Week 11 — Explain the model, analyze risk, bias and ethical considerations

1. Identification of important predictors in our model

Our group keeps the model family fixed and uses an L1 regularized linear model as the best performer from the earlier weeks. For this assignment we focus on explaining what this model is learning once we apply the Week 10 cleaned and smoothed dataset. To identify important features, we align the new design matrix with the feature order saved from training and then compute importance scores based on the absolute size of the coefficients. Larger absolute coefficients mean that small changes in that feature translate into larger changes in the prediction, while coefficients near zero have almost no effect.

The resulting importance ranking is highly concentrated. A small group of predictors dominates the list, while many others have very small weights. In our feature importance table, the top positions are occupied by variables that capture recent behaviour of the target and overall scale effects, along with a few risk or volatility indicators. In other words, the model relies mainly on short term dynamics and level effects, and treats many of the remaining engineered signals as minor adjustments. Across the full list, the top five to ten features account for the majority of total importance, and their coefficients are several times larger in absolute value than the median feature. This pattern matches our expectation that a sparse linear model will automatically shrink redundant or noisy inputs and highlight only the few that truly matter.

Looking at the signs of these top coefficients also gives us an interpretable story. Features that describe strong recent performance tend to push the prediction upward, while features that encode negative shocks or higher risk move the prediction downward. Time related dummies and weaker signals have much smaller magnitudes, so they are mainly fine tuning the output rather than driving it. This combination of sparsity and clear sign structure makes the model relatively easy for our group to reason about and to audit for potential bias in later sections.

2. Local explanations for five individual predictions

To connect global importance back to individual cases, we extract five observations from the evaluation window and examine how the model produced each prediction. For every row we compare the feature values with the overall distribution and interpret the prediction as a weighted sum of features, where each weight comes from the learned coefficient. This gives us a simple way to see which variables are pushing the output up or down for that specific case.

In the first case, the outcome is predicted to be relatively high. The row shows strong values on the top momentum type feature and a sizeable value on another positive scale feature, both far above their typical levels. Negative features such as risk indicators are close to average. Because the dominant positive coefficients are multiplied by unusually large inputs, they explain most of the prediction. To move this output to a more moderate level, those two key features would need to decrease by roughly one standard deviation each, or an equivalent increase would be required in the strongest negatively signed feature.

The second case has a mid range prediction. Here the leading momentum feature is only slightly above its mean, while a volatility feature with a negative coefficient is somewhat elevated. The opposing contributions nearly cancel, and the remaining weaker features adjust the value only slightly. A significant upward shift would require either a sizeable rise in the main momentum feature or a reduction in the volatility measure.

In the third case our group sees a low prediction. Several of the high impact positive features sit below their median levels, and one of the primary downside indicators is high. The model therefore receives strong negative contributions. To flip this case into a high prediction regime, the downside indicator would need to fall back near its lower quartile while at least one of the leading positive features moves from a low value to well above average.

The fourth case is an example where the prediction is close to the threshold that would trigger a different business decision. The top two features have moderate but same direction effects, and a third feature with opposite sign is also moderate. Because no single variable dominates, the prediction is sensitive to joint changes. A coordinated shift of several features by smaller amounts could move the output across the threshold, which is important for later risk and fairness analysis.

Finally, the fifth case shows that the model can still produce a high prediction even when one influential feature is unfavourable, as long as other strong predictors compensate in the opposite direction. This case has a large negative contribution from a risk feature but extremely strong positive contributions from recent performance variables. To reduce the prediction meaningfully, either those positive drivers must weaken, or the risk feature must become even more extreme.

Through these five local explanations, our group links the global feature importance ranking to concrete individuals in the dataset. This helps us understand where the model is stable, where it is sensitive to small movements in dominant features, and how future interventions or policy thresholds might interact with the learned coefficients.

3. Discuss whether your dataset includes protected categories and whether you are using them in your model

We first examined whether the dataset includes protected categories. According to the U.S. Equal Employment Opportunity Commission (EEOC), protected categories include race, gender, age, religion, nationality, disability, and marital status. Our dataset primarily consists of macroeconomic indicators (e.g., inflation, unemployment, treasury yields), company identifiers (e.g., SLG, BXP, TRNO), stock return variables, and time-derived features. It does not contain any legally defined protected attributes. However, fields such as company and metro (metropolitan area) may act as proxy variables that indirectly reflect demographic or regional economic differences, potentially introducing structural bias. Therefore, while the model does not explicitly use protected categories, its predictions may still be influenced by structural disparities across companies or regions.

4.Discuss the bias of the model

For bias analysis, we grouped the data by company and calculated MAE, RMSE, MAPE, and mean residuals. Results show that SLG has the highest error and a positive mean residual, indicating systematic underestimation of its returns. BXP and ELME have moderate errors but negative residuals, suggesting the model tends to overestimate their returns. REXR has relatively low absolute error but the highest MAPE (224%), due to small actual TSR values that amplify relative error. TRNO performs most consistently, with low error and near-zero residuals. EQR has the lowest MAE and RMSE, but a high MAPE due to the same small-denominator effect.

Overall, the model does not exhibit severe systemic bias, as all mean residuals fall within the range of −0.007 to +0.006. However, group-level bias is evident: SLG is consistently underestimated, BXP and ELME are overestimated, and REXR and EQR suffer from inflated MAPE due to low TSR levels. These disparities may lead to unfair resource allocation or misinformed investment decisions if not addressed.

5.Provide bias removal strategies and their impact on the predictions

To mitigate bias, we tested four strategies. First, Reweighting assigns higher training weights to companies with larger errors, improving fairness by reducing SLG’s error, though with a slight trade-off in overall performance. Second, Group-specific Standardization normalizes data within each company, eliminating scale differences and improving generalization and residual balance. Third, Residual Correction adjusts predictions post hoc by adding back the average residual per company, effectively

correcting directional bias but not improving the model itself. Fourth, Group Interaction Features introduce company × macroeconomic variable interactions to capture heterogeneous responses. This enhances model interpretability and improves accuracy for some companies but increases model complexity and variance.

Each method has distinct effects. Reweighting and interaction terms are most effective in reducing bias but may impact stability. Group-specific standardization is a robust, low-risk method. Residual correction is simple and effective for directional bias but does not address structural issues.

6. Optional

We also compared Week 11's subgroup bias analysis with Week 10's overall RMSE improvements. Week 10 showed that data cleaning and smoothing reduced RMSE from ~0.082 to 0.053. However, Week 11 revealed that this global improvement did not translate into uniform fairness across subgroups. SLG's error remained high, BXP and ELME continued to show directional bias, and REXR's inflated MAPE persisted. This highlights a key insight: improving overall model quality does not guarantee fairness across all subgroups. Subgroup-level analysis is essential to avoid unfair or risky outcomes in real-world deployment.

Finally, we identified several model risks. Decision risk arises from SLG's underestimation and BXP/ELME's overestimation, which could mislead asset allocation. Group fairness risk stems from uneven error distribution across companies. Distributional sensitivity risk is evident in REXR and EQR, where low TSR values lead to unstable MAPE. Model interpretability risk remains due to potential coefficient instability in Lasso under multicollinearity, despite Week 10's feature decorrelation efforts.

In conclusion, Week 11's bias analysis reveals significant variation in model performance across companies. While Week 10's data enhancements improved overall accuracy, they did not resolve group-level disparities. Future improvements should include company-specific features or weighting, fairness-aware loss functions, alternative modeling for near-zero targets, and potentially separate or hybrid models for high-bias companies to enhance fairness and robustness.

7.Model risk and stakeholder impact

Beyond statistical bias and subgroup fairness, the model also presents broader risks that affect real-world stakeholders. First, decision risk arises because the model systematically underestimates SLG and

overestimates BXP and ELME. If this model were used in an investment or asset-allocation context, it could shift capital away from SLG and toward BXP/ELME in ways that do not reflect true economic value, potentially harming the firm or misallocating portfolio resources. Second, the model carries generalization risk. Since training is based on a subset of time and firms, changes in macroeconomic conditions—interest rates, inflation, or credit cycles—could make the learned coefficients obsolete, causing unstable future predictions. Third, the model exhibits distributional sensitivity, particularly for firms with very low TSR values (e.g., REXR and EQR), where small absolute deviations cause extremely high MAPE. This instability could mislead analysts during periods of financial stress or abnormal performance.

There is also regulatory and governance risk. If predictions were incorporated into lending decisions, risk scoring, or real-estate investment guidelines, biased outputs could raise compliance issues under fairness regulations, even if the dataset does not contain protected classes. In addition, organizational risk arises when decision-makers place excessive trust in a sparse linear model that may omit important nonlinear interactions present in the real estate market. For residents, tenants, or municipalities, systematic mispricing of firms could indirectly influence construction, borrowing, or development plans. Finally, model maintenance risk is nontrivial: because Lasso coefficients are sensitive to data cleaning and smoothing choices, even small updates to the dataset could dramatically change the selected features. This coefficient instability underscores the need for continuous monitoring, retraining, and robust governance practices before deploying such a model in any operational environment.

8.Comparison of Week 10 improvements vs Week 11 fairness outcomes

Week 10 produced a substantial improvement in global accuracy: RMSE dropped from approximately 0.082 to 0.053 after outlier clipping, missing-value imputation, feature decorrelation, and rolling-median smoothing. However, Week 11's subgroup analysis demonstrates that these global improvements did not translate into uniform fairness across companies. Several groups continue to show meaningful differences in direction and magnitude of residuals: SLG remains consistently underestimated, BXP and ELME remain overestimated, and firms with very low TSR values (REXR, EQR) still show inflated MAPE due to the denominator effect. This gap between global performance and subgroup fairness highlights an important lesson: improving model accuracy does not guarantee equitable treatment across subpopulations. Without explicit fairness adjustments or group-aware modeling, cleaning the dataset can reduce noise but cannot remove structural disparities embedded in historical patterns. Therefore, Week 11 reinforces that fairness evaluation must complement accuracy-driven model optimization to ensure safe and balanced decision-making.