Preface

- *Salary*: continuous outcome ($, thousands)
- *Company*: 63-level nominal feature
- *Industry*: 7-level nominal feature
- *Degree Major*: 9-level nominal feature
- *Job Rank*: 8-level ordinal feature
- *Education Level*: 5-level ordinal feature
- *City Proximity*: continuous feature (miles)
- *Experience*: continuous feature (years)

0 Objectives and Strategy

1. Categorize jobs into robust, interpretable groups with distinctive compensation signatures.
2. Characterize the underlying relationship between features and outcomes within these groups.
3. Use these understandings of compensation dynamics to provide job compensation predictions.

To accomplish these objectives we parsimoniously partition the job space across *Job Rank (R)*, *Industry (I)*, *Degree Major (D)* and *Education Level (E)* in order to optimally associate *Experience (X)* and *City Proximity (M)* with *Salary (Y)*; and subsequently identify *Company (C_j)* compensation tendencies. I.e., we estimate

$$Y \sim \big(f(M, X) \mid g(R, I, D, E)\big) + C_j, \text{ where}$$

- $f(X, M)$ is the predictive association of experience and city proximity with salary outcomes,
- the notation " $\mid g(R, I, D, E)$" indicates model fitting within the identified data partitioning,
- and C_j is company effect (assumed to be constant across the levels of features and partitions).

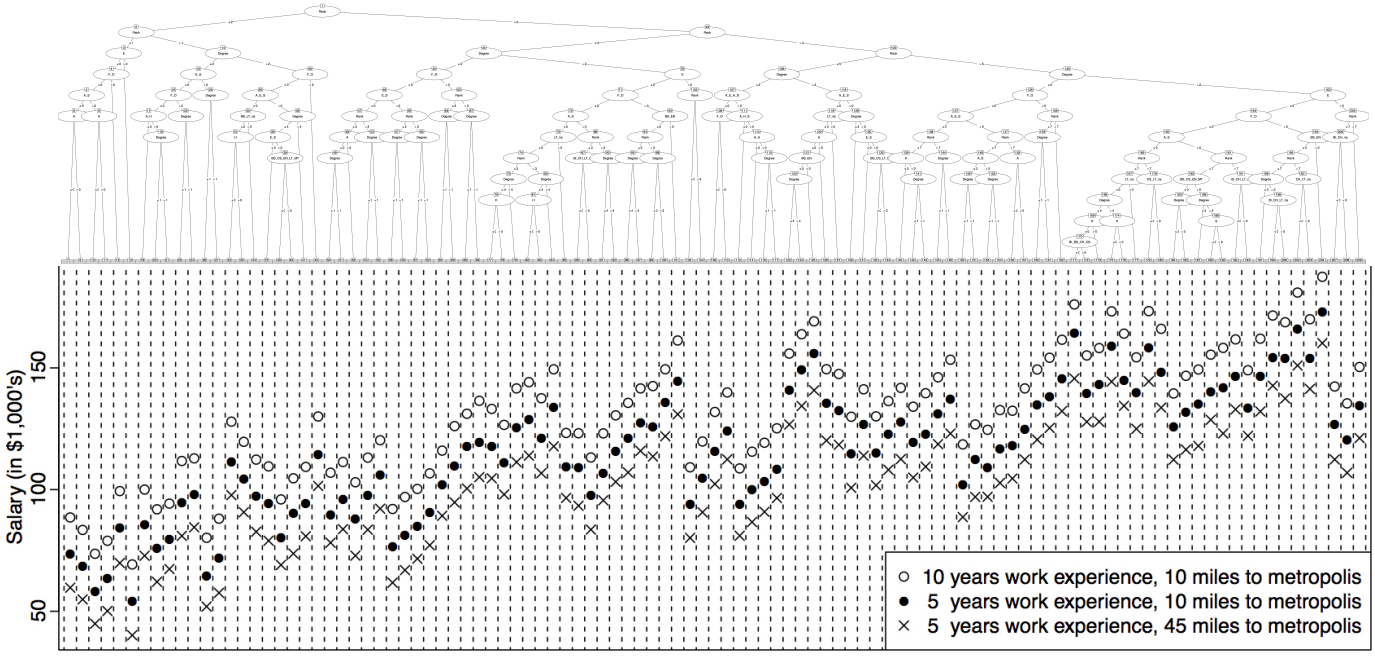
0.5 Implementation

- The `mob` function in `R`'s `party` package implements model-based recursive partitioning (MBRP).
  - \+ Separate idiosyncratic model fits within partitions provides *model by partition* interactions.
  - − Sharing parameters across partitions provides information sharing for estimation; but, model-based stratification adjustment seems unnecessarily restrictive with so much data.
- The `mob` function does not provide composite (binary tree) spliting on categorical variables, e.g., splitting degree between 'none' and ''high school', versus 'college', 'masters', 'doctorate'.
  - \+ Such splits for nominal I and D are represented using $\frac{9^2-2}{2}$ and $\frac{7^2-2}{2}$ indicator variables.
  - \+ Such splits for ordinal R and E are achieved directly via real-valued feature representation.
  - − Linear models require explicit and *post-hoc* feature engineering for composite stratification.
- C_j are evaluated diagnostically against residuals to identify partition-agnostic company trends.
- X, M & Y are transformed to improve the assumptions underlying partition subdivision testing.

1 Job Partitioning

Model-based recursive partitioning (MBRP) builds a decision tree over data partitioning features which greedily optimizes a model fit criterion within the leaves at each branch split. This simultaneously allows us to (a) learn about partitions of data that can be effectively modeled together and (b) subsequently leverage that synchronicity in model fitting. The partitioning parsimoniously concludes once the model criterion can no longer be improved within further subdivisions of the data. We use this methodology to (1) automatically identify higher order interactions of job features (e.g., *industry, discipline, education requirements* and *job rank*) that differentiate job salaries, (2) utilize these associations for improved prediction, and (3) explore the partitioning itself for informative clustering. **Figure 1** shows the partitioning and representative salaries associated with each level of the partitioning for $p = 101$ partitions MBRP model (**Table 1**); and **Figure 2** shows in the decision tree which recursively partitions the data for the same model in closer detail.

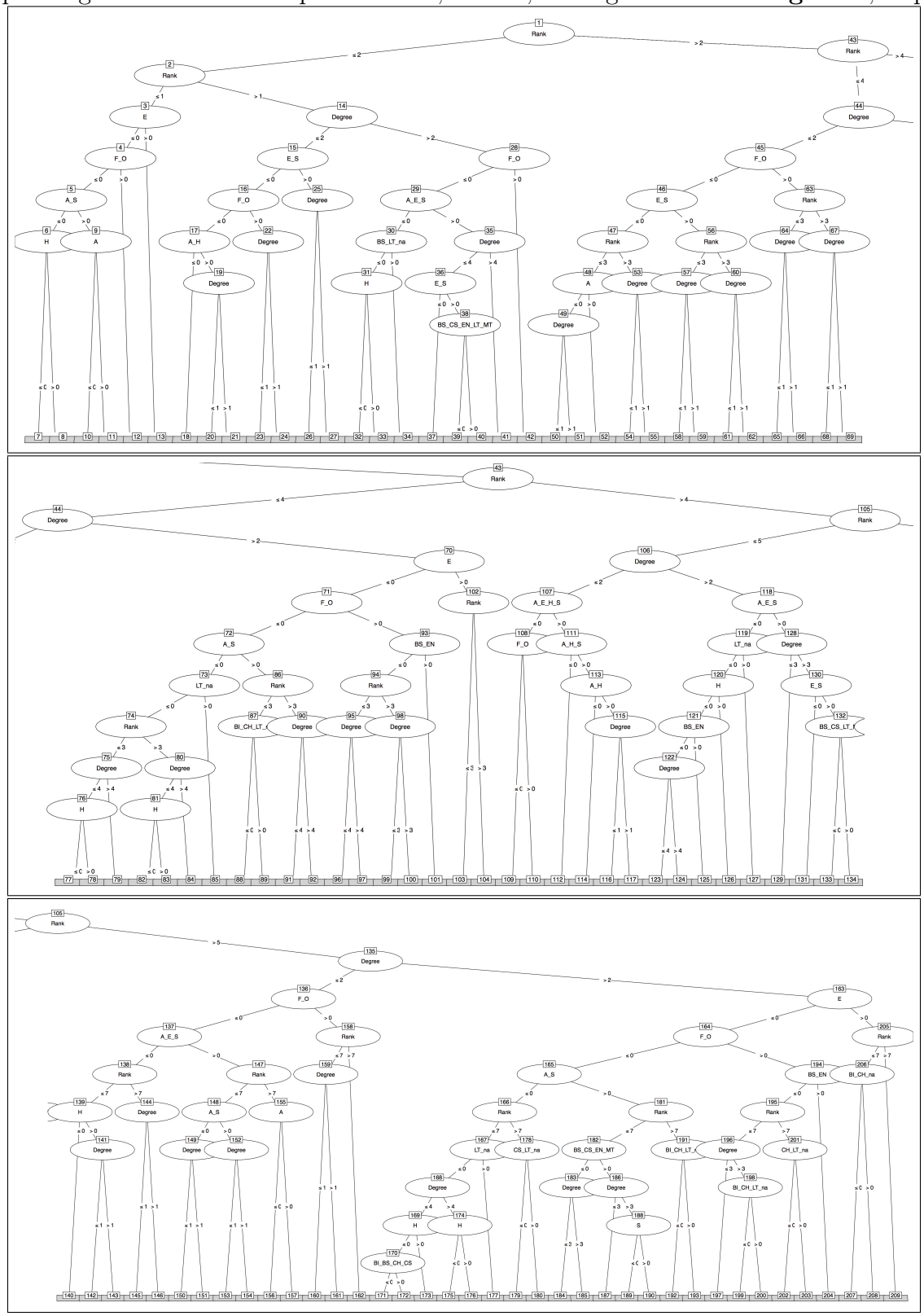
Figure 1: The top axis shows the (greedy) optimal recursive partitioning of the data via a decision tree while the plot itself shows representative compensation levels corresponding to the job partitions.



Key 1: The levels of the partitioning variables in **Figure 1** and **Figure 2** are denoted as follows.

Major		Industry		Rank		Degree	
BI	BIOLOGY	A	AUTO	1	JANITOR	1	NONE
BS	BUSINESS	E	EDUCATION	2	JUNIOR	2	HIGH SCHOOL
CH	CHEMISTRY	F	FINANCE	3	SENIOR	3	BACHELORS
CS	COMPSCI	H	HEALTH	4	MANAGER	4	MASTERS
EN	ENGINEERING	O	OIL	5	VICE PRESIDENT	5	DOCTORAL
LT	LITERATURE	S	SERVICE	6	CFO		
MT	MATH	W	WEB	7	CTO		
na	NONE			8	CEO		
PH	PHYSICS						

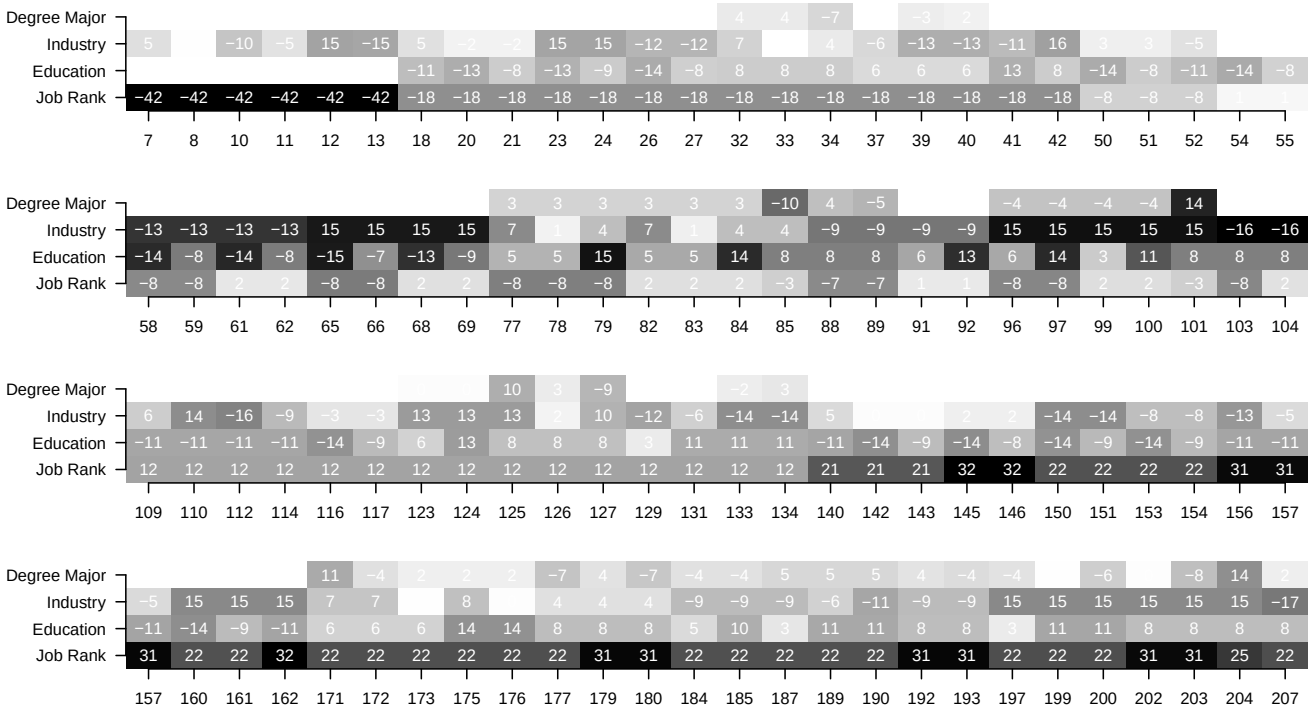
Figure 2: Data is (greedily) optimally partitioned via the decision tree. The relative selection of partitioning features (e.g., *industry, discipline, education* and *job rank*) for branching indicates how broadly these features influence compensation trends, and the resulting compositions of feature levels within leaf nodes indicate compensation trends that generalize across distinct jobs. The top, middle, and left panels give detailed close-ups of the left, middle, and right section of **Figure 1**, respectively.



2 Factors Driving Salary

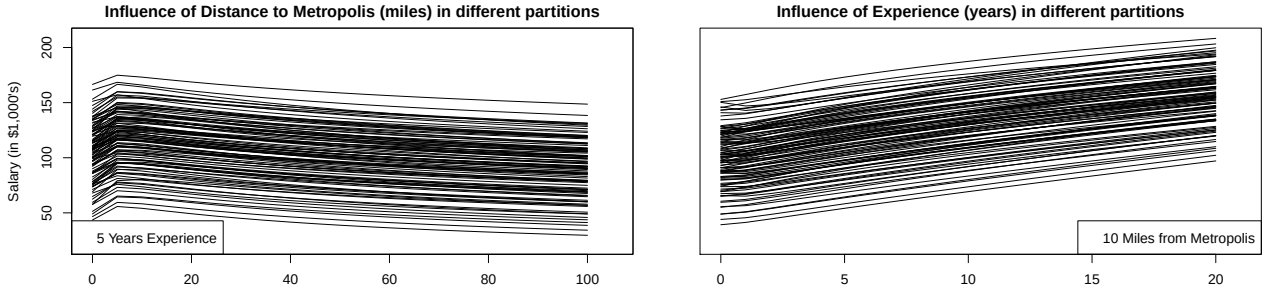
The effect of the partitions of the MBRP model can be interpreted and understood in many ways. As indicated in **Figure 1** and **Figure 2**, the MBRP model allows us to examine (1) compensation within partitions, (2) the relative influences of features driving partitioning, and (3) the composition of the partition. A deeper examination of the relative impact of each feature within each partition is also available as follows. As data traverses down the decision tree, its predictions fluctuate as a function of feature branching. These fluctuations characterize the path of differentiation away from the population average and are sequentially attributable to specific values of the responsible features. The cumulative impact of each feature can be tabulated to show the absolute and relative contributions of features to predicted values. In this manner, **Figure 2** shows the change in prediction away from the population average attributable to *degree major, industry, education level* and *job rank* within each partition. All analyses indicate the preeminence of *job rank* in driving compensation outcomes. *Industry* and *education level* also drive compensation outcomes at times, while *degree major* appears to have only a subtle role in influencing compensation outcomes. In addition to these general trends, certain synergistic and compensatory interactions appear to exist for various combinations of feature levels. Note that a deeper comparison between the levels of each factor is possible using the analyses presented here; however, we do not pursue such evaluations at this time.

Figure 3: The cumulative influence of each feature on compensation outcomes (rounded to the nearest $1,000) is shown for each partition (indicated on the x-axis). Darker squares indicate more influence.



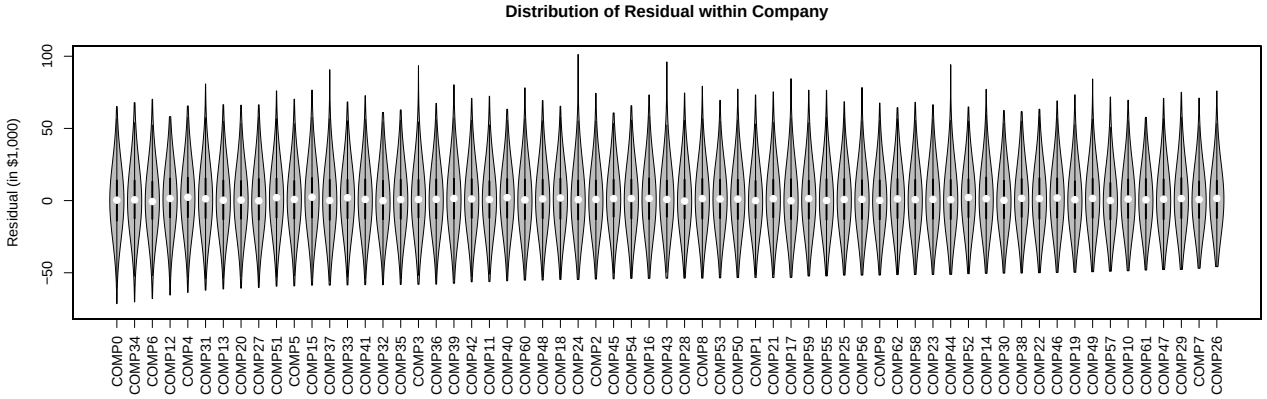
In addition to recursively partitioning the data, MBRP models associations within the identified partitions. As seen in **Figure 4**, the influence of *job experience* and *distance to a metropolis* on compensation do not vary across the partition. Thus, (a) these effects could thus be modeled parsimoniously across strata, and (b) the recursive partitioning is optimizing the additive effects of the partition. I.e., MBRP is sequentially partitioning jobs having the most similar compensations after adjusting for *job experience* and *metropolis proximity*.

Figure 4: *Job experience* and *proximity to a metropolis* do not differentially affect compensation across the identified job partition. Note the interesting "0" artifact for *years of job experience* and *miles to nearest metropolis* – further examination to better understand this phenomenon is required.



After the model was fit and predictions were made the residuals were examined to see if any specific companies showed consistent above or below market compensation tendencies. As can be seen in **Figure 5**, no definitively noticable company effect on compensation was observed.

Figure 5: Model residuals do not correlate with companies – compensation strategies appear uniform.



3 Salary Prediction

Exploratory *linear regression* and *random forest* analyses indicated (a) the presence of higher order interactions and (b) the difficulties in identifying robust and interpretable associations in the presence of overfitting. Identifying robust relationships in a linear regression context would entail a lengthy model building exercise iterating between coefficient analysis and post-hoc feature engineering. Regularization could play a role in the linear regression context as well, but it is *definitively* required for an ensemble-based approach as overfitting is *unquestionably* a risk with overly flexible models in data-rich contexts such as ours. While regularization in an ensemble context can achieve efficient predictive performance *out of sample*, when *in sample* performance still out performs *out of sample* performance model interpretations are not generalizable; and unfortunately, in our case regularization did not provide a supple enough tool to reconcile *in sample* and *out of sample* performance.

MBRP provides an automated way to identify generalizable associations in data while still optimizing predictive performance. It achieves this by recursively partitioning data into groups *specifically identified as having* statistical evidence of generalizable associations. The recursive partitioning is a natural approach to identifying meaningful "real-world interpretable" data partitions that can each be idiosyncratically (i.e., optimally) modeled; however, MBRP is a greedy algorithm and so it

does not guarantee a global optimum in terms of predictive performance. Thus, other non-recursive partitionings of the data could produce improved prediction performance. Likewise, algorithms targeting predictive performance alone (rather than optimal partitioning) could conceivably improve predictive performance over MBRP, although the improved performance might be expected to come at the cost of model interpretation generalizability (i.e., discrepancy between *in* and *out of sample* error). Due to computational and hardware limitations, our final predictive model was based on a random sample of only 40% of the data; nonetheless, we demonstrably improve upon traditional modeling approaches that might reasonably be proposed for the analysis at hand; and further, as can be seen from **Table 1**, increasing the amount of data used to fit the model both improves predictive power (i.e., *out of sample* RMSE continually decreases) as well as model interpretation generalizability (i.e., the *difference* between *out of sample* and *in sample* RMSE continually decreases) – so we expect it would be possible to further improve the predictive performance of the MBRP model. Predictions for a completely held out validation sample are available in `validation_predictions.csv`

Table 1: Model-based recursive partitioning (MBRP) improves over standard statistical and machine learning methodologies in terms of *out of sample* prediction *and interpretability* due to its more realistic and natural modeling approach. The number of partitions utilized by MBRP is given as p.

	Training RMSE	Testing RMSE	Parameterization
Linear Model	19.55212	19.60710	No interactions
Random Forest	7.78771	20.68747	$T = 100$ trees, unrestricted trees
MBRP $p = 105$	19.49403	19.7335	Min. leaf size $l_{min} = 500$; 10% data
MBRP $p = 144$	19.27032	19.66269	Min. leaf size $l_{min} = 150$; 10% data
MBRP $p = 226$	19.17442	19.35616	Min. leaf size $l_{min} = 50$; 20% data
MBRP $p = 317$	19.13249	19.24042	Min. leaf size $l_{min} = 50$; 40% data

Notes

1. Summary
    - MBRP is a decision tree that chooses a feature (and a level therein) to split upon based on maximizing a modeling criterion – a form of information gain – in the nascent nodes.
2. Process
    - I bypassed regression as requiring too much feature engineering and model selection effort.
    - I discarded random forests as tree regularization was not providing a supple enough tool to go beyond predictive performance and additionally provide robust model interpretability.
    - Other model regularization approaches, e.g., *Ridge* or *Lasso*, similarly felt insufficient.
    - I finally settled on MBRP as the modeling approach best suited to the task at hand, and I was very fortunate to have access to the `party` package in `R` to accomplish my analysis.
    - The primary benefit of MBRP is that – with proper feature initialization – it automatically detects robust higher order interactions – synergistic/compensatory relationships – present in the data (as opposed to manually identifying them *as well as* features capturing them).
    - Now that MBRP has identified the higher order interactions in the data, it would be worth exploring if a more parsimonious regression methodology could make better use of them. It might also be interesting to see if a *stepwise selection* or *Lasso* procedure can identify a model with competitive performance from a collection of candidate higher order terms.

3. Data Cleaning

    - I examined marginal distributions and feature correlations and found out that data is very balanced. I did not get around to performing the same analysis for the holdout data set.
    - I identified (and removed) five 0-valued outcomes via residual and leverage diagnostics. I didn't identify further necessary "data cleaning" tasks – but I didn't focus on this either.
    - I encoded categorical variables as indicators of splits or ordinal (continuous) variables in order to allow for a MBRP process that could naturally partition categorical features.
    - I added 1 and log-transformed all continuous valued features in order to better satisfy the linear regression assumptions utilized by the MBRP as a partition acceptance criterion.

4. Features

    - The key to the success of the MBRP in our context is the initial feature engineering which allows the recursive partition to be done in a natural and interpretable manner. This feature engineering is somewhat subtle, but is designed to provide an indicator specifying a natural division of a categorical variable. Each such possible division is then encoded as an indicator and all indicators are included in the model, thereby allowing the MBRP to choose a natural division of a categorical variable that optimizes its modeling criterion.
    - The continuous features were used as part of the modeling rather than in the partitioning.
    - The company indicators had appeared uninformative during initial analyses and were thus not used in the MBRP; instead, they were diagnostically assessed against the residuals.
    - Because of being used in the modeling (as opposed to partitioning), the continuous variables relationship with the salary outcomes was idiosyncratic in every partition. Diagnostic assessment of the company indicators against the residuals allowed evaluation of global (partition-ambivalent) rather than idiosyncratic (partition-specific) company effects.
    - Idiosyncratic continuous features were deemed admissible because they exhibited robust (partition-adjusted) associations with the salary outcomes that were sufficiently estimable.

5. Software

    - Initial EDA was done in `jupyter notebook` with `python` using `pandas`, `numpy`, `matplotlib.pyplot`, `seaborn`, `statsmodels.regression`'s `linear_model.OLS` and `stats.outliers_influence.OLSInfluence/variance_inflation_factor`, and `sklearn`'s `ensemble.RandomForestRegressor` and `cross_validation.cross_val_score`.
    - After identifying the usual "sufficient model complexity versus overfitting" challenge, I transitioned to a traditional ML regularization toolkit, complementing the `ensemble.RandomForestRegressor` and `cross_validation.cross_val_score` tools with `sklearn`'s `preprocessing.scale`, `grid_search.GridSearchCV`, `linear_model.LinearRegression` and `linear_model.Ridge` capabilities.
    - Upon feeling somewhat unenthusiastic about the process and prospects of a regularization exercise, I transitioned to a more modeling oriented framework and hit upon the idea of a "regression tree" – by which I actually mean MBRP – to automatically detect higher order interactions present in the data. Fortunately for me such a methodology is available via the `party` package in `R`, so I migrated into a generic `R/emacs` environment for the remainder of my analysis.

6. Findings

    - *Job Rank*, and *Job Experience (years)* had the strongest associations with salary outcomes.
    - *Education Level*, *Industry*, and *Metropolis Proximity (miles)* also had solid associations.
    - *Degree Major* had weaker associations, and there was no evidence of *Company* association.