Part 1.

Using earnings data from 2010-2019, we trained three models - HVZ, earnings persistence (EP), and residual income (RI) - as taught in our course. To prepare the data, we first imputed missing values by averaging data from adjacent years and removed any rows where imputation was not feasible. We then tested the accuracy of these models against actual earnings reported from 2020 to 2022, aiming to identify the most precise forecast method.

When choosing the best model, the goal is to find a balance between accuracy (precision of the forecast) and bias (direction and extent of the forecast error). Ideally, we would prefer a model with high accuracy (lower values in the heatmap Figure 1) and bias close to zero (indicating neither optimistic nor pessimistic forecasts).

- HVZ seems to be getting more pessimistic with time, and its accuracy is also decreasing, which is not an ideal combination.
- EP shows the best trend toward balanced forecasts with its bias getting closer to zero, and despite not starting as the most accurate, it becomes the most precise by 2022.
- RI makes the most significant improvement in bias but remains the least precise model throughout all the years, indicating it might not be the most reliable for forecasting purposes.

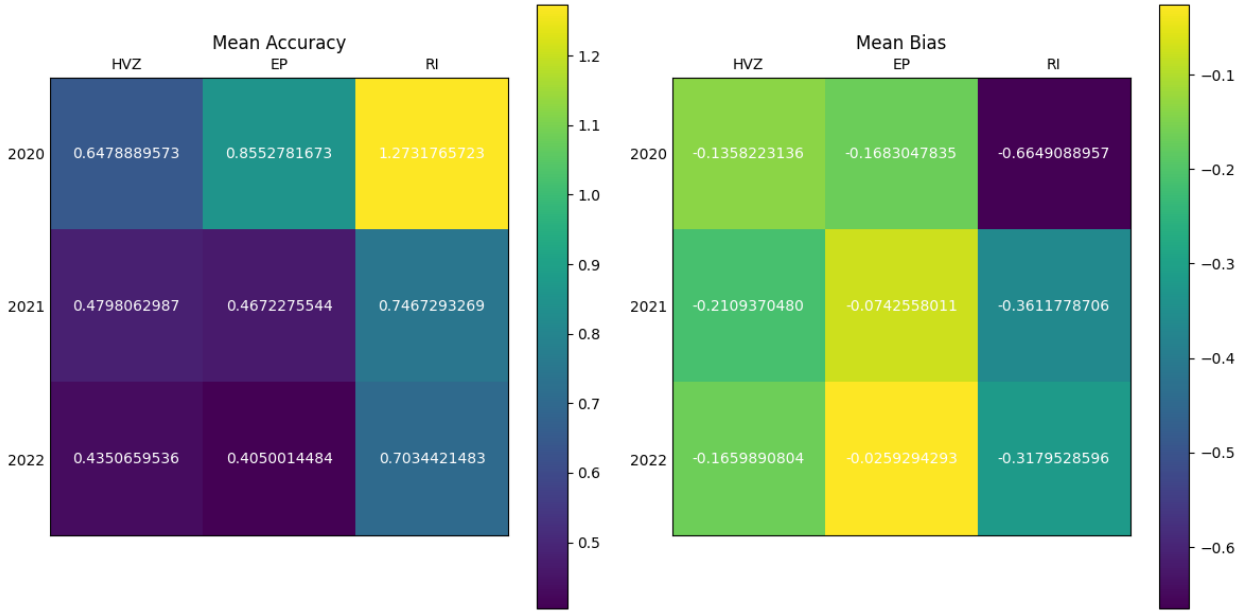

Figure 1: Mean Accuracy and Mean Bias for the three models

We also analyzed the Adjusted R Squared of each model, as shown in Table 1 below. Based on adjusted R-squared values alone, 'HVZ' appears to be the strongest model in terms of explanatory power. 'RI' is in the middle, and 'EP' has the least explanatory power.

	Year	HVZ	EP	RI
0	2020	0.828591	0.154403	0.335193
1	2021	0.742315	0.097332	0.279040
2	2022	0.715506	0.076266	0.267358

Table 1: adjusted R-square values for the three models

Overall, these findings suggest that the best model depends on the specific criteria and goals of the analysis. If the priority is explanatory power and understanding the underlying data structure, 'HVZ' might be preferred. However, our project's focus is on accurate and unbiased forecasts, 'EP' could be more suitable, despite its lower adjusted R-squared values.

Part 2.

In order to further improve the EP model as well as the prediction results from question 1, we first start by considering more factors. We first computed some extra financial ratios and metrics as well as some fundamental EDA to the new dataset. We then fit a Lasso model to determine the best alpha, as we are dealing with multiple predictors. This model grants us access to regularization and cross-validation, in addition to optimizing the alpha parameter that controls the magnitude of the penalty applied to the coefficients.

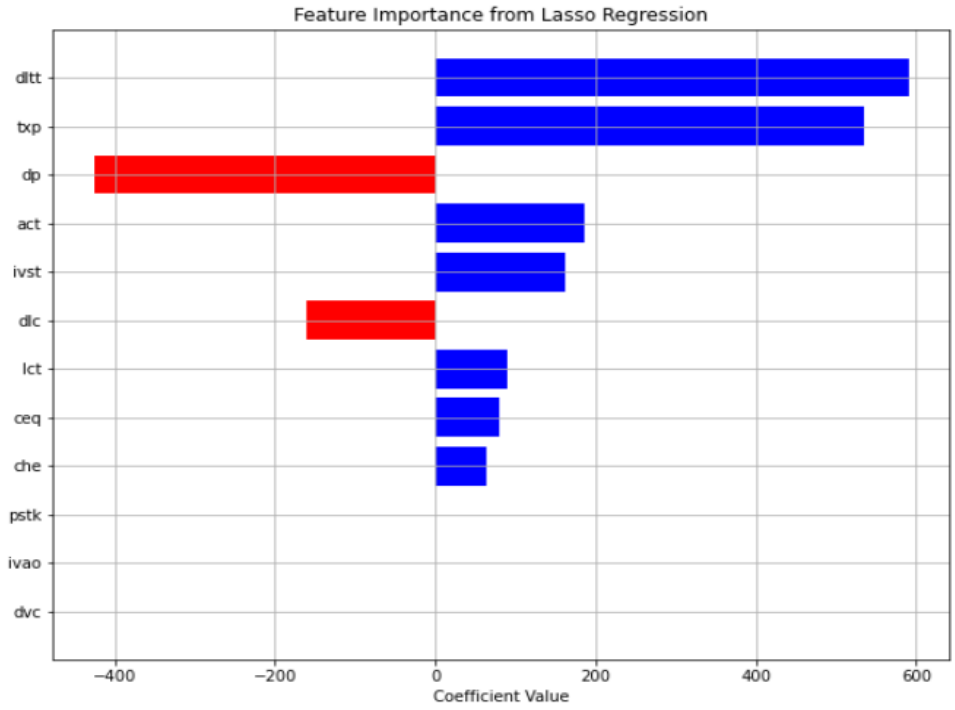

Figure 2: Feature Importance from Lasso Regression

After assessing the feature importance, we compiled the new features into the previous EP model and re-run the procedures performed in question 1. The newly adjusted R-square values are displayed below.

ep_df_2020

OLS Regression Results

Dep. Variable:	earnings_in_years_tau	R-squared:	0.162
Model:	OLS	Adj. R-squared:	0.162
Method:	Least Squares	F-statistic:	1678.
Date:	Thu, 25 Apr 2024	Prob (F-statistic):	0.00
Time:	10:55:32	Log-Likelihood:	-1.6652e+05
No. Observations:	60676	AIC:	3.331e+05
Df Residuals:	60668	BIC:	3.331e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
earning	0.0010	3.44e-05	29.658	0.000	0.001	0.001
NegE	-2.3364	0.033	-70.305	0.000	-2.402	-2.271
NegE_earnings_in_years_t_interaction	0.0031	0.000	12.812	0.000	0.003	0.004
act	2.095e-05	4.32e-06	4.848	0.000	1.25e-05	2.94e-05
txp	-0.0001	8.8e-05	-1.332	0.183	-0.000	5.52e-05
dp	-0.0002	2.71e-05	-7.099	0.000	-0.000	-0.000
industry_freq	-5.3173	0.287	-18.532	0.000	-5.880	-4.755
intercept	1.7448	0.027	63.915	0.000	1.691	1.798

Omnibus:	48729.706	Durbin-Watson:	1.176
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3643254.374
Skew:	-3.350	Prob(JB):	0.00
Kurtosis:	40.365	Cond. No.	1.25e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.25e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
adjust_r_squared is: 0.16213069910508282

ep_df_2021

OLS Regression Results

Dep. Variable:	earnings_in_years_tau	R-squared:	0.104
Model:	OLS	Adj. R-squared:	0.104
Method:	Least Squares	F-statistic:	931.6
Date:	Thu, 25 Apr 2024	Prob (F-statistic):	0.00
Time:	10:55:33	Log-Likelihood:	-1.7414e+05
No. Observations:	56091	AIC:	3.483e+05
Df Residuals:	56083	BIC:	3.484e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
earning	0.0010	4.85e-05	19.762	0.000	0.001	0.001
NegE	-2.7263	0.050	-55.005	0.000	-2.823	-2.629
NegE_earnings_in_years_t_interaction	-1.097e-05	0.000	-0.030	0.976	-0.001	0.001
act	3.32e-05	6.26e-06	5.303	0.000	2.09e-05	4.55e-05
txp	-0.0004	0.000	-2.865	0.004	-0.001	-0.000
dp	-0.0002	3.96e-05	-4.928	0.000	-0.000	-0.000
industry_freq	-7.9355	0.426	-18.620	0.000	-8.771	-7.100
intercept	1.9310	0.040	47.851	0.000	1.852	2.010

Omnibus:	58314.320	Durbin-Watson:	1.165
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6982645.603
Skew:	-5.009	Prob(JB):	0.00
Kurtosis:	56.734	Cond. No.	1.29e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.29e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
adjust_r_squared is: 0.10405471136546507

ep_df_2022

OLS Regression Results

Dep. Variable:	earnings_in_years_tau	R-squared:	0.083
Model:	OLS	Adj. R-squared:	0.083
Method:	Least Squares	F-statistic:	668.8
Date:	Thu, 25 Apr 2024	Prob (F-statistic):	0.00
Time:	10:55:33	Log-Likelihood:	-1.7548e+05
No. Observations:	51836	AIC:	3.510e+05
Df Residuals:	51828	BIC:	3.510e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
earning	0.0010	6.33e-05	15.230	0.000	0.001	0.001
NegE	-3.2730	0.068	-47.942	0.000	-3.407	-3.139
NegE_earnings_in_years_t_interaction	-0.0013	0.000	-2.671	0.008	-0.002	-0.000
act	3.338e-05	8.41e-06	3.969	0.000	1.69e-05	4.99e-05
txp	-0.0007	0.000	-3.981	0.000	-0.001	-0.000
dp	-0.0002	5.42e-05	-2.871	0.004	-0.000	-4.93e-05
industry_freq	-10.3366	0.586	-17.653	0.000	-11.484	-9.189
intercept	2.1585	0.055	39.171	0.000	2.050	2.267

Omnibus:	62615.739	Durbin-Watson:	1.121
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9788185.816
Skew:	-6.485	Prob(JB):	0.00
Kurtosis:	69.058	Cond. No.	1.33e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.33e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
adjust_r_squared is: 0.08272764229325735

Table 2: adjusted R-square values with new features included

We observe that compared to Table 1 results, we indeed saw some improvements as we added more features. Not only does this indicate a slightly better model fit but also demonstrates a reduction in potential overfitting risk. We have experimented with adding more factors that appear to have some significance from the Lasso model. To be more specific, the 2020 regression summary with an adjusted R-squared of 0.162 suggests that around 16.2% of the variability in the dependent variable is explained by the model, which is a modest amount. A high F-value (1678) and a significant p-value suggest the model is statistically significant. Most variables seem to have a significant relationship with the dependent variable, as indicated by their p-values being less than 0.05. For the 2021 model, the adjusted R-squared is lower than the 2020 model, indicating a decrease in the explanatory power of the model. A lower F-value and significant p-value suggest that while the model has explanatory power, it may not be as strong as in previous years.

Overall, there is a downward trend in the adjusted R-squared from 2020 to 2022, which may indicate that the model is becoming less effective in explaining the variance of the dependent variable over time. This could be due to changes in the underlying data or the relationships between variables. The F-statistic remains significant throughout, indicating that at least some of the independent variables are good predictors of the dependent variable.

Note that we did not use textual analysis for an excellent analysis, for a couple of reasons. First of all, the implementation of natural language processing as well as running such machine learning algorithms is timely and costly, especially in real-world situations with even more

extensive datasets. The complexity of text data conversion may not yield meaningful results. Secondly, textual analysis of management and other disclosures in addition to such a quantitative study of earnings prediction can also introduce methodological inconsistencies. We attempted to try text scraping and other basic language detection methodologies, yet we struggled to combine our findings with the prediction results. Thus, we decided to solely rely on quantitative data analytic tools and relevant methodologies.

Part 3.

Based on the coefficients estimated from question 2, we forecast 2020 annual earnings per share (EPS) for ten diverse stocks, ranging from fintech to renewable energy. This selection includes relatively young companies like Square (founded in 2009), Peloton (2012), and PagSeguro Digital (2018), reflecting various growth stages and market dynamics. High-innovation sectors are represented by Moderna, NVIDIA, and AMD, while companies like Peloton and Etsy underscore shifts in consumer behaviour from recent global events. The inclusion of firms sensitive to economic policies, such as Ford and First Solar, tests the model's adaptability to external factors. This strategic assortment ensures the model's effectiveness across different economic scenarios, providing a robust evaluation of its accuracy and versatility in real-world financial forecasting. Here is some detailed information for each firm:

- PagSeguro Digital Ltd. (PAGS) - PagSeguro Digital provides comprehensive fintech solutions targeting small to medium-sized businesses in Brazil, facilitating digital payments and financial services. Operating in a supportive regulatory and increasingly digital environment, PagSeguro navigates the volatile economic landscape of emerging markets.
- Etsy, Inc. (ETSY) - Etsy operates a specialized online marketplace for handmade and vintage items, catering to niche consumer interests in personalized and sustainable products. Despite competition from larger e-commerce platforms, Etsy maintains a loyal customer base, driven by unique product offerings and community engagement.
- Moderna, Inc. (MRNA) - Moderna is a biotechnology company focused on mRNA-based therapeutics and vaccines, notably for COVID-19. Positioned in a highly regulated and rapidly evolving sector, Moderna's success depends on its ability to innovate within stringent scientific and ethical guidelines.
- Peloton Interactive, Inc. (PTON) - Peloton merges fitness and technology by offering interactive equipment and streaming workout classes, thriving in the home fitness trend accelerated by the pandemic. The company faces competition from traditional and new fitness solutions, emphasizing the need for ongoing innovation and market adaptation.
- Advanced Micro Devices, Inc. (AMD) - AMD designs and produces semiconductors, including CPUs and GPUs, competing in a technology-driven industry marked by fast-paced innovation and global supply chain challenges. AMD's growth is contingent on its technological advancements and ability to mitigate external geopolitical and economic pressures.
- NVIDIA Corporation (NVDA) - NVIDIA Corporation dominates the semiconductor industry, specializing in graphics processing units (GPUs) for gaming, data centers, and automotive applications. The company leads in developing advanced technologies like artificial intelligence and machine learning, impacting sectors from healthcare to finance. NVIDIA's intense R&D efforts fuel innovation in computing, making it a key player in the digital landscape and a central interest for investors and technologists.

- Zillow Group, Inc. (Z) - Zillow is a leading digital real estate platform, offering services from property listings to mortgage tools. It innovates within the industry by enhancing user experiences and streamlining transactions. As a major source of real estate data, Zillow provides insights that shape consumer decisions and market trends, positioning it at the forefront of discussions on the future of real estate technology.
- The Trade Desk, Inc. (TTD) - The Trade Desk, Inc. excels in the digital advertising sector, offering a platform for programmatic ad campaigns across various formats. It leads in ad tech innovation, providing tools for data-driven targeting and real-time bidding. By analyzing vast data, The Trade Desk delivers insights that optimize marketing strategies, positioning it as a leader in navigating the fragmented media landscape and influencing industry trends.
- Ford Motor Company (F) - Ford Motor Company, a cornerstone of the automotive industry, offers a diverse vehicle lineup, including advancing electric vehicles (EVs). Industrially, it leads in vehicle innovation, focusing on safety, efficiency, and connectivity. Ford's extensive R&D in autonomous driving and mobility solutions positions it as a key influencer in future transportation technologies, shaping industry standards and evolving consumer expectations in the rapidly changing automotive sector.
- First Solar, Inc. (FSLR) - First Solar is a leader in the solar energy industry, specializing in manufacturing photovoltaic (PV) solar panels and providing comprehensive solar power solutions. In business, First Solar focuses on sustainable energy technologies with a commitment to low-cost, high-efficiency solar modules. Industrially, it drives innovation in solar power, enhancing grid reliability and performance. Through robust R&D efforts, First Solar influences renewable energy policies and market trends, positioning itself as a key player in advancing global energy transitions toward sustainability.

The forecasts are computed and displayed below.

	tic	earnings_in_years_tau	expected_earnings_in_years_tau	bias	accuracy	actual	medest	bias_analyst	accuracy_analyst
0	AMD	2.186325	2.010617	0.000005	0.000005	1.2900	1.23	-0.000022	0.000022
1	F	-1.121583	2.425411	-0.000088	0.000088	0.4100	-0.03	-0.000050	0.000050
2	ETSY	3.108802	1.588756	0.000206	0.000206	2.6900	2.13	0.000149	0.000149
3	TTD	5.328459	1.597118	0.000425	0.000425	0.6850	0.49	-0.000104	0.000104
4	PAGS	0.748230	1.824580	-0.000170	0.000170	4.3500	4.31	0.000399	0.000399
5	MRNA	-2.219857	-3.089160	0.000180	0.000180	-1.9600	-1.49	0.000234	0.000234
6	NVDA	7.514706	4.253652	0.000033	0.000033	1.4475	1.39	-0.000028	0.000028
7	FSLR	4.142808	1.734481	0.000348	0.000348	3.7300	3.95	0.000288	0.000288
8	Z	-0.414046	-2.119341	0.000261	0.000261	4.9300	4.92	0.001081	0.001081

Table 3: Difference between forecasts and predictions

We see that from Table 3, the last two columns represent the differences between analyst forecasts in IBES and the prediction results using coefficients from question 2. We see that the differences are quite small, indicating that our forecasts are quite close to that of the analysts'. The differences between our forecasts to the actual statistics are also relatively small, indicating our prediction results are indeed quite accurate.

Part 4.

The logic behind using our selected metric is to assess a firm's profitability and growth potential. EPS is a widely-used metric to gauge a company's profitability. A higher expected EPS suggests that the company is expected to generate more profit per share in the future, which could be seen as a positive sign by investors. By comparing the expected future EPS to the current EPS, investors can get an idea of the company's growth potential. A positive growth rate indicates that the company's earnings are expected to increase, which is typically a sign of a healthy, growing company. Investors often look for companies with strong growth prospects. By filtering for companies with a projected increase in EPS and calculating the growth rate, the dataset now contains firms that may be considered more attractive investment opportunities. Lastly, the growth rate of EPS can be used to benchmark a company's performance against its peers or the market average. Firms with higher growth rates might be outperforming their competitors.

This analysis can be particularly important for forward-looking assessments, strategic planning, or valuation exercises where future profitability and performance are considered. It's worth noting that while EPS growth is a useful indicator, it should not be the sole basis for investment decisions; a comprehensive analysis would consider a range of financial metrics alongside EPS growth.

	earning	NegE	NegE_earnings_in_years_t_interaction	earnings_in_years_tau	act	txp	dp	industry_freq	eps_22
18929	0.001	0	0.0	0.000037	0.208	0.000	0.000	0.009834	0.000037
14545	0.204	0	0.0	0.000055	0.543	0.175	0.540	0.011715	0.000055
78587	0.042	0	0.0	0.000112	0.208	0.000	0.000	0.011715	0.000112
90123	0.405	0	0.0	0.000238	7.502	0.095	0.082	0.009834	0.000238
99077	0.189	0	0.0	0.000369	2.686	0.000	0.050	0.075417	0.000369
100763	0.077	0	0.0	0.000554	0.208	0.000	0.000	0.003084	0.000554
19782	0.027	0	0.0	0.000508	1.138	0.000	0.003	0.095180	0.000508
71574	0.303	0	0.0	0.000588	5.741	0.000	0.403	0.109706	0.000588
794	0.003	0	0.0	0.000911	0.208	0.051	0.000	0.009996	0.000911
69973	0.112	0	0.0	0.000753	4.142	0.195	0.000	0.095180	0.000753

Table 4: Top 10 firms most likely to exhibit the largest earnings growth in 2023

Part 5.

Using qualitative factors, we picked the top three firms to be Microsoft Corp., Taiwan Semiconductor Manufacturing Company, and Pfizer Inc.

As a titan in the realms of software, cloud services, and artificial intelligence, Microsoft consistently forges paths in burgeoning sectors, with Azure and AI applications leading the charge. The company's solid financial foundation is complemented by an expansive global presence, guaranteeing stable revenue from a diversified portfolio that spans multiple industries.

TSMC stands as the preeminent independent semiconductor foundry on the globe, essential to the electronics industry's supply dynamics. In an era where semiconductor technology is ever more critical, TSMC's unparalleled manufacturing prowess and its position at the forefront of technological advancements solidify its status as a pivotal player and an attractive investment destination.

Pfizer's legacy is anchored in its innovative drug development and notable achievements, including the rapid development and deployment of a COVID-19 vaccine. Its commitment to pushing the boundaries in biopharmaceuticals is backed by an extensive research apparatus, efficacy in global market penetration, and a well-established proficiency in scaling production. Pfizer's proactive approach to R&D and its adeptness in regulatory compliance make it a standout in the pharmaceutical sector, promising sustained growth and investment potential.