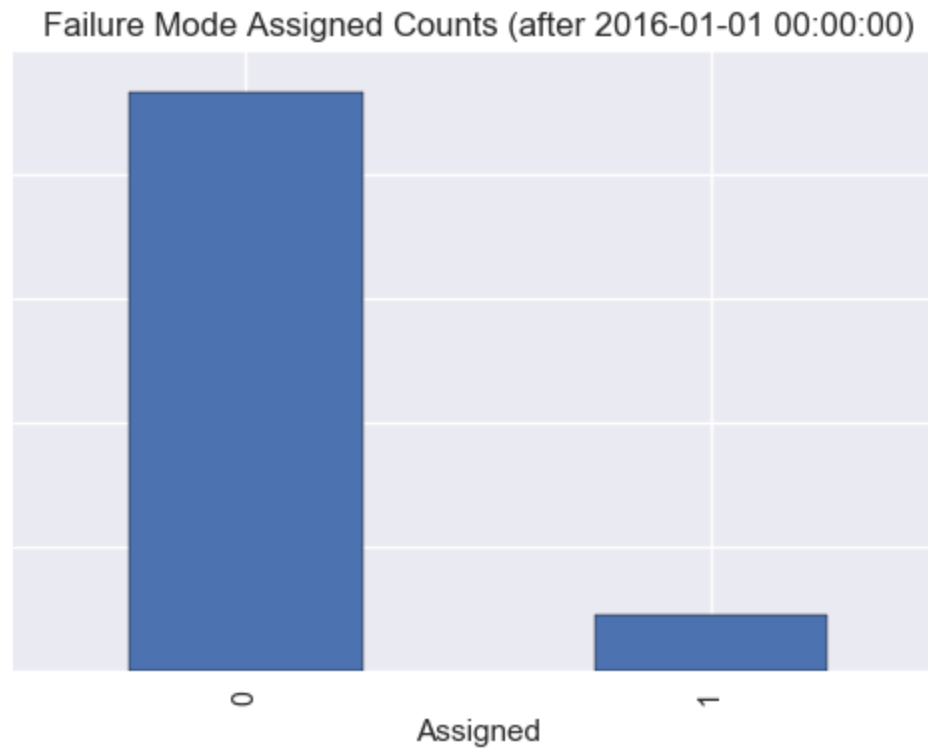
Out[12]:

	Vin	ComplaintNarrative	CauseDescription	CauseNarrative	FaultModelID	FailureMode	RollupSympt
14	5YJSA1H21FFP77752	Customer states: High Power Wall Connector (HP...	Inoperative	NaN	471	Wall Connector Pilot Fault	High Power V Connector (H fails to charg.
20	5YJSA1H43FF081959	Customer states rear right door is leaking air...	Poorly Fitted	misaligned	346	Primary Seal Poor Fitment	Window seal leaks air/wate
25	5YJSA1H10FF081469	Customer states: Guest reports HPWC is failing...	NaN	NaN	470	Wall Connector Loose Crimp	High Power V Connector (H fails to charg.
27	5YJSA1H2XFF081410	Customer states: the passenger side visor does...	Distorted Warped / Bent	Sun visor retaining clip fatigued.	437	Sun Visor Clip Broken	Sun visor will latch into hoo
32	5YJSA1H2XFFP78740	Customer states: vehicle has 3 alerts, will no...	Seized/Stiff/Sticking/Binding	Vehicle towed in with BMS_f152_SW_pos_contacto...	214	HV Contactor - Stuck	12v Battery A

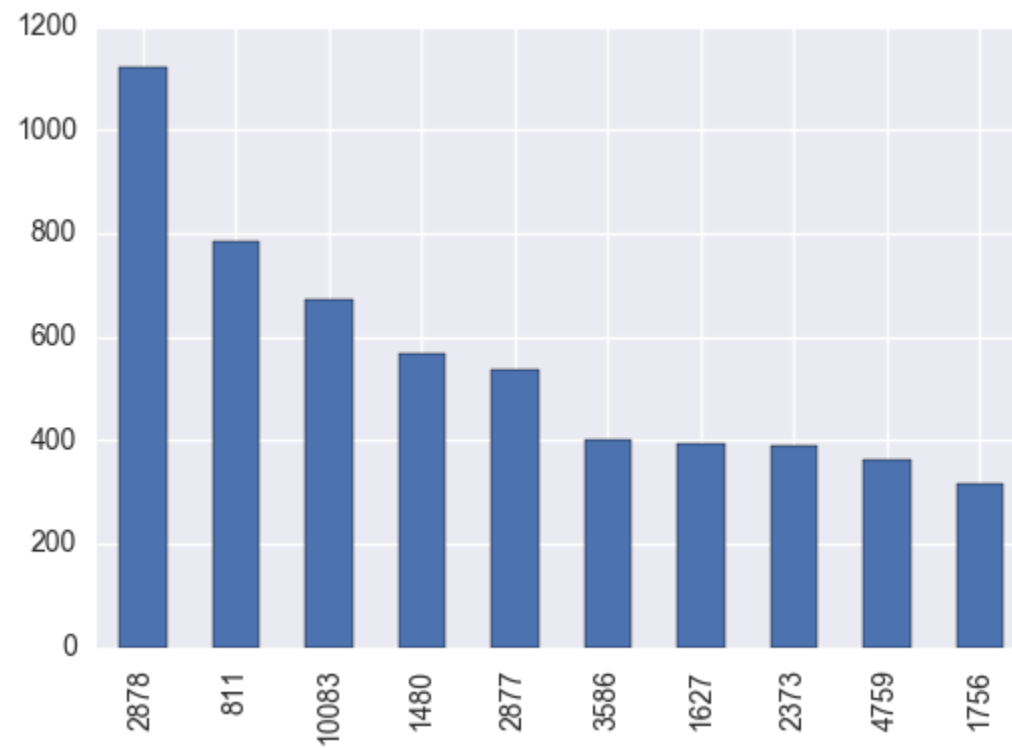
DATA PRE-PROCESSING

- Data is sparse and messy
- Some steps to cleanse and prep the data include:
 - Eliminating claims in other languages
 - NaN values are dropped from the data set
 - Free texts fields are pivoted and concatenated
 - Special characters are removed to improve the vectorizer
 - N/A failure modes are eliminated
 - Long tail of 'one-offs' are removed

DATA EXPLORATION



DATA EXPLORATION (CONT.)



FEATURE SELECTION

- Decided to use features that are required to be inputted
- Added additional features to split models (e.g. vehicle model)

MODELING PROCESS

- Started with all data (all vehicle models)
 - NULL accuracy = 4.9%
- Train test split
- Count vectorizer
 - Stop words
 - Lemmatize
 - Stemming
- Document term matrix
- Naive Bayes model
- Result (accuracy score) = 37.0%

MODELING PROCESS (CONT.)

- Decided to split the data sets by vehicle model
- NULL accuracy = 13.4% for Model S
- Results = 50.2% for Model S
- NULL accuracy = 6.3% for Model X
- Results = 42.2%
- Improvement for both

MODELING PROCESS (CONT.)

- Decided to try Logistic Regression
 - Takes a bit longer but could be worth it
- Results = 60.3% for Model S
- Results = 54.5% for Model X
- Again, improvement for both (with a little performance hit)

SUCCESSSES

- Was able to cleanse the data fairly well
- Was able to beat the NULL accuracy by quite a bit
- It was a fun project

CHALLENGES

- Messy, sparse data
- Some features did not work (n_grams, spell check, stop_words)
- Tons of options to predict ~ 3,500 unique failure modes (and counting)
- Difficult to implement
 - Confidence level for issue tagging
 - New failure modes are created as cars age
 - Not a lot of bandwidth for tagging these issues

POTENTIAL IMPROVEMENTS

- Pipeline (probably still running on my machine at home)
 - Feature union, count vectorizer, tf_idf
- Cross validation for more accurate model evaluation
- Further data cleaning (seems endless)