# Final Project Guidelines Spring 2022

Dr. Joseph P. Yurko (Pitt)

Anastasia Sosnovskikh (PPG, Pitt 2021)

# Final project sponsored by PPG Industries

- Fortune 500 Company
- Global supplier of paints, coatings, and specialty materials
- Largest coatings company in the world by Revenue
- Headquarters in Pittsburgh, PA and operates in 70+ countries around the world
- Please see ppg.com to learn more!

# Data science, analytics, and machine learning, play an important role throughout PPG

For example, machine learning techniques are used to:

- Optimize logistics and supply chain
- Design products
- Improve quality
- Analyze and improve customer experience

# Data science, analytics, and machine learning, play an important role throughout PPG

- Machine learning can also benefit sales and procurement groups and thus directly impact the company’s bottom line!

- An obvious use-case is to forecast sales and purchasing needs, but there are other applications to consider.

# Large corporations like PPG have large Sales groups with many employees

- The Sales representatives and associates directly interface with customers.

- They meet with customers to better understand the customer’s needs, and reasons for purchasing products from the company.

- Sales representatives (Sales Reps) are an integral part of any company!

# Sales Reps are responsible for specific products being sold to specific customers

- Every product has a sales goal or target that the company hopes to achieve over some period of time.
- Some products require a lot of effort by the Sales Reps to achieve the intended goal.
- Effort corresponds to calls, meetings, and other communications with the customer.
- The more communication that is required the more time is required to meet the goals.

# Problem motivation: Company goals

- Ideally, the company wants the Sales Reps to spend the most time on the products that generate the most revenue!
- After all, more effort and thus time means greater cost and thus lower profit per product!
- Also, more effort and thus time with one product means less time is available for other products!

# Problem motivation: Company goals

- Sales departments define goals for each product, and those goals are used to project revenue and profit targets for the company.

- Those goals are based on historical experience, and the Sales Reps opinions about the customers.

- Although these opinions can be based on simple “yes/no” style questions, all Sales Reps create reports detailing their interactions with the customers.

- Those reports are full of text that better describe the customer interaction!

# Problem motivation: How do we help?

- Features are derived from the text within the Sales Reps' reports.

- These features will serve as inputs to machine learning models in order to:

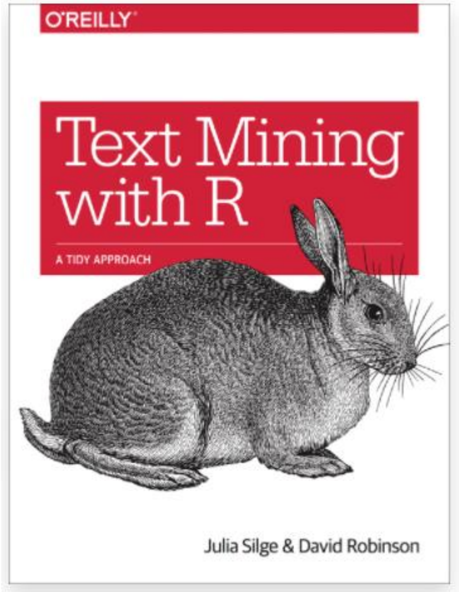
1. Predict the average hours per week Sales Reps spend on a product.
2. Classify if the product achieves the specified sales target.

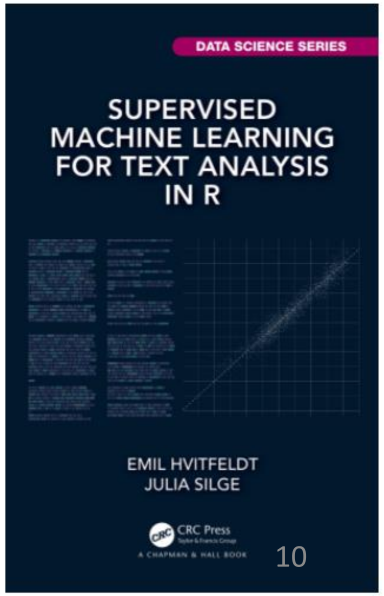
- **Ultimately, the company wants to know which customers are the most difficult to predict in terms of time and achieving sales goals.**

# Text analysis is beyond the scope of this course, and so all text derived features are provided to you

- If you would like to learn more about working with text, I recommend:
  - Text Mining with R: A Tidy Approach by Julia Silge and David Robinson
  - Supervised Learning for Text Analysis in R by Emil Hvitfeldt and Julia Silge

- Both books are freely available online at the sites linked above.
- The tidytext book is a great introduction to text analysis from an exploration and unsupervised viewpoint.
- The SMLTAR book specifically focuses on extracting features from text in support of predictive models.
- **IMPORTANT**: you do **NOT** need to read either of these books for this project. This is just if you are interested to learn more.





# Text derived features usually represent word counts or frequencies within a text

- Because of that, it is quite common to have thousands of features associated with a text analysis project.

- Each feature essentially represents the frequency of a word.

- The more unique words used in the text, the greater the number of features will be required!

- However, I wanted to start simpler with this problem…

# I am curious if the text **sentiment** is related with the response behavior!

- Sentiment represents that the words, phrases, sentences, and/or paragraphs of text are “positive” or “negative”.

- Using more “positive” words to describe something in text is associated with greater “positive” sentiment.
  - For example: I really like pizza! I would love to eat pizza everyday!

- Conversely, using more “negative” words to describe something is associated with greater “negative” sentiment.
  - For example: Traffic is awful! I hate sitting in traffic!

# Why are we using sentiment derived features?

- My intuition is, greater positive sentiment will be associated with greater probability of achieving the sales targets!

- Also, I am curious if greater positive sentiment **and** greater negative sentiment are associated with spending more time on the product.
    - Represents we spend lots of time on things we know are going well, and we spend lots of time on things we know are not going well. Are we therefore not spending enough time with things that seem “ok”?

- **Secondary project goal: I want you to find out if I am right!**

# Data description

- You are given a CSV file containing the features and outputs.
  - You must download the data from the Canvas page.
  - All sentiment derived features have been provided to you.
  - You do **NOT** need to perform sentiment analysis.

- One row corresponds to a product sold to a customer.

- Variables (columns) are divided into 3 categories:
  - Identifiers
  - Sentiment derived features
  - Outputs

# Data description: Identifiers

- There are 3 identifiers in the data.

- `rowid`: uniquely defines each row in the data. Do **NOT** use this column in any model. It helps with book-keeping.

- `region`: the global region for the customer purchasing the product. The region label has been anonymized and is categorical. You will consider this variable in your models (more details on this later).

- `customer`: the company purchasing the product. The customer label has been anonymized and is categorical. You will consider this variable in your models (more details on this later).

# Data description: Sentiment derived features

- There are multiple ways of conducting sentiment analysis.
- Several different approaches have been applied to each report, generating multiple types of sentiment derived features.

- There are many different reports associated with each product.
- The sentiment derived features associated with each report have been summarized appropriately for each product.

# Data description: Sentiment derived features

- Fives types of sentiment derived features have been summarized for each product.

- Three types correspond to different Lexicons (dictionaries) used to determine if a word has “positive” or “negative” sentiment.
    - If you are interested, the Bing, NRC, and AFINN lexicons were used. More information is available in Chapter 2 of the tidytext book, but you are NOT required to read about this material.

- One of the five types are associated with word counts, including the impact of “stop words” on the text.

- The last type corresponds to features produced from the sentimentr package which has more powerful Natural Language Processing (NLP) capabilities than simple lexicons.

# Data description: Sentiment derived features

The five types are distinguished based on the column (variable) name, which follows a particular naming convention:

- Bing lexicon derived features have names starting with `xb_`
- NRC lexicon derived features have names starting with `xn_`
- AFINN lexicon derived features have names starting with `xa_`
- Word count derived features have names starting with `xw_`
- `sentimentr` derived features have names starting with `xs_`

- All five types of sentiment derived features are continuous variables.

# Data description: Outputs

- There are 2 outputs in this project, one is continuous while the other is categorical.

- Continuous output: `response`
  - Average hours per week associated with a product sold to a customer.

- Categorical output: `outcome`
  - Binary variable with values `event` and `non_event`
  - The `event` value represents the product did **NOT** achieve its sale goal.

# Output considerations: `response`

- Although the problem is formulated as predicting the hours per week, you should NOT predict the response directly.

- The `response` is lower bounded at 0, after all no one can work NEGATIVE hours!

- You should instead transform `response` by applying the (natural) log-transformation!

- **Your regression models should be trained to predict the log-transformed output!**

# Output considerations: `outcome`

- Pay close attention to the empirical proportion of the `event` value when interpreting the Accuracy and other classification performance metrics!!!!!

# The project therefore consists of regression and classification tasks

**Regression**

- Predict the log-transformed `response`, as a function of the categorical inputs (`region` and `customer`) and continuous sentiment derived features.

**Classification**

- Classify if the binary `outcome` is `event` as a function of the categorical inputs (`region` and `customer`) and continuous sentiment derived features.

# The project is open ended

- No template is provided.

- An Rmarkdown is provided to give an example of reading in data.
    - It also shows how to save a model object and load that model in again.

- Specific requirements are listed next, and those requirements can help guide you through the predictive modeling application.

# Project consists of 4 main areas

- Part i: Exploration
  - It is always important to explore and study your data before starting any modeling exercise.
- Part ii: Regression
  - Fit non-Bayesian and Bayesian linear models.
  - Train, tune, and assess performance of simple and complex models with resampling.
- Part iii: Classification
  - Fit non-Bayesian and Bayesian generalized linear models.
  - Train, tune, and assess performance of simple and complex models with resampling.
- Part iv: Interpretation
  - Identify the best models, most important features, and the hardest to predict customers for the regression and classification tasks.

# Part i: Exploration

- Visualize the distributions of variables in the data set.
    - Counts for categorical variables.
    - Distributions for continuous variables. Are the distributions Gaussian like?
- Consider conditioning (grouping or “breaking up”) the continuous variables based on the categorical variables.
    - Are there differences in continuous variable distributions and continuous variable summary statistics based on region or customer?
    - Are there differences in continuous variable distributions and continuous variable summary statistics based on the binary outcome?
- Visualize the relationships between the continuous inputs, are they correlated?
- Visualize the relationships between the continuous outputs (`response` and the log-transformed `response`) with respect to the continuous inputs.
    - Can you identify any clear trends? Do the trends depend on the categorical inputs?
- How can you visualize the behavior of the binary outcome with respect to the continuous inputs?

# Part ii: Regression – iiA)

- Before using advanced methods, you need to develop a baseline understanding of the log-transformed `response` as a function of the inputs using linear modeling techniques.

- Use `lm()` to fit the following linear models:
  - Categorical variables only – linear additive
  - Continuous variables only – linear additive
  - All categorical and continuous variables – linear additive
  - Interact `region` with continuous inputs, do not include `customer`
  - Interact `customer` with continuous inputs, do not include `region`
  - All pairwise interactions of continuous inputs, do not include categorical inputs
  - 3 models with basis functions of your choice
    - Can include interactions of the basis functions with other basis functions and with the categorical inputs!

# Part ii: Regression – iiA)

- You must therefore train 9 different models!
- Which of the 9 models is the best? What performance metric did you use to make your selection?
- Visualize the coefficient summaries for your top 3 models.
- How do the coefficient summaries compare between the top 3?
- Which inputs seem important?

# Part ii: Regression – iiB)

- You have explored the relationships; next you must consider the UNCERTAINTY on the residual error through Bayesian modeling techniques!
- Fit 2 Bayesian linear models – one must be the best model from iiA) and the second must be another model you fit in iiA).
  - State why you chose the second model.
- You may use the Laplace Approximation approach we used in lecture and the homework assignments.
- Alternatively, you may use `rstanarm`'s `stan_lm()` or `stan_glm()` function to fit full Bayesian linear models with syntax like R's `lm()`.
  - How to Use the rstanarm Package (r-project.org)
  - Estimating Regularized Linear Models with rstanarm (r-project.org)

# Part ii: Regression – iiB)

- After fitting the 2 models, you must identify the best model. Which performance metric did you use to make your selection?

- Visualize the regression coefficient posterior summary statistics for your best model.

- For your best model: Study the posterior uncertainty in the noise (residual error), $\sigma$. How does the `lm()` maximum likelihood estimate (MLE) on $\sigma$ relate to the posterior uncertainty on $\sigma$?
  - Do you feel the posterior is precise or are we quite uncertain about $\sigma$?

# Part ii: Regression – iiC)

- You must make predictions with your 2 selected linear models in order to visualize the trends of the log-transformed `response` with respect to the inputs.

- You may use non-Bayesian or Bayesian models for the predictions.

- You must decide which inputs/features you wish to visualize the trends with respect to.

- You must visualize your predictive trends using the following style:
  - The primary input should be used as the x-aesthetic in a graphic.
  - The secondary input should be used as a facet variable – it is recommended to use 4 to 6 unique values if your secondary input is a continuous variable.
  - You must decide what values to use for the remaining inputs.

- You MUST include the predictive mean trend, the confidence interval, and the prediction interval whether you use non-Bayesian or Bayesian models.

- You MUST state if the predictive trends are consistent between the 2 selected linear models.

# Part ii: Regression – iiD)

- You must train, evaluate, tune, and compare simple to complex models via resampling.
  - You may use either `caret` or `tidymodels` to handle the preprocessing, training, testing, and evaluation.
- You must train and tune the following models:
  - Linear models:
    - All categorical and continuous inputs - linear additive features
    - All pairwise interactions of continuous inputs, include additive categorical features
    - The 2 models selected from iiA)
  - Regularized regression with Elastic net
    - All pairwise interactions of continuous inputs, include additive categorical features
    - The more complex of the 2 models selected from iiA)
  - Neural network
  - Random forest
  - Gradient boosted tree
  - 2 methods of your choice that we did not explicitly discuss in lecture

You must use all categorical and continuous inputs with non-linear methods

# Part ii: Regression – iiD)

- You must decide the resampling scheme, what kind of preprocessing options you should consider, and the performance metric you will focus on.

- You must identify the best model.

# Part iii: Classification – iiiA)

- Before using advanced methods, you need to develop a baseline understanding of the event probability as a function of the inputs using generalized linear modeling techniques.

- Use `glm()` to fit the following linear models:
  - Categorical variables only – linear additive
  - Continuous variables only – linear additive
  - All categorical and continuous variables – linear additive
  - Interact `region` with continuous inputs, do not include `customer`
  - Interact `customer` with continuous inputs, do not include `region`
  - All pairwise interactions of continuous inputs, do not include categorical inputs
  - 3 models with basis functions of your choice
    - Can include interactions of the basis functions with other basis functions and with the categorical inputs!

# Part iii: Classification – iiiA)

- You must therefore train 9 different models!
- These models are consistent with the regression portion. Did you experience any issues or warnings while fitting the generalized linear models?
- Which of the 9 models is the best? What performance metric did you use to make your selection?
- Visualize the coefficient summaries for your top 3 models.
- How do the coefficient summaries compare between the top 3?
- Which inputs seem important?

# Part iii: Classification – iiiB)

- Next, you need to consider uncertainty via Bayesian methods!
- Fit 2 Bayesian generalized linear models – one must be the best model from iiiA) and the second must be another model you fit in iiiA).
  - State why you chose the second model.
- You may use the Laplace Approximation approach we used in lecture and the homework assignments.
  - Alternatively, you may use `rstanarm`'s `stan_glm()` function to fit full Bayesian linear models with syntax like R's `glm()`.
- After fitting the 2 models, you must identify the best model.
  - Which performance metric did you use to make your selection?
- Visualize the regression coefficient posterior summary statistics for your best model.

# Part iii: Classification – iiiC)

- You must make predictions with your 2 selected generalized linear models in order to visualize the trends of the event probability with respect to the inputs.

- You may use non-Bayesian or Bayesian models for the predictions.

- You must decide which inputs/features you wish to visualize the trends with respect to.

- You must visualize your predictive trends using the following style:
  - The primary input should be used as the x-aesthetic in a graphic.
  - The secondary input should be used as a facet variable – it is recommended to use 4 to 6 unique values if your secondary input is a continuous variable.
  - You must decide what values to use for the remaining inputs.

- You MUST include the predicted mean event probability and the confidence interval whether you use non-Bayesian or Bayesian models.

- You MUST state if the predictive trends are consistent between the 2 selected generalized linear models.

# Part iii: Classification – iiiD)

- You must train, evaluate, tune, and compare simple to complex models via resampling.
  - You may use either `caret` or `tidymodels` to handle the preprocessing, training, testing, and evaluation.
- You must train and tune the following models:
  - Generalized linear models:
    - All categorical and continuous inputs - linear additive features
    - All pairwise interactions of continuous inputs, include additive categorical features
    - The 2 models selected from iiiA)
  - Regularized logistic regression with Elastic net
    - All pairwise interactions of continuous inputs, include additive categorical features
    - The more complex of the 2 models selected from iiA)
  - Neural network
  - Random forest
  - Gradient boosted tree
  - 2 methods of your choice that we did not explicitly discuss in lecture

You must use all categorical and continuous inputs with non-linear methods

# Part iii: Classification – iiiD)

- You must decide the resampling scheme, what kind of preprocessing options you should consider.

- You must identify the best model.

- Which model is the best if you are interested in maximizing Accuracy compared to maximizing the Area Under the ROC Curve (ROC AUC)?

# Part iv: Interpretation – ivA)

- With the model training completed, you can now answer meaningful questions associated with the data!

- You must identify the best regression model and the best classification model.

- Identify the most important variables associated with your best performing models.

- Are the most important variables similar for the regression and classification tasks?
    - Does one of the sentiment derived feature types “dominate” the most important variables?
    - Does one of the sentiment derived feature types appear to be not helpful at all?

- Based on your modeling results, do you feel these sentiment derived features are helpful at predicting the outputs?
    - Essentially, do you feel the model results are “good” and was I right about the relationship between report sentiment and the outputs?

# Part iv: Interpretation – ivB)

- You must identify which customers appear to be the hardest to predict in the regression and classification tasks.

- Base your conclusions on the best performing regression and classification models.

- You should base your conclusions on the resampled hold-out sets and **NOT** on the training set!
  - Thus, save your hold-out set predictions!

# Part iv: Interpretation – ivC)

- You must visualize the trends associated with the hardest to predict customer with respect to the most important sentiment derived feature.

- You are free to select whether you wish to visualize the regression model trends (log-transformed `response`) or the classification model trends (event probability).
  - Predictions should be made using the best performing model.

- You must visualize your predictive trends using the following style:
  - The primary input should be used as the x-aesthetic in a graphic.
  - The secondary input should be used as a facet variable – it is recommended to use 4 to 6 unique values if your secondary input is a continuous variable.
  - You must decide what values to use for the remaining inputs.

- What conclusions can you draw from the predictive trends?

# Two additional methods

- You may use the same two methods for both the regression and classification portions of the project.
  - If, however, you select a method that cannot be used for both regression and classification, then you will need to select an additional method.

- Potential methods to consider:
  - Support Vector Machines (SVM) – classification and regression
  - Naïve Bayes – classification
  - Generalized Additive Models (GAM) – classification and regression
  - Multivariate Additive Regression Splines (MARS) – classification and regression
  - Partial Least Squares (PLS) – classification and regression
  - Deep Neural Network – classification and regression
  - K-nearest neighbors – classification and regression
  - Stacked models

- Please see Ch 6 in the caret documentation for a complete list of available methods in `caret`.
- Please see the tidymodels parsnip list of available models for models available in `tidymodels`.

# Interpretation and visualization help

- Chapter 16 in HOML provides useful discussion on interpretable machine learning.

- Provides code examples for visualizing model behavior and interpreting the graphics.

# Homework assignments include examples working with caret

- You may use `caret` to perform all preprocessing, resampling, tuning, and evaluation for the project.

- However, you may use `tidymodels` instead of `caret`.

- `tidymodels` provides modeling aligned with the philosophy of the `tidyverse`, created by the developers of `caret`.

- If you are interested to learn `tidymodels`, please see the homepage, and try some of the "Get Started" tutorials.

# Applied machine learning examples available on Canvas provide both `caret` and `tidymodels` examples

- Week 01 – Airfoil example problem
  - Example EDA, linear models, and regression models with `caret`

- Week 02 and Week 03 – examples
  - Regression application with `tidymodels` – concrete data
  - Binary classification application with `tidymodels` – ionosphere data

# Test set predictions

- A test set of input values will be provided in April.

- You must predict the continuous response and the event probability using this test set.

- You will upload your predictions to a website. The website will provide performance metrics associated with your predictions.

- More to come on this later!

# BONUS points – report to PPG

- Create a PowerPoint presentation highlighting the major results, findings, and conclusions from your work.

- **You may earn up to 10 BONUS points.**

- Background/motivation material is not needed for the presentation.

- Presentation is open ended, but some recommendations to include:
  - Interesting visualizations from the EDA portion of the project.
  - Visualizations comparing model performance and selection of the best model.
  - Visualizations showing variable importances and/or coefficient summaries.
  - Visualizations showing model predictions with text interpreting the trends.

# BONUS points – synthetic data

- In lecture, we have discussed the importance and usefulness of synthetic (fake) data in a complete Bayesian model workflow.

- **You may earn up to 12 BONUS points if you create your own synthetic data and demonstrate the ability to recover the model parameters that generated the data.**

- You may consider the regression problem OR the classification problem, but you must use Bayesian modeling techniques.

- Regardless you must use the following to earn the maximum bonus:
  - You must consider 1 categorical variable with 4 levels (unique values)
  - You must consider 3 continuous variables
  - You must specify the true functional (basis) relationship between the linear predictor and the inputs. You must specify the true parameter values.
  - You must generate small, medium, and large sample size data sets.
  - You must fit your model, assuming the correct functional (basis) relationship for the small, medium, and large sample sizes.
  - How well are you able to recover the true parameter values given the three training sample sizes?

# BONUS points – imbalanced data

- **IMPORTANT**: As you SHOULD see in your data exploration, the binary outcome is imbalanced.
- I did not give you the real data set for this problem, because the real problem is even more imbalanced and thus more challenging!
- I did something you should **NEVER** do, I "cherry picked" the data.
- The "cherry picked" data removed some of the sampling imbalance issues including:
  - Low frequency categorical levels (unique values) – the customer variable includes an Other category which lumped together low frequency levels.
  - No columns have zero or near zero variance.
  - The event has an empirical proportion greater than 15%.
- **This was only "ok" because we are using the data for educational purposes.**

# BONUS points – imbalanced data

- However, if you would like learn how to deal with imbalanced data, you may work with the BONUS data set provided in Canvas.

- You only need to focus on the classification task with the bonus data.

- **You may earn up to 15 BONUS points** if you fit 3 classification models which account for:
  - Low frequency categorical input classes via lumping – you may test if this helps or not!
  - Near zero variance features.
  - Output class imbalance via SUBSAMPLING methods.

- Subsampling is a technique which artificially balances the output classes. This artificial sampling must be accounted for and should NEVER be done manually.
  - Please see Ch 11 from the caret documentation if you would like to learn about subsampling in caret.
  - Please see the tidymodels Learn page on subsampling for dealing with class imbalance in tidymodels.
  - Please see this Julia Silge blogpost for another tidymodels example with class imbalance.

# Project submission

- You must submit the Rmarkdown source .Rmd files and the associated rendered HTML documents.

- It is recommended that you create separate Rmarkdowns for the different portions of the project. This way you can work in a modular fashion and will not have a single enormous file.

- You must upload the separate HTML files. Do NOT zip files! Zipping files together will result in a 15% penalty!!!!!!!!!!!
  - You may only zip the saved model objects.

- **Project must be submitted not later than Wednesday April 27, 2022 at 11PM EST (Pittsburgh, PA local time).**